DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

MACHINE LEARNING BASED FUSION OF DIFFERENT SEGMENTATION TECHNIQUES FOR LIVER VISUALIZATION FOR ENHANCED ACCURACY AND SENSITIVITY

by

Ali Emre KAVUR

August, 2020

İZMİR

MACHINE LEARNING BASED FUSION OF DIFFERENT SEGMENTATION TECHNIQUES FOR LIVER VISUALIZATION FOR ENHANCED ACCURACY AND SENSITIVITY

A Thesis Submitted to the
Graduate School of Natural And Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Electrical and Electronics Engineering

by

Ali Emre KAVUR

August, 2020
İZMİR

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**MACHINE LEARNING BASED FUSION OF DIFFERENT SEGMENTATION TECHNIQUES FOR LIVER VISUALIZATION FOR ENHANCED ACCURACY AND SENSITIVITY**" completed by **ALİ EMRE KAVUR** under supervision of **ASSOC. PROF. DR. M. ALPER SELVER** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

.....

Assoc. Prof. Dr. M. Alper SELVER

Supervisor

.....	.....
Prof. Dr. Oğuz DİCLE	Assoc. Prof. Dr. Ahmet ÖZKURT
Thesis Committee Member	Thesis Committee Member
.....	.....
Prof.Dr. Gözde ÜNAL	Prof.Dr. Ludmilla KUNCHEVA
Examining Committee Member	Examining Committee Member

Prof.Dr. Özgür ÖZÇELİK
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

First of all, I would like to thank my Ph.D. supervisor Assoc. Prof. Dr. M. Alper SELVER for his consistent support and guidance during this thesis. The conversations and his opinions were vital in inspiring me to think outside the box, from multiple perspectives to form a comprehensive and objective critique.

I would like to thank Prof. Dr. Ludmilla KUNCHEVA, who is my supervisor during my academic visit to Bangor University, for her dedicated support and guidance. She continuously provided encouragement and was always willing and enthusiastic to assist in any way she could throughout the research project.

I would like to thank Prof. Dr. Oğuz DİCLE and Assoc. Prof. Dr. Ahmet ÖZKURT for guiding me as the jury in my Ph.D. committee.

I would like to thank Assoc. Prof. Dr. Sinem GEZER and Dr. Mustafa BARIŞ for their valuable contributions to my studies.

I would like to thank Özge CANLI, Ekrem YAVUZ, Gizem KALENDER, Enes DURBABA, İpek LÖK, Nail AKÇURA, Tuğçe TOPRAK, Ulaş YÜKSEL, and Esranur KAZAZ whose support as part of my Ph.D. allowed my studies to go elevated.

Last but not least, I express my sincere thanks and gratitude to my family and Esin CANDEMİR for all the support they have shown me through this thesis.

Ali Emre KAVUR

MACHINE LEARNING BASED FUSION OF DIFFERENT SEGMENTATION TECHNIQUES FOR LIVER VISUALIZATION FOR ENHANCED ACCURACY AND SENSITIVITY

ABSTRACT

Medical imaging is a vital resource in medicine, improving quality, sensitivity, and objectiveness of diagnosis by providing unique information. Medical image analysis tools provide segmentation capabilities to focus on a specific structure of interest. Thus, segmentation methods play a crucial role to support a variety of imaging operations such as diagnosis, structural analysis, treatment and surgery planning.

The liver is one of the abdomen organs with the highest imaging demand. This situation increases the need for the segmentation of the liver. Despite the various segmentation methods in the literature, the success of them developed up to the last decade is now outperformed by Deep Learning Models (DMs). However, DMs' effectiveness is highly dependent to user experience, specific design procedures, and the data characteristics. Besides, DMs tend to overfit to the training data due to the fundamentals of their designs. Thus, the reproducibility of analytical studies involving DMs are still limited. To eliminate all problems of DM based segmentation methods; fusion of multiple segmenters can be used as an alternative approach.

In this thesis, novel studies on creating new benchmark platforms for segmentation methods, analysis and adaptation of ensemble methods to medical image segmentation applications, and designing of a new ensemble method are presented. It is expected that the findings and proposed solutions in this thesis will help to remove barriers between academic studies and their implementations in real-world applications. (This work is supported by TUBITAK ARDEB-EEEAG under grant number 116E133 and TUBITAK BIDEB-2214 International Doctoral Research Fellowship Programme.)

Keywords: Classifier ensembles, abdominal imaging, image segmentation, medical imaging challenges

KARACİĞERDE BAŞARIMI VE DUYARLILIĞI GELİŞTİRİLMİŞ BÖLÜTLEME İÇİN FARKLI YÖNTEMLERİN MAKİNA ÖĞRENMESİ TABANLI EN İYİLENMİŞ FÜZYONU

ÖZ

Tıbbi görüntüleme, sağladığı anatomik veriler sayesinde kliniklerde kullanılan önemli bir araçtır. Tıbbi görüntü analiz araçları istenilen yapıya odaklanmak için segmentasyon yöntemleri sağlarlar. Bu yöntemler; tanı, anatomik yapı analizi, cerrahi planlama gibi çeşitli işlemlerinde çok önemli bir role sahiptir.

Karaciğer, görüntüleme ve segmentasyon talebi en yüksek olan organlarından biridir. Literatürdeki çeşitli segmentasyon yöntemlerine rağmen, artık Derin Öğrenme Modelleri (DM'ler), geliştirilen diğer yöntemlerden çok daha başarılı sonuçlar üretmektedir. Bununla birlikte, DM'lerin etkinliği büyük ölçüde kullanıcı deneyimine, özel tasarım prosedürlerine ve veri özelliklerine bağlıdır. Ayrıca, DM'ler tasarımları gereği eğitim verilerine oturma eğilimindedir. Bu nedenle, DM'leri içeren çalışmaların tekrarlanabilirliği hala sınırlıdır. DM tabanlı segmentasyon yöntemlerinin sorunlarını ortadan kaldırmak için; farklı segmentasyon metotlarının füzyonu alternatif bir yaklaşım olarak kullanılabilir.

Bu tezde, segmentasyon yöntemleri için yeni karşılaştırma platformlarının oluşturulması, füzyon yöntemlerinin tıbbi görüntü segmentasyonu uygulamalarına analizi ve adaptasyonu, ve yeni bir topluluk yönteminin tasarlanması ile ilgili yeni çalışmalar sunulmaktadır. Bu tezde elde edilen bulguların ve önerilen çözümlerin akademik çalışmalar ile bunların gerçek hayattaki uygulamaları arasındaki engelleri ortadan kaldırmaya yardımcı olması hedeflenmektedir. (Bu çalışma, TÜBİTAK ARDEB-EEEAG tarafından 116E133 numaralı proje ve TÜBİTAK BIDEB-2214 Uluslararası Doktora Araştırma Burs Programı kapsamında desteklenmiştir.)

Anahtar kelimeler: Sınıflandırıcı topluluklar, abdominal görüntüleme, görüntü bölütleme, tıbbi görüntüleme yarışmaları

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

Due to the continuous increase of the human population, the demand for medical imaging applications is increasing. In the light of these needs, the capabilities of medical imaging devices are constantly increasing. Owing to technological advancements in medical imaging technologies, detailed and accurate information of the human anatomy can be provided by modalities such as Computed Tomography (CT) and Magnetic Resonance Imaging (MR). Thus medical imaging is a fundamental resource in medicine. The methods in medical image analysis are often used for improving understanding of anatomy, thereby promoting diagnosis and treatment preparation. Modern techniques providing more comprehensive and insightful examinations need larger data. Besides, the rich content of medical images makes them hard to interpret and analyze. Therefore, computer algorithms are widely used to process medical images in order to accelerate analyses as well as improving their accuracy.

For all these reasons, the need for medical image processing tools (i.e. software) is becoming vital. Such tools should include essential features to help and automate particular parts of the workflow in radiology such as diagnosis, volume measurements, tissue quantification, pathology location, surgery, and/or treatment planning. One of the most used features offered by medical imaging tools is segmentation. Segmentation can be defined as separating an image into relevant divisions (Shapiro & Stockman, 2001). Such divisions might refer to specific types of tissues, pathologies, or other biologically important structures in medical images.

1.1 A General Overview of Medical Image Segmentation Literature

For several years, segmentation has become an important research field. The segmentation of various structures or organs in the human body is needed in clinical applications. Recent studies show that segmentation became the most researched area

of biomedical image processing and accounted for around 70.0% of all research (Maier-Hein et al., 2018). While several image segmentation methods are available for computer vision, some of them are uniquely tailored for medical image analysis. One of them is called the atlas-based segmentation method which uses several manually labeled medical images. These labels are summed and extrapolated to create single atlas (Gee et al., 1993). For different sizes of data, this method needs image registration to adjust images. In shape-based segmentation methods, a shape of reference is altered by some features in the images especially along borders. The deformation of the shape continues to fit a new object. Active Shape Models (Cootes et al., 1995) is one of the most popular shape-based segmentation techniques. On the other hand, the deformation of a reference shape can be controlled by integral error metrics such as in Active contour models (Goldenberg et al., 2001).

Recent improvements in Machine Learning (ML) have changed the segmentation studies as well as other fields. On the last decade, Deep Learning-based segmentation methods, specifically Deep Learning Models (DM), has clearly outperformed traditional segmentation methods (Zhou et al., 2017; Kavur et al., 2020b). DMs automate the majority of the stages performed by users in other image segmentation methods. For example, they automatically create feature maps to classify the targets. Despite feature extraction followed by classification-based segmentation in traditional ML approaches, once a DM is trained, it can be used to segment new unseen data in the light of the features automatically obtained during the training session. Accordingly, the need for big data of DMs is higher than other segmentation methods. Another key requirement of DMs is the high level of user experience. Implementations of DMs are distinctly harder than other segmentation methods. The intuitiveness of the traditional segmentation methods are higher compared to DMs. On the other hand, DMs need tailored fine tuning for the specific data. Another drawback of DMs is that they tend to overfit to the training data. Overfitting decreases the generalization capabilities of DMs when they are used for unseen data. All of these disadvantages make the implementation of DMs developed in academic studies to real-life tasks harder.

1.2 Liver Segmentation Literature

Many of the organs in the human abdomen such as liver, kidneys, spleen, pancreas, prostate need to be segmented in clinical routines. Segmentation of the liver is one of the most requested tasks due to the liver's several clinical operations such as volumetric measurements (Lu et al., 2017), tumor detection (Li et al., 2018; Chlebus et al., 2018; Christ et al., 2016; Vorontsov et al., 2019), diagnosis (Moghbel et al., 2018), disease detection (Bal et al., 2018), surgery planning as well as treatment planning (Yang et al., 2018). As segmentation of other structures in the human body, DMs have dominated the methods for liver segmentation. However, the availability of extensively annotated abdomen image datasets limits the capabilities of DMs.

1.3 A Review of Grand-Challenges

The continuous requirement for new data and increasing competition between scientific studies around the world have made the grand challenges on biomedical imaging very important. In the past, it was possible to make a study and to publish its outcomes with private data and self-evaluation. However, nowadays the proposed algorithms should prove their success on open benchmark platforms with public datasets. Thus, challenges are now very vital organizations for academic studies ever than before. On the other hand, while the number of competitions organized is increasing every day, there are gaps in the literature regarding making these competitions fairer and more convenient. Fortunately, new studies and regulations for these deficiencies are now being studied (Reinke et al., 2018a; Maier-Hein et al., 2019).

1.4 Literature on Ensembles and Classifier Fusion

Results of many grand challenges clarified that DMs dominated the segmentation studies in medical imaging (Kavur et al., 2020b; Kamnitsas et al., 2017; Zhou et al.,

2017). In so many recent medical imaging challenges related with abdomen organ segmentation, participants used variations of proposed DMs such as Deepmedic, U-Net, V-Net, Dense V-Networks, etc. (Kamnitsas et al., 2016; Ronneberger et al., 2015; Milletari et al., 2016; Gibson et al., 2018). However, there are still fundamental problems in these studies. As explained before, the lack of reproducibility and generalization due to overfitting and requirement for big data are two main obstacles for DM studies. Here an alternative approach may help to overcome these problems: fusion/ensemble of multiple classifiers. Classifier ensembles are being used for years for many fields such as pattern recognition (Kuncheva, 2014; Oza & Tumer, 2008; Rokach, 2010). Adaptation of these techniques in the medical image segmentation problem can reduce overfitting and extensive training need of DMs. For example, the reports of some grand challenges on medical imaging revealed that the winner algorithms use ensembles (Kamnitsas et al., 2018; Isensee et al., 2019). Although ensemble methods such as STAPLE (Warfield et al., 2004) have been studied in the literature, there are no sufficient studies such as their application to deep learning techniques.

1.5 Thesis Statement and Contributions

In this thesis, the problem of liver segmentation is addressed from several different perspectives. Novel studies on creating broadly annotated abdomen image data, creating extensive benchmark platforms to compare state-of-the-art segmentation methods, extensively analyzing classifier ensemble methods to eliminate problems of DMs, and offering a new fusion method for more accurate and robust liver segmentation. All these studies helped to analyze the weak points of existing segmentation methods, to improve their weaknesses, and to develop more robust methods. The main contributions in this thesis can be described as:

1.5.1 Creating a new public dataset for abdomen imaging

Our literature researches showed that there are only a few publicly available datasets for academic studies about abdomen imaging. Due to the high costs of processing and annotating, these volumetric data sets only include a few tens of images (Heimann et al., 2009; Bilic et al., 2019; Menze et al., 2015). This amount of data may be inadequate for convenient examinations of DMs. Therefore, in the scope of this thesis two novel datasets have been prepared and donated to scientific studies. The first dataset contains abdomen CT scans of 40 patients with annotation of livers. The second dataset includes abdomen MRI scans with T1-DUAL and T2-SPIR sequences of 40 patients. The liver, both kidneys, and spleen were annotated in this unique data. According to our knowledge, this dataset is the only set that covers four abdomen organs acquired by MRI scans. The details of both sets are presented in Section 3.4.2.

1.5.2 Organizing grand challenges on abdomen organ segmentation

After creating unique datasets, the second topic was organizing new grand challenges on the abdomen organ segmentations. Our literature analyses revealed that the challenges for abdomen organ segmentation may be considered as outdated. Also, there was not any kind of challenge for medical image segmentation in Turkey. Therefore two challenges have been organized. The first challenge was a nation-wide organization and it is called "Karaciğer Bölütleme Algoritmaları Yarışıyor!". The task for this challenge was the segmentation of the liver from CT scans. 11 teams from different universities participated. The details of this challenge are presented in Section 3.3 as well as its results in Section 3.3.4.

The second challenge, CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation, is an international event. CHAOS has started in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI), 2019, in Venice, Italy. Now it is an online challenge where more than 1500 participants have been registered

to the challenge. Unlike other abdomen challenges, CHAOS has five competitive tasks to push DMs to their limits. These tasks use CT and MRI data as single or combined. This makes the CHAOS challenge is the first challenge that uses abdomen MRI scans in various tasks. According to our knowledge, the CHAOS challenge is the most popular abdomen segmentation challenge in the world. The details and discussions can be found in Section 3.4 and 3.4.7.

1.5.3 Adapting fusion methods to segmentation of liver problem

Although the success of the fusion/ensemble methods has been proven in many fields (Kuncheva, 2014; Zhang & Ma, 2012; Rokach, 2010), it has not been used in medical image segmentation yet. In this section, appropriate ensemble methods for liver segmentation in the literature have been examined. These methods were then applied to the probability maps (Chen et al., 2020) produced by DMs with their vanilla style. In other words, the publicly available DMs were used as their downloaded native versions. Thus, there was no need to make heavy optimizations, parameter tuning, and domain adaptation. Our results showed that it is possible to achieve similar scores with the ensemble of multiple DMs as heavily optimized single DM. Also, classifier ensembles showed less overfitting than individual DMs.

After the discovery of DMs' potentials, a new fusion method has been designed for liver segmentation problem. This new combiner is called Logit combiner which was inspired by adaptation linear regression methods for mapping purposes (Hilbe, 2009). The control parameter of the mapping has improved segmentation accuracy. All studies in this chapter were done by using two public datasets. The evaluation results and their statistical significance are discussed in Chapter 4 and Section 4.6

1.5.4 Adapting fusion methods to segmentation of liver veins problem

The promising results obtained from the fusion approach for liver segmentation have created motivation for looking for a harder problem. While this thesis was

proposed, vessel segmentation from CT angiography images was not planned before. Since there is not yet sufficient solution in the area of vascular segmentation of the liver, it was decided to prolong the study.

Here instead of using different base DMs, the ensemble of the same DM used with rotated volumes has been used due to the shape and complexity of the liver vessel tree. Our findings showed that individual DMs are not ready to segment the vascular tree in the liver yet. Therefore, using classifier fusion seems like a requirement beyond a preference for this task. Again ensemble methods improved the overall accuracy while eliminating overfitting of DMs. The ensemble design and the results are presented in Section 5.3.5.

To sum up, all contributions of the thesis, bringing new data to the literature, analysis of the existing segmentation methods proposed so far and creating new methods to overcome the deficiencies of existing methods have been achieved. With these contributions, it is expected that the obstacles between academic studies about organ segmentation and their real-life applications will be overcome.

The thesis is organized as follows. Chapter 2 introduces the abdomen of the human body, medical imaging modalities, medical image processing techniques, analysis of all organ segmentation methods from past to present. Chapter 3 describes two organized grand challenges in biomedical image analysis. Chapter 4 includes the fusion of different methods for the segmentation of the liver as well as designed new fusion methods. Chapter 5 extends the ensemble approach to a harder and novel problem: segmentation of the liver veins from CT angiography. Finally, Chapter 6 contains discussion of the results and final thoughts.

CHAPTER TWO
BACKGROUND

2.1 Abdomen of Human Body

The abdomen area of the human body is located between thorax and pelvis. It contains vital organs for digestive, exocrine, defecation, endocrine systems. The liver, kidneys, spleen, pancreas, stomach, colons with an attached appendix, and gallbladder are placed together in the abdomen. Besides, main blood vessels such as the aorta and vena cava are located from top to bottom in the abdominal region and provide blood flow to these critical organs. Illustrations from different views are presented in Figures 2.1, 2.2, 2.3.

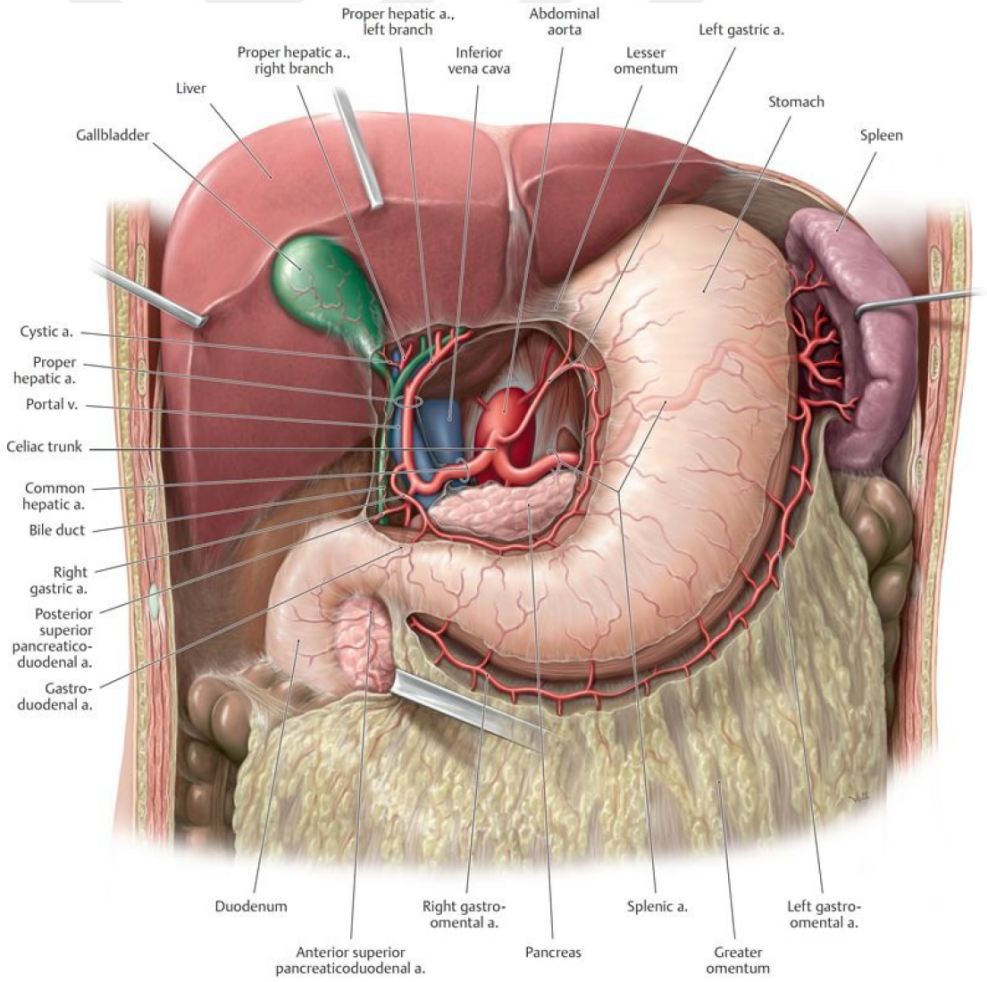

Figure 2.1 Anterior view of human abdomen (Gilroy, 2013)

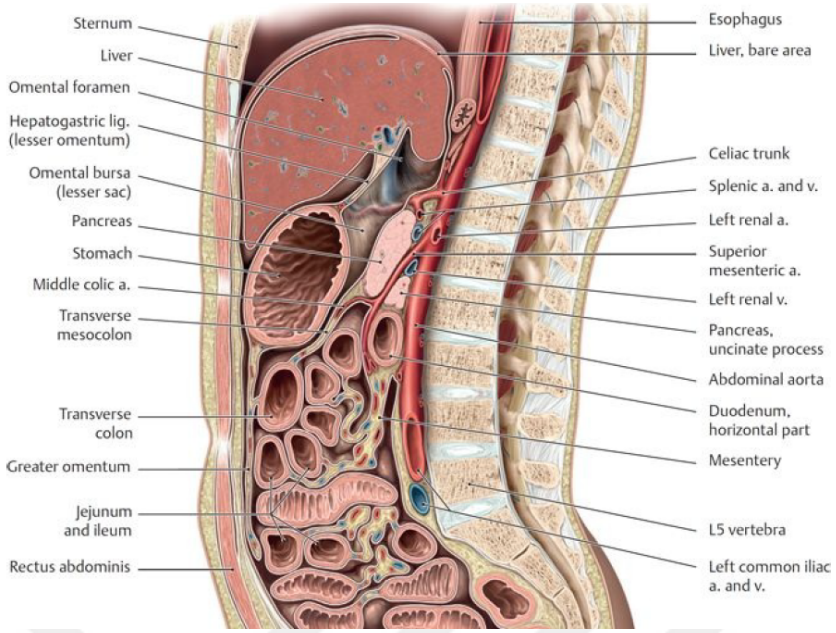

Figure 2.2 Sagital view of human abdomen (Gilroy, 2013)

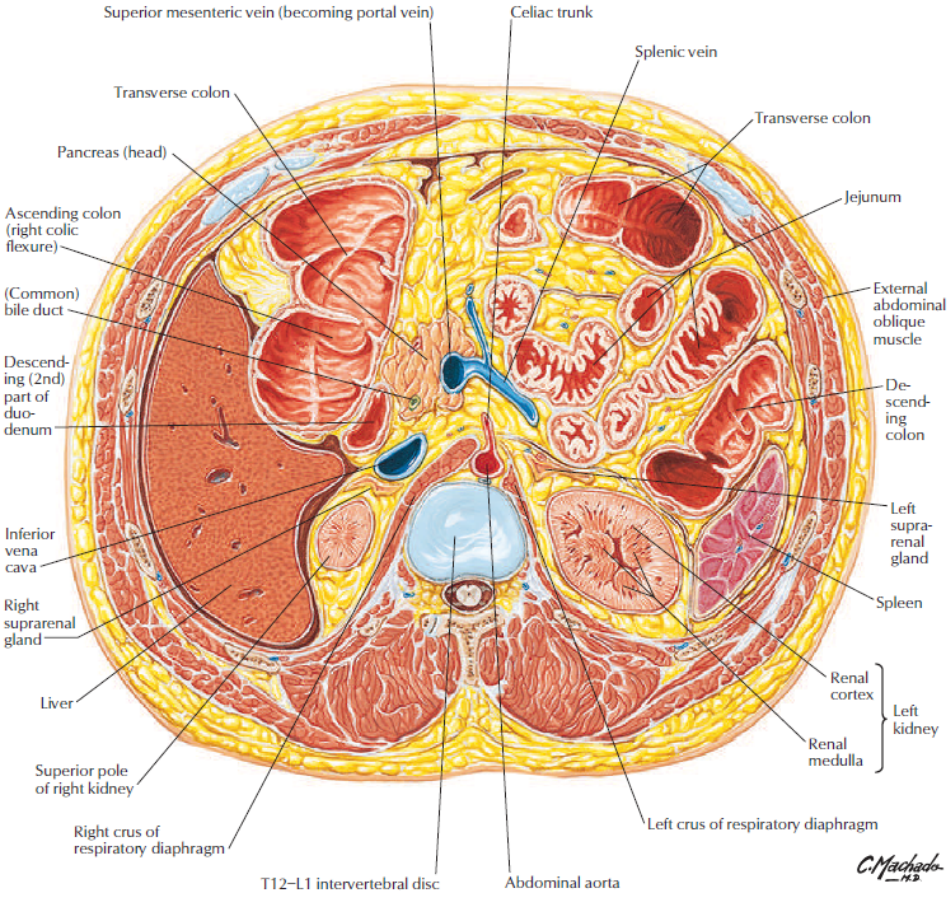

Figure 2.3 Axial view of human abdomen (Netter, 2018)

In the abdomen many organs are connected to each other. The gallbladder is attached to the liver. The liver and the pancreas work closely via ducts in digestive and endocrine systems. The peritoneum, that is a membrane, is located in the abdomen to wrap abdomen organs. The peritoneum holds abdomen organs together.

The part of the aorta in the abdominal cavity is the biggest vein in this region. Aorta begins from the hearth, passes through the thorax area and it reaches to the abdomen. Abdomen aorta passes through the posterior wall of the abdomen. The second-largest vein, inferior vena cava (IVC), is placed just near of the aorta. IVC is located parallel to the aorta. The both veins are responsible for transmitting and receiving blood from vital abdomen organs.

2.1.1 Liver and Blood Vessels in the Liver

The liver is the largest gland in the body. The liver is a wedge-shaped organ located on the right side of the abdominal cavity. The average weight of the liver is 1.5 kg (Kumar et al., 2010). The liver is responsible for many vital processes in the human body such as protein synthesis, secretion of biochemicals for digestions, excretion of hormones, detoxification of many metabolites, regulation of glycogen storage, decomposition of red blood cells. The liver is an external digestive organ that produces bile, (a substance that includes cholesterol and bile acids), and an alkaline compound that helps to break down fat.

There are multiple veins inside the liver as shown in Figure 2.4. Portal vein and hepatic artery are two large blood vessels that carry out blood circulation from the entire gastrointestinal tract and aorta respectively. The structure of these two main blood vessels separates the liver into lobes. A thin, dense, fibroelastic connective tissue layer (known as Glisson's capsule) which extends from the fibrous capsule covering the whole liver keeps the lobules together.

Due to its role in many essential systems in the human body, the liver is a crucial organ. Unfortunately, there is no human-made alternative to compensate for the liver's

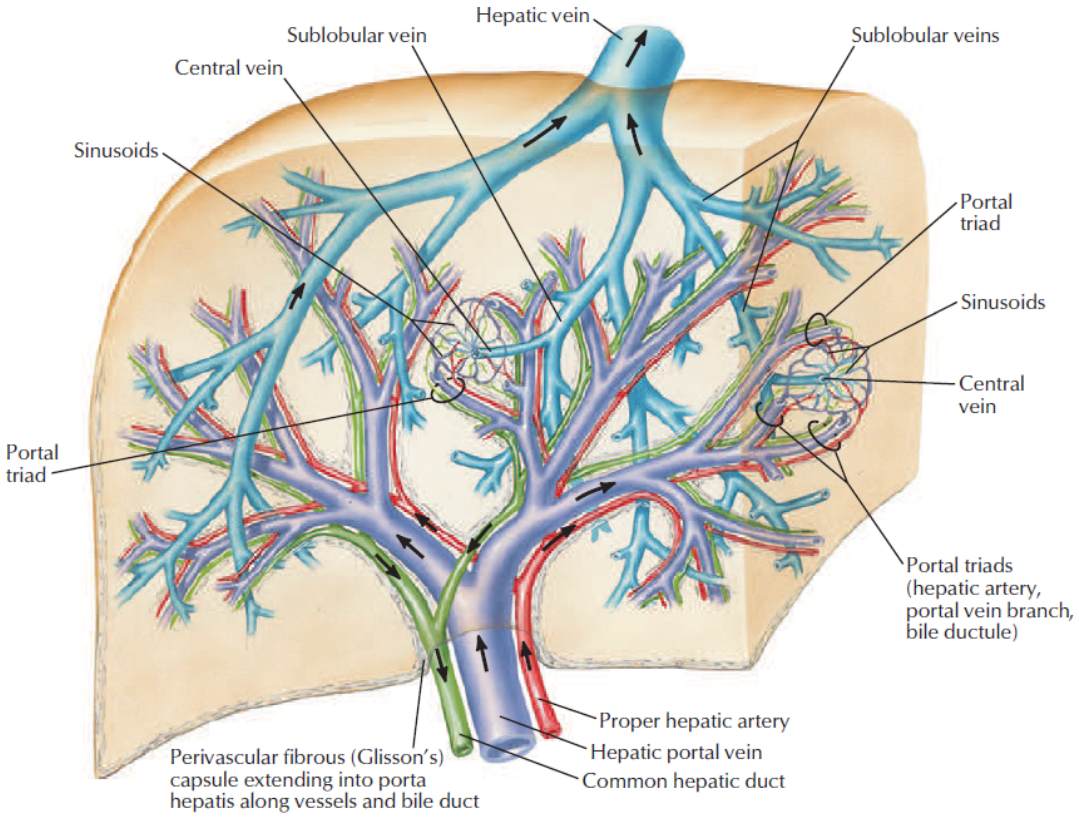

Figure 2.4 Overview of the veins inside of the liver (Netter, 2018)

functions. In other words, the diseases in the liver have to be cured in order to keep the human body functional and healthy. That is why there are many treatment and surgical operations for the liver. These operations have to be handled very carefully because of the complexity of the liver. This makes the liver is one of the most scanned organs in medical imaging.

2.1.2 Kidneys and Spleen

There are other essential organs in the abdominal cavity such as kidneys and spleen shown in Figure 2.5. Kidneys are located left-back and right-back side of the spine in a pair. The left kidney lies behind the diaphragm and the back of the spleen. The right kidney lies slightly behind the diaphragm and the back of the liver. The right kidney is marginally lower than the left one, while the left kidney is more vertical than the right, because of the irregularity in the abdominal cavity created by the liver. Kidneys

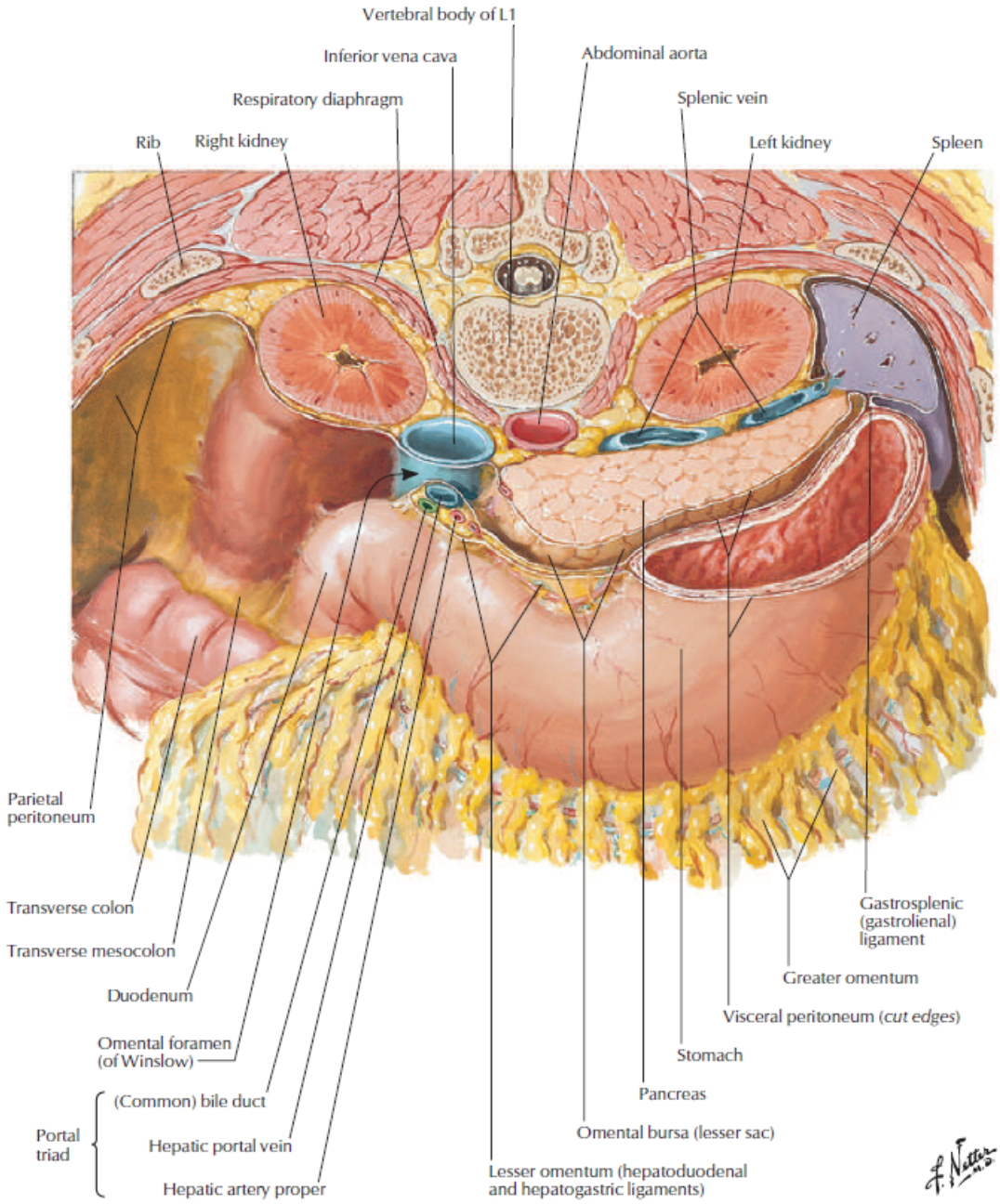

Figure 2.5 Axial view of kidneys, spleen and other abdominal organs (Netter, 2018)

are a relatively smaller organ than the liver. Each adult kidney has an average of 129 g (range, 79-223 g) (Molina & DiMaio, 2012). The primary function of the kidneys is in the urinary system for excreting wastes in the blood. Kidneys are also directly or indirectly responsible for keeping acid-base balance and blood pressure stability as well as managing of electrolytes. Besides, the kidneys have a role in the production of many hormones.

The spleen is in the left upper abdominal quadrant. Spleen is also a large organ in the abdomen with an average weight of 139 g (range, 43-344 g) (Molina & DiMaio, 2012). Spleen primarily serves as a blood filter. It is responsible for the regeneration of iron in the blood. This makes spleen is a crucial organ for regulating red blood cells. Spleen also has roles in the immune system with producing antibodies.

2.2 Medical Imaging Systems for Abdomen

Since Wilhelm Conrad Roentgen discovered X-rays in 1895 by, the capabilities of medical imaging systems have expanded tremendously. Current imaging technologies and techniques enable the processing of anatomical and physiological details from the human body. Medical imaging systems can be considered as a chain operation from the acquisition of the image to processing them in order to extract information. Digital image processing methods have critical importance for further detailed analysis of radiology experts. Experts may need various details derived from the images in order to plan treatments, surgeries, and other operations.

Nowadays, there are many modalities for abdomen imaging in clinical usage. Computer Tomography (CT) and Magnetic Resonance Imaging (MRI) can be considered the most commonly used techniques. Such approaches utilize various features of the human body to collect data and to turn them into images. For example, CT can be explained as a combination of multiple X-ray scans from different angles. On the other hand, MRI uses completely different features which are proton density and relaxation mechanisms. Many different scanning protocols are created such as injection of contrast agents during the scan.

Each modality has advantages and disadvantages against other ones. For instance, an abdomen CT scan of an adult human takes less than 30 seconds while MRI for the same region may need 20 minutes. Nonetheless, CT may cause effects of radiation emission in the body due to the usage of X-rays. MRI does not have such an effect on the body because it uses electromagnetic waves.

Both CT and MRI are capable of creating 3-dimensional images of the targeted region in the human body. In the following subsections, details of CT and MRI are explained.

2.2.1 Computer Tomography (CT)

CT uses multiple X-ray scans from different angles as mentioned previously. Radon transform which was created by Johann Radon in 1917 constitutes the fundamentals of CT imaging. Godfrey Hounsfield developed the first successful CT scanner in 1967. He adapted the algebraic reconstruction (ART) technique of Allan McLeod Cormack.

Figure 2.6 Photo of a modern CT scanner (Wikimedia Commons, 2020)

The attenuation of X-rays into the human body creates slice images of the target location. Rotation of the X-ray tube 360° around the body creates multiple angular projections. These projections are called sinogram and they are used to construct a 3-dimensional image of the target region with the help of ART. ART makes possible for the reconstruction of an image from a sequence of angular projections. In 1979, Allan McLeod Cormack and Godfrey Hounsfield were shared Nobel Prize for Physiology or Medicine.

As many modalities, CT scanners are evolved during time. The first CT scanners

were using 360° rotated X-ray tubes. Now, new techniques such as helical scanning developed and CT scans are faster with using lower X-ray beams. These innovations made CT scans safer without occupying much time. Increasing the acquisition speed is not important for patient comfort, but also gives CT scans availability of eliminating the artifacts caused by patient movements and activities inside of the body (such as breathing, digesting). Such advances in CT imaging have a significant impact on the capabilities of volumetric applications. These 3-D imaging and image processing techniques will be explained in "Medical Image Processing" section.

The intensity values of different tissues are defined as Hounsfield values. Hounsfield values have lower-range (between -1000 and 1900 for organic elements in the body) which makes a dynamic range of CT images smaller. This range is enough for distinguishing of hard tissues from soft ones. That is why CT can create high-contrasted images for hard tissues in the body such as bones. On the other hand, it may not create detailed images of soft tissues due to the intensity range limitations. One way to overcome this problem is by using more X-ray beams. However, the dosage and energy of X-ray beams should be taken carefully to make CT safe.

2.2.2 Magnetic Resonance Imaging (MRI)

MRI is a relatively new technique for medical imaging with respect to CT. Developments of MRI contain the study of several pioneers who led to the development of nuclear magnetic resonance (NMR) beginning in the early 20. century. In September 1971, MR imaging was discovered by Paul C. Lauterbur. He created a method to translate spatial knowledge through an NMR signal utilizing magnetic field gradients. The first full-body MRI scanner was developed by John Mallard and his team at the University of Aberdeen, Scotland in 1973. On 28 August 1980. This MRI scanner was used to produce the first clinically usable image of a patient's internal tissue. Peter Mansfield also improved the methods used in the collection and analysis of MR images. Peter Mansfield and Paul C. Lauterbur were given the Nobel Prize in Physiology and Medicine in 2003.

MRI utilizes a strong magnetic field to coordinate the body's hydrogen atoms, which are made mostly of water and therefore contain hydrogen. Unlike CT that uses radiation, MRI uses a powerful and constant magnetic field about 1.5 or 3 Tesla (B_0) to arrange the hydrogen atoms in the body. The main source of hydrogen atoms is water which is the main molecule in the human body. B_0 keeps magnetic moments of hydrogen atoms steady and aligned parallel with the direction of the field. External and smaller but precisely targeted magnetic field M is the second key element in the MRI scanner. M is created by Radiofrequency (RF) coils. This external magnetic field is turned on and off to create pulses during the MRI scan. While M is being applied, the alignment of magnetic moments of hydrogen atoms in a specific small area changes. This is called *excitation*. When M is turned off, the alignment of magnetic moments suddenly come back parallel to B_0. This is called *relaxation*. The time period between excitation and relaxation creates a signal which is received by the antennas on the scanner. These signals are unique for different atoms/molecules. This specification is used to create a contrast for different tissue types in the target region. That is how it is possible to construct images from anywhere in the body without using an X-ray or any other beam. Iterative utilization of this technique in small subareas of the target region creates a 3-dimensional image.

Unlike CT, MRI is capable of high-contrasted and detailed images for soft tissues. The technology behind MRI has the ability of more precise acquisition than CT scans. Therefore it is possible to obtain high contrast between different soft tissues of the body. Despite the time drawback of MRI scans, it is the most preferable way for scanning of soft tissues.

Another advantage of MRI is the capability of obtaining different types of images with different parameters (excitation-relaxation times, the magnitude of magnetic pulses, etc.) via a single scanner. These protocols create different *sequences* during the same scan session. This feature of MRI scans makes possible to obtain high-contrasted images of almost all different types of tissues in the human body. Repetition Time (TR) is the period between consecutive pulse sequences added to the same slice. Time to Echo (TE) is the interval between the transmission of the RF

pulse and the reception of the echo signal. The duration of TE and TR determines the sequence type.

T1-weighted and T2-weighted sequences comprise the most widely used MRI series. Using short TE and TR periods, T1-weighted images are produced. The contrast and luminance of the image is primarily determined by the tissue characteristics of T1. T1 images can be considered as a proton energy diagram within the body's fatty tissues. Conversely, by using longer TE and TR times, the T2-weighted images are generated. T2 photos are a proton-energy diagram of the body's fatty *and* water-based tissues. Fatty tissue can be differentiated from water-based tissue by comparing with the T1 images. For example a tissue that is bright on the T2-weighted images but dark on the T1-weighted images is fluid-based tissue. To reflect the dominant image contrast of spin-density, double echo pulses are preferred in T2-weighed images as shown in Figure 2.7.

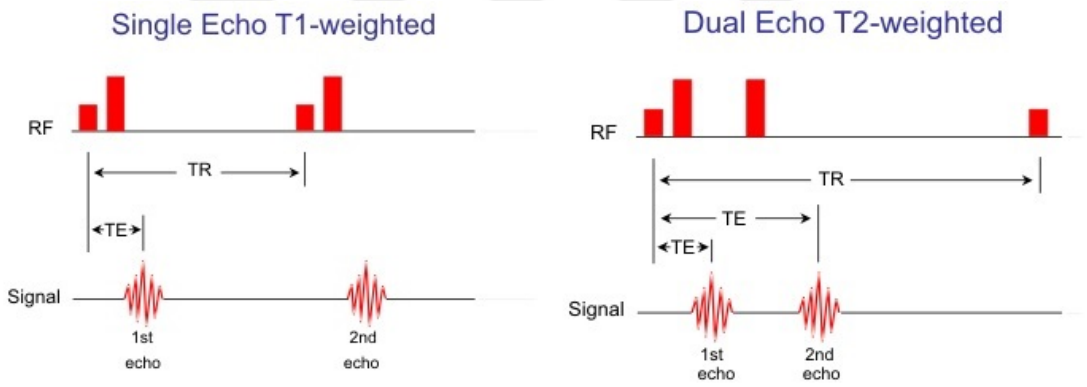

Figure 2.7 Comparison of pulse sequences for T1 and T2 images (Hesselink, 2020)

2.3 Clinical Usage of Acquired Medical Images

After imaging from any modality, the use of these images in the clinic begins. At this stage, it is necessary to study and analyze the images. This step is performed by experts, specifically radiologists. Radiologists need to display the images from many aspects. They need to filter images, change displaying properties and obtain 3-dimensional data from 2-dimensional slices if the region of interest on the human body is a volumetric area such as the abdomen. With the help of some tools (software),

radiologists can obtain a 3-dimensional volume image without doing extra work.

After sorting out images, experts may need additional operations on the image to see suppressed information from the images. Images are handled to maximize the most critical aspects in order to highlight details of importance prior to the show. At this stage there may be single or multiple cascade operations (Selvi et al., 2015) depending on the desired information and the modality. All operations at this step are examined under "Medical Image Processing" section. All operations under medical image processing must be handled very carefully. Otherwise, the primary information inside of the data may be overlooked or there may be pseudo-information that can mislead experts during analysis. The details of medical image processing and commonly used techniques are explained in the following section.

2.3.1 Medical Image Processing

Medical image processing techniques are adapted methods from digital image processing for clinical aspects. Methods in digital image processing can be used to derive essential and important information from the image required to carry out measurements or other analysis. The most common analyses are planning treatments, surgeries, and other surgical operations.

There are many advanced algorithms in digital image processing branch. However, it is not possible to adapt all of them to medical image processing by reason of protecting original information in the image data. Therefore the techniques in medical image processing tools are limited. In the following sections of this chapter, the most common techniques (from simple to advanced) are briefly explained.

2.3.1.1 Windowing, Filtering, and Multi Planar Reconstruction

Windowing: Nowadays all of the screens attached to the computers, phones, TVs, etc. have a digital panel. Almost all of these panels have an 8-bit color range. In other

words, they are capable of displaying $2^8 = 256$ intensity levels for each color channel. On the other hand, images acquired from some modalities such as X-ray, CT, MRI have more than 8-bit data. This means that it is not possible to show all the information inside of the images at the same time. Although there are some special displays (such as 16-bit panels), sometimes experts need to suppress irrelevant information in the image to focus on the region of interest. Windowing is an easier approach to handle this problem. Windowing gives the opportunity to the users to emphasize the preferred region in image histogram while suppressing the unwanted areas. Radiologists usually adjusts *window level* and *window width* values with mouse shortcuts to use windowing effectively.

Filtering: After adjusting display parameters with windowing, experts may need to study out detailed information that is not seen clearly. Here the suppressed information can be strengthened with filtering techniques. In signal processing, a filter is a system or mechanism that eliminates any unnecessary components from the signal. The filter can be the total or partial removal of any component of the signal. Filters can be defined on spatial or frequency domains. Correlations can be eliminated for some elements and not for others. For example, windowing can be classified as a basic filtering method. However, filters may have complicated designs.

Multi Planar Reconstruction (MPR): As explained before, modalities that are capable of imaging a volume of interest save the data as a 2-dimensional image series. These images can be considered as slices and stacked together to create a volumetric image data. Although the original images are acquired from only one direction (axial, coronal, or sagittal), it is possible to examine 3-dimensional data from different angles of interest. This procedure is called Multi Planar Reconstruction (MPR). MPR gives users the opportunity of displaying non-acquired orthogonal orientations like acquired ones. MPR reorders voxels to create images of the desired view.

MPR is one of the most preferred image processing methods for daily clinical usage. Experts have the capability of examining volumetric data from infinite possible angles. An arbitrary plane can be selected by the expert at some oblique angle. All medical

image processing techniques are also available for use on reconstructed MPR images.

2.3.1.2 Image Segmentation

Image segmentation is the method of splitting an image into diverse and meaningful segments. In other words, segmentation means masking a group of pixels/voxels from the whole image to focus on a region of interest. In general, these pixels/voxels belong to a structure such as an object or an organ. Hence examination of the target object is getting easier with segmentation operation. Experts can analyze the target structure more in-depth. They can make measurements on it such as width/length measurements or volume calculation. They can also use the segmented object for visualization.

Suppose that the input is a 3D image $A = \{a_{\{i,j,k\}}\}$, where $i = 1, ..., R$, $j = 1, ..., C$ and $k = 1, ..., K$, where R is the number of rows of pixels, C is the number of columns of pixels, and K is the number of slices in the 3D volume. We introduce probability map $P = \{p_{\{i,j,k\}}\}$, where the indices vary in the same intervals, and $p_{\{i,j,k\}} \in [0, 1]$ is the probability that a voxel with coordinates (i, j, k) belongs to class "foreground". Denote the ground truth as $G = \{g_{\{i,j,k\}}\}$, where $g_{\{i,j,k\}} \in \{0, 1\}$ is zero if the voxel is labelled as background by the expert radiologist or 1 if the voxel is labelled as foreground.

Segmentation is the most used and studied medical image processing operation (Maier-Hein et al., 2018; Guo & Ashour, 2019). It is possible to classify segmentation methods in digital image processing into different categories. One of the most common categorizations is defining them as manual, semi-automatic, and fully-automatic.

Manual segmentation is segmenting the target structure(s) by hand with a preliminary software (Starmans et al., 2020). Manual segmentation by experts guarantees the precision of the segmentation. Although it is the safest way to handle the segmentation process, it may be tremendously time-consuming depending on the region of interest and the modality. For example, an abdomen CT scan of an adult patient with 1mm slice thickness can produce more than 200 2-D images with

512x512 resolution. Segmenting the liver from these slices can take several hours to complete. Another drawback of manual segmentation is that the accuracy and precision of the results are strictly dependent on the conditions of the operator. These conditions can be physical conditions of the environment or experience of the operator. Manual segmentation is generally handled by drawing borders of the targeted structure. After that, the area inside of the borders is filled by image processing methods to obtain a mask of segment(s).

Semi-automatic image segmentation approaches also need user input (Fischer et al., 2010). However, there are some tools available to make segmentation reliable and fast. The most common method is selecting the start point(s) of the segmentation algorithm manually. After that, the proposed algorithm starts segmenting target objects with defined methods such as region growing and fast marching. After the segmentation finished, the outputs may need some post-processing operations such as removing over-segmented areas or completing miss-segmented regions. Despite the need for user supervision, semi-automatic methods may reduce the effort and time significantly.

Fully-automatic segmentation methods do not need any user supervision or interaction (Moghbel et al., 2018). However, their generalization capabilities are limited to pre-defined structures. The main reason for this drawback is that fully-automatic methods use multiple cascaded methods to construct segmentation masks. Before the last decade, generalization abilities, success, and performance of the fully-automatic methods were behind the semi-automatic algorithms. However, along with major developments in the machine learning area, this situation is changing in favor of fully-automatic methods (Kavur et al., 2020b). This development will be discussed in the following sections/chapters of the thesis.

2.4 Metrics for Evaluation of Segmentation Accuracy

2.4.1 Sørensen–Dice coefficient (DICE)

DICE is an overlapping-based metric. Assume that V_{Seg} represents the voxels in a segmentation result, V_{Ref} represents the voxels in the ground truth. Both are 3D binary segmentation mask image. DICE generates values between $[0-1]$ scale (the larger, the better). DICE coefficient is calculated as

$$DICE = \frac{2 \mid V_{Seg} \cap V_{Ref} \mid}{\mid V_{Seg} \mid + \mid V_{Ref} \mid} \tag{2.1}$$

where $\mid . \mid$ denotes cardinality:

$$\mid V \mid = \sum_{i=1}^{N} v_i \tag{2.2}$$

where V is 3D volumetric binary object, N is total number of voxels, and $v_i \in \{0, 1\}$

2.4.2 Volumetric Overlap Error (VOE)

Volumetric Overlap Error uses the intersection of two objects, reference V_{Ref} and segmentation V_{Seg}. The volume in the intersected zone is divided by the volume of union:

$$VOE = \frac{\mid V_{Seg} \cap V_{Ref} \mid}{\mid V_{Seg} \cup V_{Ref} \mid} \times 100 \tag{2.3}$$

Here $V_{Seg} \cap V_{Ref}$ and $V_{Seg} \cup V_{Ref}$ symbolize the number of voxels in the intersection and union of the segmented and reference (Ground Truth) objects. An ideal segmentation gets value of 100 while zero grade is calculated if there isn't any intersection between two objects.

2.4.3 Relative absolute volume difference (RAVD)

The whole volume difference between the segmentation and ground truth is divided by the whole volume of the reference object. Average of the division is converted to percent value (Eq.2.4).

$$RAVD = \frac{abs(\mid V_{Seg} \mid - \mid V_{Ref} \mid)}{\mid V_{Ref} \mid} \times 100 \tag{2.4}$$

An whole accurate segmentation gets RAVD value of 0. Higher errors get higher RAVD values. The disadvantage of this metric is that any segmentation with same volume with reference may get 0 because the metric does not use any topological value. That is why many different error calculation metrics should be preferred.

2.4.4 Average symmetric surface distance (ASSD)

ASSD uses symmetric surface distance (SSD) to compare two volumes. Let a distance measure for a voxel x from a set of voxels A to be defined as:

$$d(x,A) = \min_{y \in A} d(x,y) \tag{2.5}$$

where $d(x,y)$ is the Euclidean distance of the voxels incorporating the real spatial resolution of the image. To calculate symmetric surface distances border voxels that are the voxels at the shell of the 3D object are used. First, the border voxels of the segmented object (B_{seg}) and reference (B_{ref}) are determined. Then for each voxel in these sets, the closest voxel in the other set is determined as shown in Figure 2.8. All these distances are stored, for all border voxels from both reference and segmentation.

The average of all the distances, $d(x,y)$, gives the averages symmetric absolute surface distance as shown in Eq.2.6.

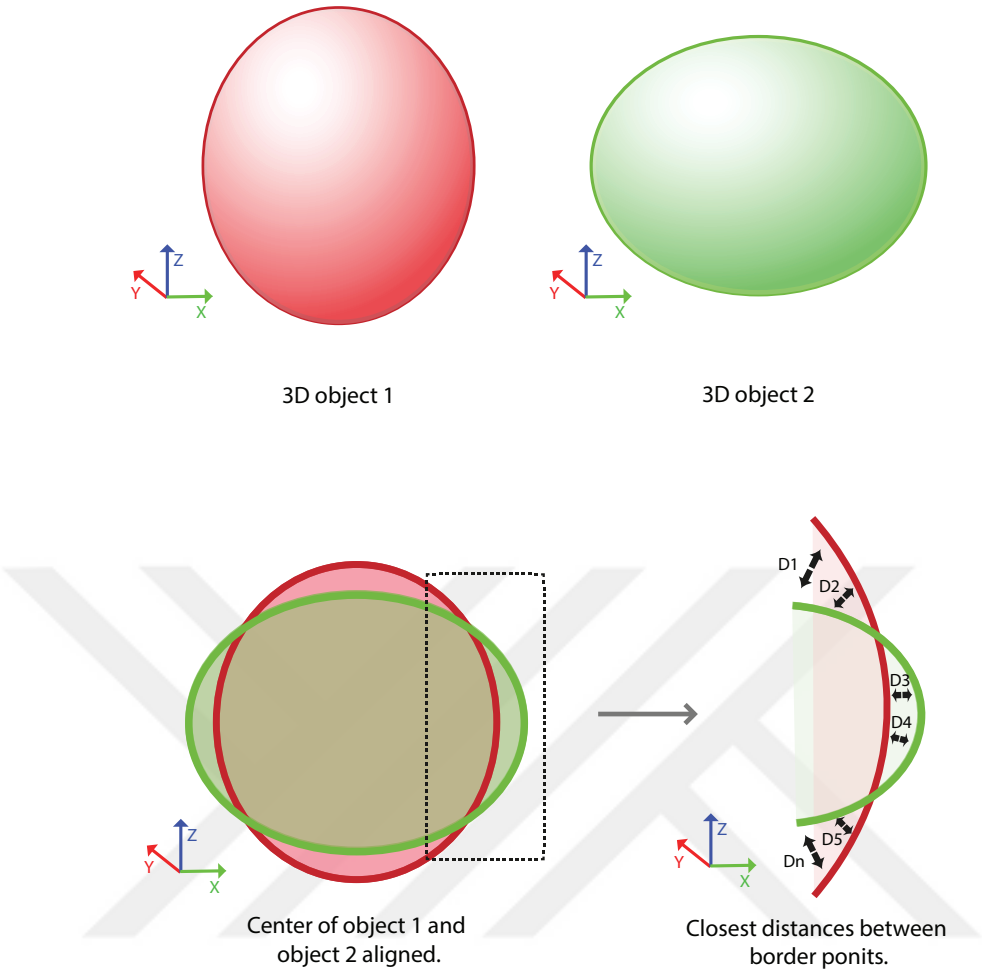

Figure 2.8 Illustration of symmetric surface (Hausdorff) distance

$$ASSD = \frac{1}{|B_{seg}| + |B_{ref}|} \times \left(\sum_{x \in B_{seg}} d(x, B_{ref}) + \sum_{y \in B_{ref}} d(y, B_{seg}) \right) \tag{2.6}$$

ASSD is 0 for a perfect segmentation. There is no upper limit.

2.4.5 Root mean square symmetric surface distance (RMSSD)

RMSSD is identical to ASSD but retains the square distances among points on edge of the two objects. After averaging the squared values, the root is extracted and gives the symmetric RMS surface distance:

$$RMSSD = \sqrt{\frac{1}{|B_{seg}| + |B_{ref}|}} \times \sqrt{\sum_{x \in B_{seg}} d^2(x, B_{ref}) + \sum_{y \in B_{ref}} d^2(y, B_{seg})} \tag{2.7}$$

This value is 0 for a perfect segmentation without any highest limit.

2.4.6 Maximum symmetric surface distance (MSSD)

Here only the highest voxel distance is used instead of applying average or RMS. In some studies, MSSD is called Hausdorff Distance.

$$MSSD = max\left(\min_{x \in B_{ref}, y \in B_{seg}} d(x, y)\right) \tag{2.8}$$

This value is 0 for a perfect segmentation without any highest limit.

2.5 Organ Segmentation Methods for Abdomen Imaging

In previous sections, segmentation methods in medical image processing were briefly explained. Also, they were categorized as traditional way (manual, semi-automatic, and fully-automatic). However, it is necessary to separate image segmentation methods into different classes now. In the last decade, the advancements in machine learning, specifically deep learning are changing the fundamentals of image segmentation solutions as well as many other fields. It is possible to be said that developments in the medical image segmentation field are dominated by Deep Learning based methods. Therefore new categories were preferred. These are *Image Processing-based Segmentation Methods* and *Deep Learning-based Segmentation Methods*.

2.5.1 Image Processing-based Segmentation Methods

These segmentation methods depend on developed algorithms in the computer vision field for image processing. These algorithms can be used alone or together to segment the targeted structure. The common factor of these methods is that they are using one or multiple features to perform segmentation. These features can be ready for use or they are extracted by a series of operations. The overview of these algorithms is well explained in Song & Yan (2017). The summary of them is presented in Figure 2.9.

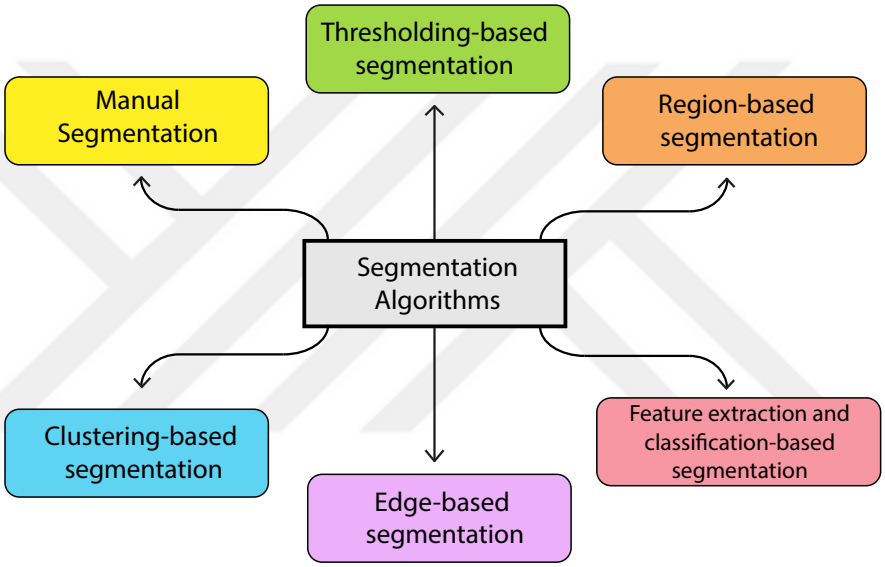

Figure 2.9 Commonly used image processing-based segmentation methods

2.5.1.1 Classifiers

Another approach for image segmentation is classifying the structures inside of the image. This classification uses algorithms from the pattern recognition field. In pattern recognition, the information which is used for classification is called features. These features can be pre-defined ones in the image or they are "extracted" by feature extraction strategies. Feature extraction was a widely studied field before the machine learning era. The features can be extracted not only from the spatial domain but also from different domains such as frequency domain. The feature spaces may be one or multiple dimensional (Alpaydin, 2014).

After obtaining features, the main part of the classifiers is performed. The classification algorithms try to find the best and the optimal way to "separate" the features in order to classify the data. The most common way is searching similarity between a group of objects. That is how they can divide feature space into meaningful parts and perform classification procedures.

Medical image segmentation classifiers are commonly supervised methods. That means they need training data to adjust classifier parameters. Also, they need an unseen data, called test data to evaluate their performance.

2.5.1.2 Clustering

Clustering algorithms work very similarly to classification algorithms but the main difference is that they do not need training steps. They are looking for a similarity between objects. That is why clustering methods can be defined as fully-automatic methods.

K-means algorithm is one of the most commonly used clustering algorithms (Pelleg & Moore, 1999). K-means tries to separate feature space into "K" parts. The only user-defined parameter is the number of the classes, "K". Classes are defined by elements with the closest mean in the same cluster. The generalized version of K-means is the fuzzy C-means algorithm which uses fuzzy set theory. Another popular method for clustering is the Expectation-Maximization (EM) algorithm. EM is an iterative method. Firstly, EM tries to "expect" the clusters. Then, it computes the posterior probabilities. Finally, it modifies clusters to "maximize" posterior probabilities.

2.5.1.3 Artificial Neural Networks

Although Artificial Neural Networks (ANNs) are a subbranch of the machine learning field (Goodfellow et al., 2016), their utilization in the image processing field

is as classifiers for feature space. Hence they are explained under the image processing-based segmentation methods section.

The design of ANNs are based on the implementation of real neurons with a very simple mathematical model. There are several artificial neurons inside of the ANN system. These neurons are fully connected inside of the "layers". The learning process is handled by a training session that needs training data. Each connection in the ANN has a unique weight. These weights are the places where the information is "stored" in the trained ANN. After the training session, trained ANNs can be used to classify unseen data. ANNs' capability of storage depends on their size, more clearly the number of neurons/connections. More connections hold more data but they increase time and memory needs. The design and implementation of ANNs from scratch can be complicated depending on their design. However, there are some useful publicly available tools/libraries to use them on a specific problem in an easy way.

2.5.2 Deep Learning-based Methods

Deep learning (DL) is part of a wider class of methods in machine learning (Goodfellow et al., 2016). The fundamentals of DL are based on ANNs. Therefore they need to be trained with dedicated data. After that, they are ready to use on unseen data.

DL has different frameworks such as deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). CNNs and RNNs can be considered as subclasses of DNNs. Applications of DL architectures are widely used in medical image analysis as well as computer vision, voice recognition, audio processing, natural language processing, machine translation, bioinformatics, etc. DL has gained huge popularity in literature by surpassing alternative methods in different fields, notably in the field of medical image analysis (Greenspan et al., 2016; Shin et al., 2016).

DNNs use numerous layers to derive features from the raw data. The key difference between ANNs and DNNs is the number of layers between input and output. The number of hidden layers describes "depth" of the model. Since there is no absolute definition of the layer size to separate ANNs and DNNs, it is not possible to make a definition. However, an ANN with three or more layers is called DNN (Albawi et al., 2017). A simple illustration of ANN and DNN is presented in Figure 2.10.

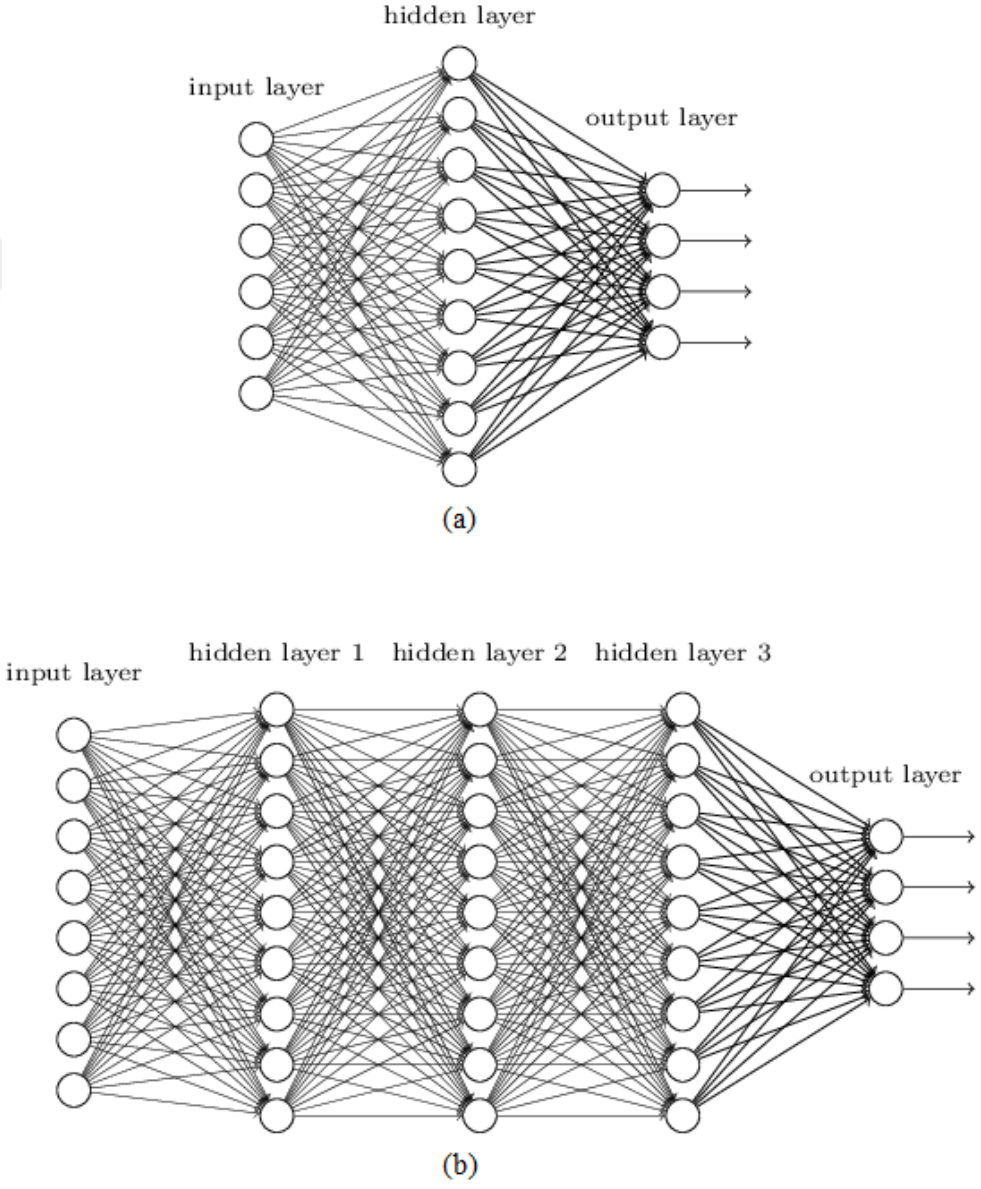

Figure 2.10 Architecture of an (a) ANN and (b) DNN (Nielsen, 2019)

DNNs have the capability of storing huge information inside of the structure. However such amount of information does not always mean successful classification.

Overfitting is the biggest problem in DNN and Dnn related studies. Simply, overfitting is defined as too much learning from the train data. In other words, the system perfectly adjusts itself to the train data. This causes losing the generalization capability of the system. As a result, the system fails to perform on a new, unseen data. Overfitting makes impossible for utilization of the proposed DL system on real-life data even it performs very successfully on train data.

RNNs' design is based on ANNs. However, there is an important difference. ANNs can only accept inputs with a fixed size of vectors while RNNs do not have such a limitation. RNNs are configured to take a set of inputs with no fixed size limit. Besides, RNNs uses previous outputs of the system to be used as inputs. Therefore RNNs have a memory of past decisions. The decision of RNNs is affected by the previously learned information in the past stages. That means RNNs can learn not only during training sessions but also even generating outputs for test data. The mainly used fields of RNNs are in speech recognition, handwriting recognition, and natural language processing. On the other hand, it is not preferred for image segmentation applications.

CNNs can be defined as a modified version of DNNs to reduce overfitting. A DNN has fully connected neurons. That means every neuron in a single layer is connected to all neurons in the next layer. This design is one of the important factors that causes overfitting. CNNs also have additions to the loss function to ensure regularization. Regularization adds additional conditions (or information) to the system to reduce overfitting.

Another key difference of CNNs is the usage of convolution operation instead of matrix multiplication between layers. Using convolution gives CNNs the opportunity of using hierarchical pattern information in data. Besides, it makes possible to store complex patterns in single ones which also reduces overfitting.

Working principle of a typical CNN in an image segmentation application can be explained in the following stages:

1. Tensors are generated by input image

2. Convolutional layers convolve the tensors
3. The results are passed to activation function
4. Pooling applied the results and passes its result to the next layer.
5. All steps are repeated according to the design of CNN
6. The final tensor is flattened and classified by a fully connected layer.
7. The result of the final layer defines the output of the CNN.

Tensors are multi-dimensional matrices which are constructed by multiple images. The size of a typical tensor is created by (number of images) x (image width) x (image height) x (image depth).

In convolutional layers, the tensors are convolved by *kernels*. Kernels store the learned information. In other words, the memory of the CNN is inside of their kernels. During training progress, a cost function is calculated via reference data in the train data. With the help of the values of the cost function, kernels are updated to perform better segmentation in the next epoch.

The results are transformed into the activation layer. The choice of activation function depends on the application of CNN. For example Rectified Linear Unit (ReLU) are widely used in computer vision applications (such as segmentation) due to elimination of less relevant signals. Softmax is another popular activation function that gives probability distribution because it maps each output in such a way that the total sum is 1. Therefore it is often used in the final layer of a neural network-based classifiers. Sigmoid functions are used for the logistic regression problems. After the activation function, pooling layers are used to downsize the tensors. The shrinking of the information into smaller tensor reduces overfitting.

As mentioned before, the proposed image segmentation methods are dominated by DL-based methods in the last decade. Hence many various system designs have different advantages and disadvantages to each other. In the following sections, the

most widely-used and successful DL-based solutions in the medical image segmentation field will be explained.

2.5.2.1 U-Net

One of the first CNN designed for the segmentation of biomedical images is U-Net. The first theoretical architecture of U-Net was created by Long and Shelhamer (Long et al., 2015). In 2015, the first usage of U-Net in the biomedical imaging field is the segmentation of neuronal patterns in electron microscopic slices (Ronneberger et al., 2015). U-Net is the most popular CNN design in medical image processing due to its capability of operating with fewer training data than other models without losing segmentation precision.

The name 'U-Net' comes from the shape of the architecture (Figure 2.11). The network consists of two main parts, which are compression and decompression. The layers of the compression part make the right arm of the 'U' shape. They perform repeated convolution operations. ReLU and a max-pooling operation follow each convolution. That is how spatial information is shrinking while preserving the most important information. The decompression part, forming the right arm of the 'U' shape, collects feature maps from each stage of the compression part. Upsampling is applied at each stage of decompression to match the final resolution. Segmentation is handled by a final layer. Due to the huge popularity of U-Net, there are variously modified versions of it in literature.

2.5.2.2 DeepMedic

DeepMedic uses a combination of 3D CNN architecture with a fully connected Random Field (Kamnitsas et al., 2016). Deepmedic was initially built to segment brain lesions from MRI scans. It won BraTS 2017 and ISLES 2015 challenges (Kamnitsas et al., 2017).

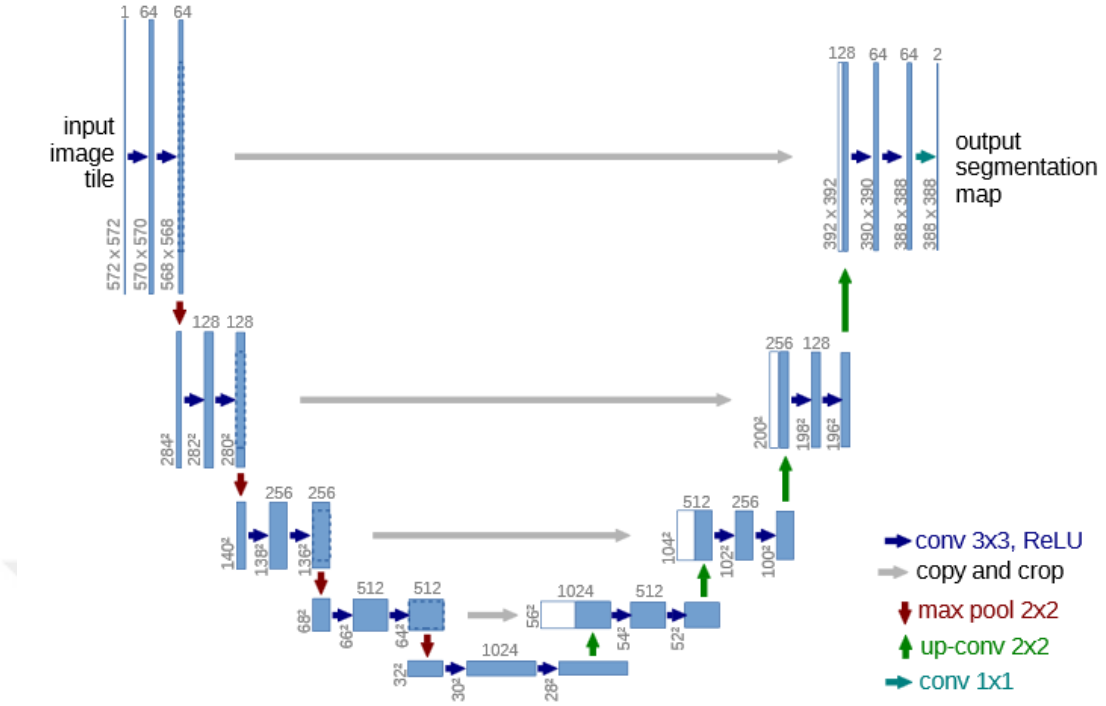

Figure 2.11 Architecture of U-Net (Ronneberger et al., 2015). Multi-channel feature maps (with channel numbers on the top) are represented by blue boxes while copied feature map are represented by whiteboxes

To use multi-scaling characteristics for organ segmentation, DeepMedic uses a dual 3D CNN pathway with 11 layers shown in Figure 2.12. DeepMedic comprises the simultaneous training of one network that uses the full-resolution image and another network on the down-sampled version of the image. The feature maps coming from the two paths are concatenated at the final stage. Deepmedic relies on a 3D fully-connected Conditional Random Field (CRF) for post-processing. This approach is used for clearing false positive background voxels wrongly labeled as foreground. Deepmedic also uses cross-entropy as a loss function (L_{CE}) in Eq 2.9.

$$L_{ce} = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{K} g_{\{i,j,k\}} \log\left(p_{\{i,j,k\}}\right) - \left(1 - g_{\{i,j,k\}}\right)\log\left(1 - p_{\{i,j,k\}}\right) \tag{2.9}$$

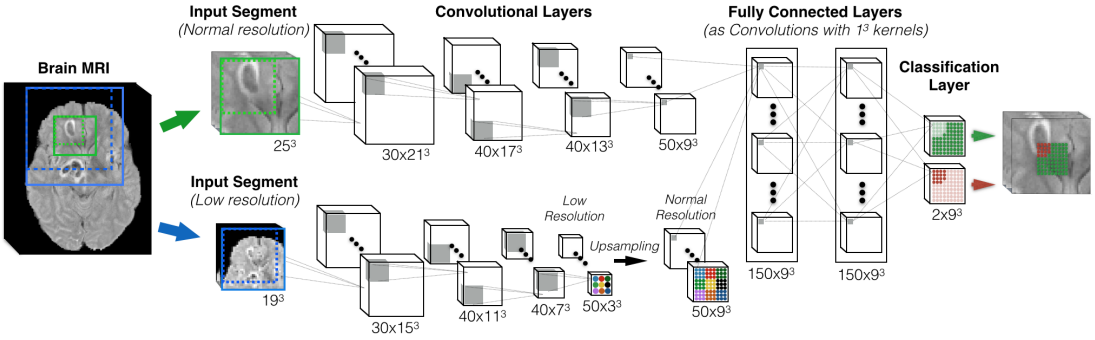

Figure 2.12 Architecture of DeepMedic (Kamnitsas et al., 2017). Multi-channel feature maps in normal and low resolution channels are represented by boxes

2.5.2.3 V-Net

V-Net has been designed for volumetric segmentation of the prostate from MR scans (Milletari et al., 2016). Like U-Net, the name of V-Net is coming from its architecture that has a V-shape (Figure 2.13). The design of V-Net was inferred by U-Net with slight architectural differences.

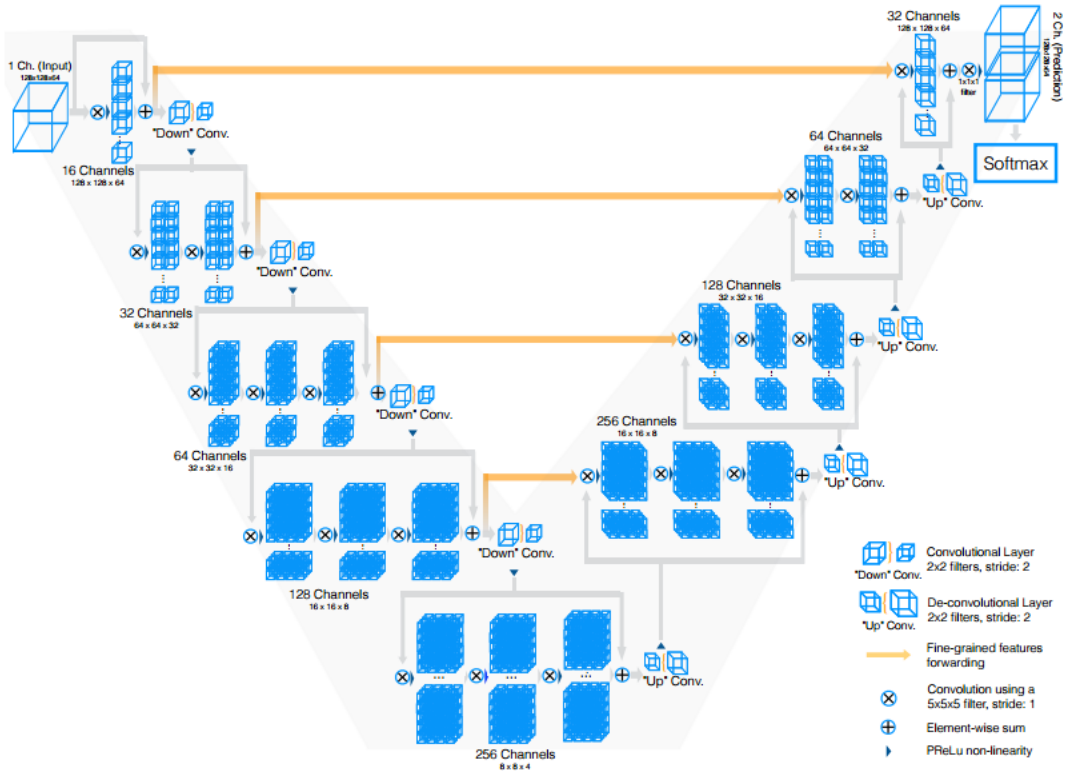

Figure 2.13 Architecture of V-Net (Milletari et al., 2016)

The left part of V-Net handles compression for extracting features. At each iteration, the resolution is reduced by a pre-determined stride. The right part of V-Net handles the decompression of the feature maps until reaching the original resolution of the input image. V-Net segments the target object from the volumetric data using directly 3D convolutions instead of 2D convolutions for each slice.

The loss function based on the DICE overlap coefficient and it is created for medical image segmentation. If the target organ has a relatively small volume against the whole volume, the training process can get stuck in a local minimum of the standard DICE loss function. V-Net uses a modified DICE loss which is defined by a gradient of the DICE score with respect to the predicted voxels. This loss metric amplifies the performance of the system according to its creators. The DICE loss formula is presented in Eq. 2.10 taken from reference article (Milletari et al., 2016).

$$\frac{\partial D}{\partial p_j} = 2\left[\frac{g_j\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right) - 2p_j\left(\sum_i^N p_i g_i\right)}{\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right)^2}\right] \tag{2.10}$$

where $N = R \times C \times K$, p is the probability map of the network, g is the ground truth.

2.5.2.4 Dense V-Networks

Dense V-Networks have been designed for the automatic segmentation of abdominal organs from CT image series (Gibson et al., 2018). The most distinctive feature of Dense V-Networks is three dense feature blocks at each encoding stage.

Dense V-Networks use a fully convolutional neural network architecture (Figure 2.14). The convolution process contains 3D convolution, batch normalization, and rectified linear unit. There are three dense feature blocks with different resolutions. At each stage of dense feature blocks, strided convolution is applied to compute feature maps. As a result, there are three future maps for different resolutions. In order to decrease the number of feature maps, convolution is applied for each resolution. At the final stage, future maps that have lower resolutions are upsampled and all maps are added.

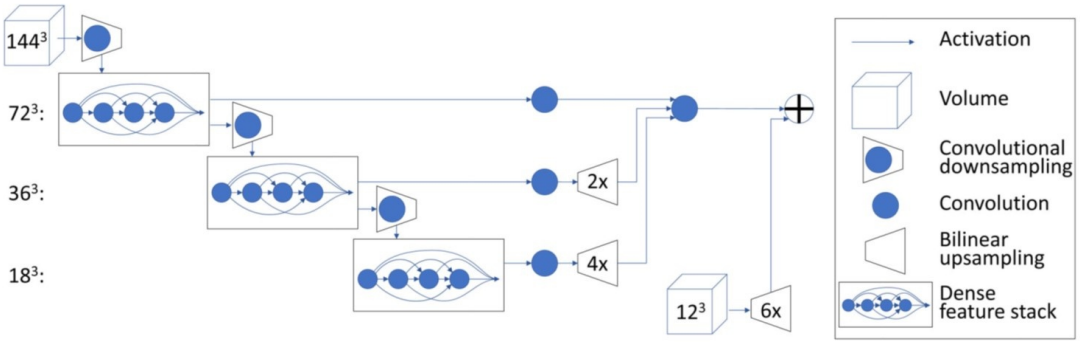

Figure 2.14 Architecture of Dense V-Networks (Gibson et al., 2018)

Dense V-Networks uses probabilistic DICE score ($p\,\mathrm{DICE}_l$) (Eq. 2.11) for calculating the loss function ($\mathrm{loss}(p,)$). The function is the weighted sum of L2 regularization (also called least-squares error (LSE)) shown in Eq. 2.12.

$$p\,\mathrm{DICE}(p,g) = \overline{\left(\frac{\min(p,0.9)\cdot g}{\|g\|_2 + \|\min(p,0.9)\|_2}\right)} \tag{2.11}$$

where p is probability map of segmentation, g is the ground truth.

$$\mathrm{loss}(p) = \sum_{\forall W} \frac{\overline{w^2}}{40} - p\,\mathrm{DICE}(p,g)) \tag{2.12}$$

where $w \in W$ are kernel values. Note: the Eq. 2.11 and Eq. 2.12 were taken from reference article (Gibson et al., 2018).

2.5.3 Classifier Ensembles (Fusion)

Classifier ensembles (or fusion of classifiers) are an alternative way to improve the accuracy of any system that has a classification step. It is an efficient way to produce improved outcomes by integrating multiple results from different models to obtain a consistent final outcome (Kuncheva, 2014; Oza & Tumer, 2008; Rokach, 2010; Zhang & Ma, 2012). Classifier ensembles have been used most respected grand challenges such as Imagenet (Deng et al., 2009) and Kaggle (Google Inc, 2020) that the winner algorithms use ensembles of deep learning architectures (Huang et al.,

2017). Object detection (Razinkov et al., 2018), aerial scene classification (Dede et al., 2019), video classification (Zheng et al., 2019), and diagnosis and prediction in commercial systems (Ma & Chu, 2019; Zhang et al., 2017) are other popular topics where ensembles were used.

The integration of ensemble methods in the proposed segmentation solutions can be handled in different ways. A reasonable number of ensemble members are the most commonly preferred approach. Another approach is using multiple DMs as ensemble inputs. The training of the DMs can be done individually (Kamnitsas et al., 2017; Warfield et al., 2004) or at the same time (Ma & Chu, 2019; Zheng et al., 2019). Another ensemble method for DMs is using the outputs of the same DM with different training stages (Dede et al., 2019). The members in the ensemble design can be different/individual methods. These kinds of ensemble strategies are called a heterogeneous ensemble. On the other hand, using the same method with different parameters and/or training strategies is called a homogeneous ensemble. For example, separating training data for K-fold cross-validation, training multiple DMs with the same model on a different portion of the data, and the ensemble of their results can be used as a homogeneous ensemble. Another approach is using a combination of multiple DM results obtained by stopping in different local minima during training. This strategy is called "snapshot ensembling" (Huang et al., 2017; Dede et al., 2019). The undertrained condition of DMs eliminates problems coming from overfitting to data. Here, diversity helps to create superior results than a typical train-test strategy of DMs.

Common ensemble methods can be explained as majority vote (Ortiz et al., 2016), average (Kamnitsas et al., 2018; Maji et al., 2016; Codella et al., 2017), product and more (Warfield et al., 2004; Ju et al., 2018). Besides basic methods, there are some advantaged ensemble rules such as STAPLE (Warfield et al., 2004) that uses the expectation-maximization method. Another category of ensemble rules is trained combination rules such as stacked generalization, Bayes models, and "super learner" (Ju et al., 2018).

Studies, where the same data are used by many different systems, provide the necessary information for the analysis of ensemble potentials. At the end of public challenges, the capabilities of DM ensembles in medical imaging are presented (Prevedello et al., 2019). In this kind of analysis, all or just top of the proposed algorithms are combined through basic ensemble methods (such as majority voting). Usually the final result outperforms individual results (Menze et al., 2015; Jimenez-del-Toro et al., 2016; Bilic et al., 2019; Kamnitsas et al., 2018; Kavur et al., 2020b). However such results can be misleading if the sum of the challenge results is affected by the "peeking" problem that is explained in Section 3.1.1.

The success potential of classifier ensembles has led these methods to be preferred in the field of medical image processing as well as in many other fields. Several ensembles have been particularly proposed for the segmentation of medical images (Kamnitsas et al., 2017). Especially the ensemble of multiple DMs is getting more popular (Ju et al., 2018; Codella et al., 2017). The article of Liver Tumour Segmentation Benchmark (LiTS) challenge (Bilic et al., 2019), reveals some probable need for ensemble strategy to create a generalized segmenter for medical images. In this article, the winner models were analyzed and the following important facts were emphasized:

1. Although many successful DM designs were proposed in the LiTS17 challenge, it is not quite possible to recommend certain DM design as well as its parameters, training strategies, modifications, and so on. The proposed models do not have strict, proven guidelines. In other words, their success is coming from rough ideas. The main reason for this is that researching possible choices for each task requires, long training times and computational power.

2. Another problem is only a few of the proposed DMs have a 3D architecture for working with volumetric image series. Other DMs process volumetric images slice-by-slice which increases computational costs. On the other hand, 3D models can have many more parameters than 2D ones.

3. In general, an ensemble of multiple DMs outperforms individual DMs. However,

the time costs of designing and fine-tuning individual DMs is one of the biggest drawbacks of preferring ensemble approaches.

Due to the notices in this study (Bilic et al., 2019), the ensemble methods that seemed simple should actually be handled very carefully. The computational cost and the time for tuning individual DMs can be reduced with smart design choices. For example, keeping ensemble methods and usage of individual DMs as simple as possible is a preferable way. Also, the advantages of ensembles against individual models must be analyzed carefully before proposing an ensemble method.

CHAPTER THREE
GRAND CHALLENGES IN THE BIOMEDICAL IMAGE ANALYSIS

3.1 Introduction

Many sophisticated methods are being proposed for organ segmentation problems due to the necessity of robust segmenters in clinical usage. Besides, there has been tremendous progress in deep learning (DL) in many fields of science in recent years. These developments in DL studies are continuously adopted to organ segmentation solutions as expected. Thanks to the capabilities provided by DL methods, the theoretical successes of DL-based segmenters increase more than ever.

On the other hand, despite the success of previous algorithms was surpassed by DL, it is difficult to analyze the effects of DL parameters on the performance without making comprehensive evaluations. For this reason, comparative analyses have become an essential mechanism to explain systems better. To find the most successful among the many recommended DMs, it is necessary to push the methods to their limits. New data with new tasks are useful methods to handle this mission.

The methods for comparing the efficiency of various segmentation techniques in medical imaging are very crucial in clinically important tasks (Ayache & Duncan, 2016). Providing a new dataset to create a new benchmark platform has gained significance in the analysis of proposed algorithms (Simpson et al., 2019). These benchmark platforms, namely grand challenges, report the results in a methodical way (Kozubek, 2016). Therefore grand challenges in biomedical image analysis are getting more important than ever. For example, there are specific websites for hosting grand challenges such as grand-challenge.org (van Ginneken & Kerkstra, 2015). This site currently includes more than 200 challenges in biomedical image analysis.

While the impact factor of grand challenges is increasing, the design of the challenges has a major effect on the true potential of such contests (Reinke et al., 2018b). For example, the decisions on ground truth generation, evaluation metrics,

ranking, criteria, and construction of the datasets must be handled very carefully in order to make challenge results significant. In the literature, these issues have been studied in detailed reviews (Maier-Hein et al., 2018; Reinke et al., 2018a) to enhance the quality of challenges. New challenges are being designed to solve the flaws of the current ones and provide new data to the field of interest. The majority of challenges are one-time events. In addition, some of them are continuously updated (Menze et al., 2015) while some of them are repeated after some time (Staal et al., 2004).

3.1.1 Peeking problem

In addition to problems in grand challenges, there is an underestimated problem called 'peeking'. Peeking is optimizing the proposed system by fine-tuning parameters and modifying designs on the 'test data'. Peeking is done by making several iterative evaluations on the test data showing in Figure 3.1. Here, there is even no need for direct access to ground truths of test data. The results of each submission are used to optimize the system.

After this stage, the test data becomes a kind of validation data indirectly. Peeking

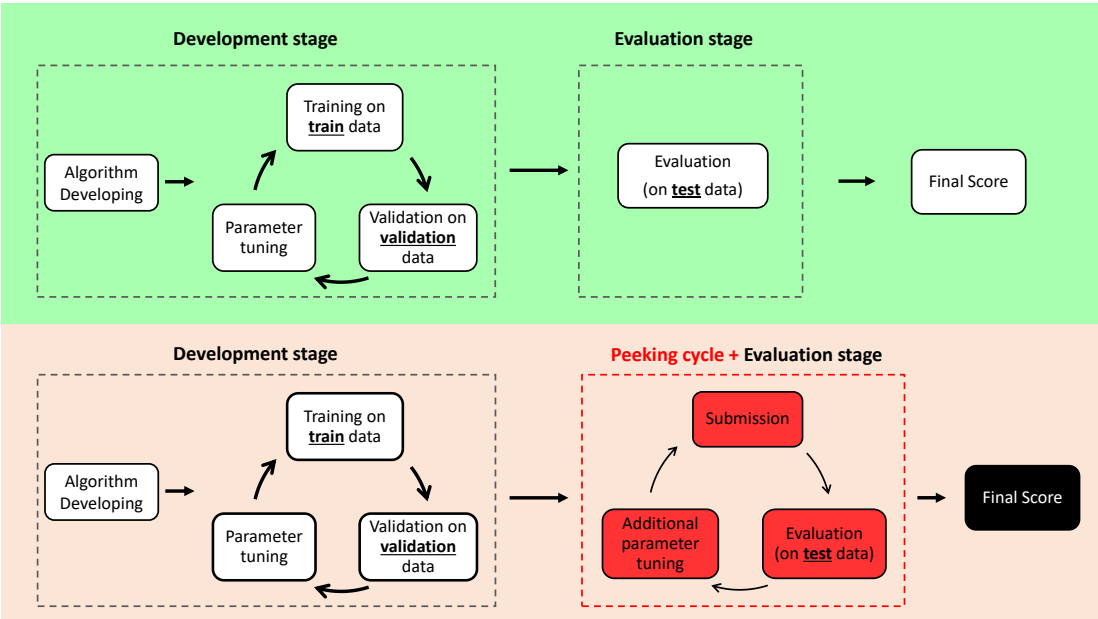

Figure 3.1 Comparison of a proper study (in green) and peeking attempts (in red)

can mask the true capabilities of proposed systems especially DMs. In other words, a DM which gives very accurate segmentation results on specific data/challenge may be ineffective for other datasets. Considering the main reason for creating such segmenters is to solve problems in radiology, the promising results in specific datasets may not mean anything for real-world utilization. Surprisingly peeking is an underestimated problem in many challenges. It is one of the biggest reasons for the obstruction between academic studies in machine learning and their real-world implementations.

In the following sections of this chapter, the overview of previous abdomen related challenges are introduced in 3.2 Related Work. After that, the two organized challenges (national and international), are presented in detail in Sections 3.3 and 3.4.

3.2 Related Work

It has been revealed through a detailed literature review that current challenges with abdominal organs focus significantly and tumor/lesion classification tasks from CT scans. However, there were only a few challenges that included the abdominal MRI series. This is an expected situation since because CT is preferred more than MRI in abdominal imaging. On the other hand, the recent advances in MRI technology make it an alternative method for a detailed analysis of the abdominal region. Significant improvements in MRI technology in terms of resolution, dynamic range, and speed make a joint analysis of both CT and MRI possible (Hirokawa et al., 2008).

Currently, there exist 9 international challenges (rather than CHAOS) focusing on abdominal organs (van Ginneken & Kerkstra, 2015). The summary of these challenges is presented in Table 3.1. SLIVER07 (Heimann et al., 2009; van Ginneken et al., 2007) can be considered the most important challenge because of being the pioneering one. In 2007, SLIVER07 has one task with is the segmentation of the liver from abdomen CT images. A comparative analysis of a number of liver segmentation algorithms was conducted under many obstacles, such as patient orientation

differences or tumors and lesions. In 2008, the same team created a new challenge, "3D Liver Tumor Segmentation Challenge (LTSC08)" (Deng & Du, 2008). They expanded the task of SLIVER07 to segment liver tumors from the abdomen CT scans. Shape 2014 and 2015 challenges (Kistler et al., 2013) targeted on liver segmentation from CT data. There are also such challenges that focus on multiple organs as well as abdomen ones. Anatomy3 challenge (Jimenez-del-Toro et al., 2016) is one of these challenges that provided a broad benchmark opportunity. It covers the segmentation of the left/right lung, urinary bladder, and pancreas in addition to the liver. Some challenges target both liver and tumors in the liver at the same time. LiTS - Liver Tumor Segmentation Challenge (Bilic et al., 2019) focuses on the segmentation of liver and liver tumors from CT scans. Pancreatic Cancer Survival Prediction (Guinney et al., 2017) covers a rare task that is the segmentation of pancreas cancer tissues in CT scans. Among with liver, kidneys are also highly examined organs in medical imaging. KiTS19 challenge (Weight et al., 2019) has the task of kidney tumor segmentation from CT data. Some abdomen related challenges use different modalities such as whole slice images. PAIP 2019 challenge (Choi et al., 2019) targets detecting liver cancer from these images.

Besides the challenges with local organizers, there are also important ones organized by the community of multiple scientists around the world. Medical Segmentation Decathlon (MSD) (Simpson et al., 2019) is one of these. MSD was organized in 2018 and focused segmentation of several organs/structures from multiple diverse datasets. The targeted areas are liver parenchyma, hepatic vessels and tumors, spleen, brain tumors, hippocampus, and lung tumors. The dataset of MSD includes both CT and MRI scans. These multiple modalities are used to evaluate the performance of proposed methods along with repeatability, reproducibility, and generalizability of the algorithms. MSD is a successful challenge to reveal important factors of DL-based methods and to push the methods into their boundaries.

Table 3.1 Overview of challenges that have upper abdomen data and task

Challenge	Task(s)	Structure (Modality)	Organization
SLIVER07	Single model segmentation	Liver (CT)	MICCAI 2007, Australia
LTSC08	Single model segmentation	Liver tumor (CT)	MICCAI 2008, USA
Shape 2014	Building organ model	Liver (CT)	Delémont, Switzerland
Shape 2015	Completing segmentation	Liver (CT)	Delémont, Switzerland
Anatomy3	Multi-model segmentation	Kidney, urinary bladder, gallbladder, spleen, liver, and pancreas (CT and MRI for all organs)	VISCERAL Consortium, 2014
LiTS	Single model segmentation	Liver and liver tumor (CT)	MICCAI 2017, Canada
MSD	Multi-model segmentation	Liver (CT), liver tumor (CT), spleen (CT), hepatic vessels in the liver (CT), pancreas and pancreas tumor (CT)	MICCAI 2018, Spain
KiTS19	Single model segmentation	Kidney and kidney tumor (CT)	MICCAI 2019, China
PAIP 2019	Detection	Liver cancer (Whole-slide images)	MICCAI 2019, China
CHAOS	Multi-model segmentation	Liver, kidney(s), spleen (CT, MRI for all organs)	ISBI 2019, Italy

3.3 "Karaciğer Bölütleme Algoritmaları Yarışıyor!" Challenge

Our research has shown that there is no recent challenge for the healthy liver segmentation problem in Turkey although challenges are very important events today.

It was decided to organize a new challenge to examine the up-to-date methods in segmentation algorithms and to bring new data to the literature. Therefore the first challenge in biomedical image analysis field in Turkey, "Karaciğer Bölütleme Algoritmaları Yarışıyor!" was organized. The challenge was organized under the supervision of the Turkish Medical Informatics Association (TURKMIA) at the BioİzmirIzmir building in Dokuz Eylul University Medical School Campus, İzmir, Turkey. This was the first competition about liver segmentation in Turkey. In addition, since 2007, there was not any challenge specially focused on healthy liver segmentation in the world. The most famous competition was organized in MICCAI 2007, "3D Segmentation in the Clinic: A Grand Challenge, on October 29, 2007" (Heimann et al., 2009). The idea of organizing a challenge was inspired by that grand challenge.

The challenge was organized nationwide and a one-time event. It was announced three months before. The participants registered to the challenge one month before the challenge day. After completing registration, the train set was shared with participants.

3.3.1 Aims and Data Information

The challenge has a single task: segmentation of healthy liver from CT images. The reason for this choice is that we want to examine the interest in a subject that has not been organized for many years.

The challenge data contains abdomen CT scans of 20 patients. The dataset includes healthy abdomen organs without any tumors, lesions, etc. The datasets were collected from the Department of Radiology, Dokuz Eylul University Hospital, Izmir, Turkey. CT image series were obtained at the portal phase during the injection of the contrast matter. At this stage, liver parenchyma reaches the biggest volume because of blood circulation inside of the organ from the portal vein. With the help of the contrast agent, the portal veins can be seen in detail. This protocol is one of the most performed liver CT imaging for both liver and veins. The details of the data are presented in Table 3.2.

Each image slice in the patient sets was annotated manually in order to guarantee

Table 3.2 Statistics about dataset in "Karaciğer Bölütleme Algoritmaları Yarışıyor!" challenge

Number of sets (Train + Test)	10 + 10
Spatial resolution of files	512 x 512
Number of files in all sets [min-max]	[78 - 264]
Average number of files in a set	90
Total files in the whole dataset	1207
X space (mm) [min-max]	[0.59 - 0.79]
Y space (mm) [min-max]	[0.59 - 0.79]
Slice thickness (mm.) [min-max]	[2.0 - 3.2]

the quality of annotation. The number of cases in the data was divided equally (10+10) for train and testing stages. Train data was shared as anonymized DICOM images and their annotations (ground truth). Test data was shared with only anonymized DICOM images.

3.3.2 Participants

Eleven different teams have participated in the challenge. They are from; Boğaziçi University, İstanbul Technical University, Yıldız Teknik University, Middle East Technical University, Hacettepe University, Eskişehir Osmangazi University, Bursa Uludağ University, Abdullah Gül University, Celal Bayar University, and Dokuz Eylül University. The cities of participants are shown in Figure 3.2.

In the competition day, all teams were informed about the rules of the challenge. After that, the whole dataset was shared with all teams at the same time and the competition started. Teams were free to use any kind of segmentation methods and tools. Also, there was no strict rule on the number of members in each team. The challenge lasted seven hours. Eight of eleven teams delivered results. The other three could not finish the segmentation of all sets. However, they were allowed to send their results after the challenge. Their grades were not included in the rating of the competition but they are valuable for further analysis of various segmentation results.

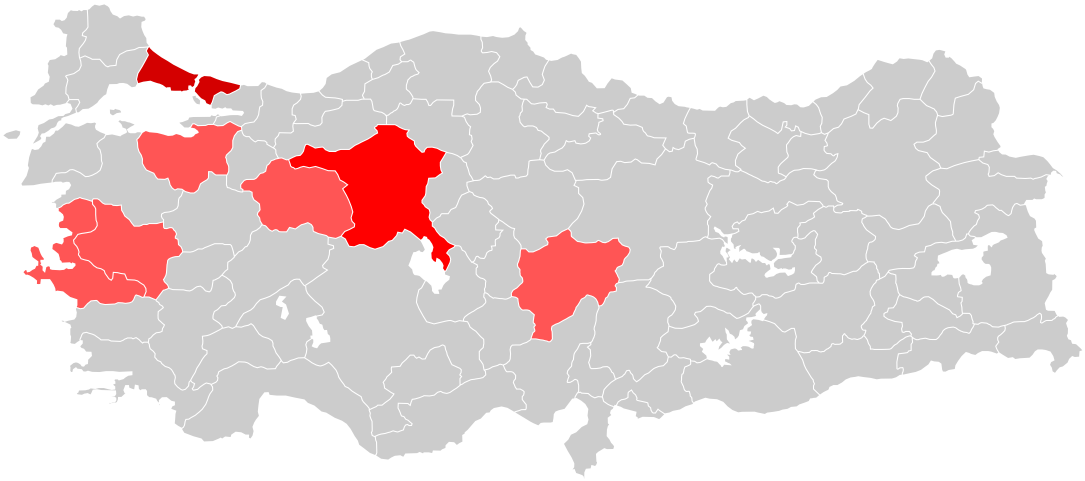

Figure 3.2 Distribution of universities participating in the challenge

3.3.3 Evaluation

After collecting the results from teams, their performances were evaluated. Selecting the proper evaluation metric(s) is the most critical point in these kinds of competitions. In literature, some metrics compares the two 3D objects to analyze how similar they are. On the other hand, none of them are sufficient alone (Taha & Hanbury, 2015). In order to overcome this problem, we determined to use five different performance metrics at the same time. Their average gives the final grade of the segmentation. This approach was also used previous segmentation challenges in the world (Heimann et al., 2009). The five different performance metrics are:

1. Volumetric overlap error (VOE)
2. Relative volume difference (RAVD)
3. Average symmetric surface distance (ASSD)
4. Root mean square symmetric surface distance (RMSD)
5. Maximum symmetric surface distance (MSSD)

The calculated error metrics were not used according to their value. To get proper analysis, their values are mapped between 0-100 points as other grand challenges.

However, our segmentation challenge has a different approach for evaluating the results at this point. In SLIVER07, the thresholds had a very narrow band. In other words, only very successful results could obtain a grade. Otherwise, they will be evaluated as zero points. Unlike SLIVER07, we extended the threshold limits wider because the main goal of the challenge is obtaining many different segmentation results with many different algorithms. The mapping thresholds are explained below:

1. VOE: The threshold for mapping is determined as 50%. If a VOE of result has lower than 50%, the score will be 0. If it is higher than 50% the score will be the same as calculated.

2. RAVD: The values higher than 10 get a grade of 0. The RAVD values between 10 and 0 are mapped between 50 and 100 score range. Since lower RAVD represents higher performance the mapping calculation from actual value to score has an inverse proportion.

3. ASSD: The values greater than 10 gets a grade of 0. The values between 10 and 0 are mapped between 50 and 100 score range. Again there is a inverse proportion between ASSD and scores.

4. RMSSD: The values greater than 15 gets a grade of 0. The values between 15 and 0 are mapped between 50 and 100 score range.

5. MSSD: The values greater than 50 gets a score of 0. The values between 50 and 0 are mapped between 50 and 100 score range.

After calculating case results from submissions, their average over the test data determines the final scores of the participants. All results and analyses of the challenge are discussed in Section 3.3.4.

3.3.4 Results

"Karaciğer Bölütleme Algoritmaları Yarışıyor" challenge was a one time and on-site event. The test set was shared with participants on the challenge day. After six

hours, eight of eleven teams made submissions. Segmentation results were evaluated according to the metrics that explained in Section 3.3.3. In order to make fairer analysis, results of automatic segmentation and semi-automatic segmentations examined individually. Hence, results were divided into semi-automatic methods and automatic methods as presented in Table 3.3 and 3.4.

Table 3.3 Results of teams using semi-automatic methods

Team Name	**Grade**
Team 3 - Boğaziçi University	72.57
Team 11 - Uludağ University	56.14
Team 5 - Abdullah Gül University	48.97

Table 3.4 Results of teams using automatic methods

Team Name	**Grade**
Team 1 - İstanbul Technical University Vision Lab	79.68
Team 4 - Middle East Technical University MM LAB	79.47
Team 8 - Yıldız Technical University	55.05
Team 7 - Hacettepe University	37.56
Team 6 - Osmangazi University	20.89

Team 3, the winner of the category of the semi-automatic algorithm used an algorithm based on the active contours method. This is a traditional algorithm for segmentation problem but it generated satisfying results. On the other hand, all teams in the automatic segmentation category used Deep Learning algorithms with a variation of U-Net (Ronneberger et al., 2015) model. This is a very interesting outcome because all the teams came from different universities with the same base DM. Hopefully, this situation gives us a very rare chance to examine the deep learning methods under different optimizations.

In order to observe the differences properly, the analyzes were handled via both qualitative and quantitative evaluation methods. First, all results of segmentation

algorithms were summed cumulatively to obtain a heatmap. The values were mapped to the virtual color scheme for qualitative examination. One of the examples can be seen in Figure 3.3.

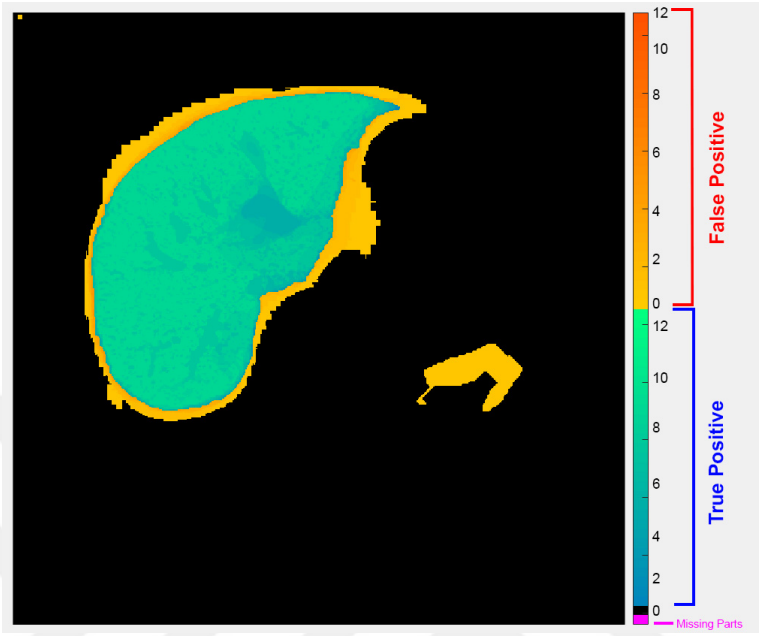

Figure 3.3 Colored heatmap of all segmentation algorithms on a sample slice

Figure 3.3 was colored according to two different color maps. The first color map that has greenish colors, represents inside of the ground truth mask. In this area, we want to examine True Positive (TP) performance of segmentation algorithms. Since the inside of the liver is not homogeneous, segmentation results have different characteristics. For example, a segmentation algorithm that is sensitive to intensity changes of voxels can miss the veins inside of the liver. In figure 3.3, it can be clearly observed that nearly half of the segmenters have some issues at vein regions.

On the other hand, a second reddish color map was chosen to explore situations of segmenters outside of the ground truth. This area is related to False Positive (FP) value of segmentation results. FP area has very critical importance for maximum error margin.

In addition to examining all segmenters, we also wanted to analyze semi-automatic results and automatic results separately. The qualitative analysis example of the two groups are presented in Figure 3.4 and 3.5.

The comparison between Figure 3.4 and 3.5 reveals that Deep Learning-based automatic approaches have lower FP results except for one submission. In addition, they performed more successfully on heterogeneous structures in the liver such as veins inside of the parenchyma tissue.

After qualitative analysis, the results were examined quantitatively. Again the results were analyzed into three different sections; All segmentation results, automatic segmentation results, and semi-automatic segmentation results. The results were merged via logic OR operator. The meaning of OR operator is summing all results together and it gives the complementarity of different of results. After that FN and FP numbers of voxels were count. Unsegmented voxels show the regions that none of the segmentation algorithm could find. The results were presented in Table 3.5.

According to qualitative and quantitative analyses, it can be clearly observed that automatic and semi-automatic methods complete their results and decrease FN error if they are summed. The summation of their results covers almost all voxels inside the liver. The percentage of FN voxels is under 0.3% in all sets. However, the situation of FP voxels tells another story. The segmentation results of all methods show complementing behavior inside the liver while they dramatically increase FP voxels.

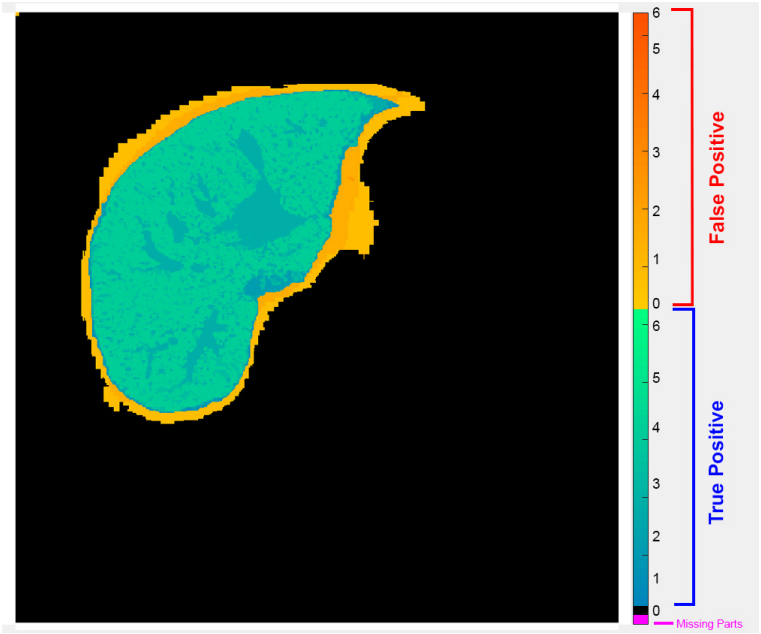

Figure 3.4 Colored heatmap of semi-automatic segmentation algorithms on a sample slice

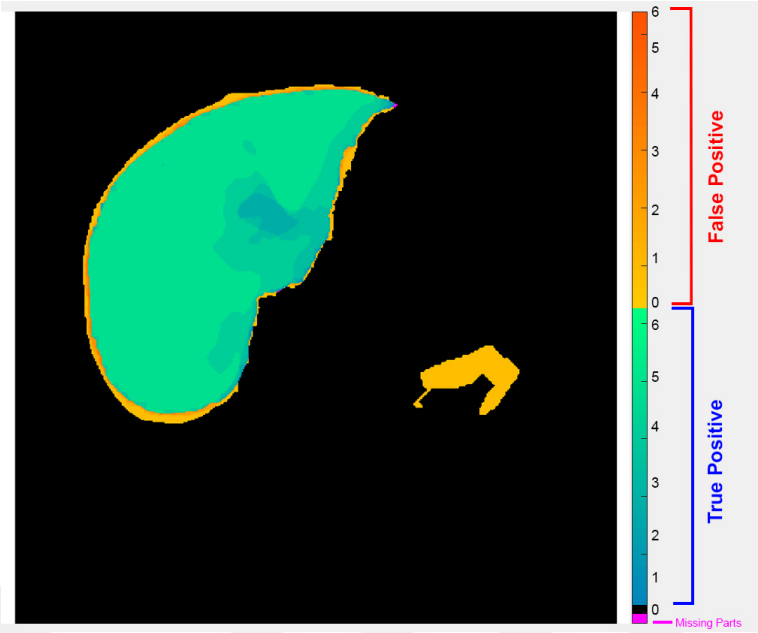

Figure 3.5 Colored heatmap of automatic segmentation algorithms on a sample slice

Therefore, the evaluation of these results with metrics used in our liver segmentation challenge gives a result of zero points almost every set. Hereby, the summation of all results cannot be used as an ensemble solution.

In addition to complementary analyses, diversity of results with AND operator was calculated to analyze the intersection of all results. The results are presented in Table 3.6. Table 3.6 shows that segmentation methods have a diverse characteristic. This causes two results. The first outcome is the diversity of results is very effective while trying to decrease false-positive (FP) voxels that belong to outside of the liver. Another outcome of the OR operator is that this diversity dramatically increases false negative (FN) voxels that are inside the liver.

To sum up, despite a small size of the event, "Karaciğer Böltüleme Algoritmaları Yarışıyor" challenge has attracted the medical imaging teams at national level. Since it was one time and on-site challenge, there was not any possibility for peeking. Therefore the impartiality of the results could be validated. The results revealed an important truth about organ segmentation analysis: Up to this challenge, semi-automatic segmentation methods generated more precise and accurate results than automatic ones. Automatic segmentation methods, generally needed additional

Table 3.5 Quantitative analysis of OR operator in three group; FN, FP and Total number of voxels

	All Results		Automatic Results		Semi-Automatic Results		
Set	FN Voxels	FP Voxels	FN Voxels	FP Voxels	FN Voxels	FP Voxels	All Voxels
1	2	1662601	643	629442	27	1407447	2086084
2	175	967416	941	259452	5540	893253	1159488
3	175	2579244	13658	630565	565	2469255	1351846
4	2069	3276054	14604	1205892	5865	3067914	2377134
5	662	5516280	20184	633903	2160	5333218	2214195
6	3729	4269658	27506	1159656	7408	4057793	2572660
7	2066	2208197	11998	1016063	27882	1889581	1495046
8	251	3058514	882	1515891	13419	2840737	1907329
9	434	3646886	25954	761290	1225	3500360	1608064
10	815	1810019	18931	726851	2119	1569483	2157990
11	434	2186193	7915	684263	782	1925671	1514495
12	2158	3288102	27380	1223672	3683	2933552	1940596
13	1661	2331382	43136	800343	1773	2151439	1877030
14	9	3560787	6767	875526	271	3403366	2019476
15	884	2737363	13039	934882	1400	2499171	1788934
16	13	3739378	2401	659399	325	3623870	1788285
17	453	2246968	33164	759332	919	1834653	1622016
18	152	2644044	5447	720631	5236	2579218	2436574
19	1176	3790495	32109	1025182	3000	3595310	2073975
20	210	7301275	15594	1136240	511	7152412	2368338

post-processing to obtain clinically acceptable results. However, the tremendous developments in machine learning, specifically DL, has changed the situation. "Karaciğer Bölütleme Algoritmaları Yarışıyor" challenge shed light on the developments in this field and constituted the main motivation of our next contest, the CHAOS challenge. All analyzes made from the challenge results were compiled and published in Kavur et al. (2020b). The experience obtained during this challenge made possible to design more sophisticated and different tasks in the CHAOS challenge.

Table 3.6 Quantitative analysis of AND operator in three group; FN, FP and Total number of voxels

	All Results		Automatic Results		Semi-Automatic Results		
Set	FN Voxels	FP Voxels	FN Voxels	FP Voxels	FN Voxels	FP Voxels	All Voxels
1	945070	5	488781	894	760987	1096	2086084
2	634303	0	271008	503	580302	21	1159488
3	732750	177	408931	1205	601994	1330	1351846
4	1693867	514	1466430	2755	790809	21285	2377134
5	1071209	447	733280	1214	742672	2804	2214195
6	1831292	274	903207	4261	1625700	2160	2572660
7	910377	313	534426	2485	751089	4065	1495046
8	1463873	16	1015110	835	1414284	222	1907329
9	954982	195	916518	381	409433	8435	1608064
10	1624076	277	1017123	810	1332932	930	2157990
11	994169	423	612560	4407	753398	2742	1514495
12	1471657	145	722448	2458	1343178	811	1940596
13	1428023	8	1089869	47	1053893	422	1877030
14	1004651	90	654437	1658	750820	3161	2019476
15	897802	196	625249	1880	682405	829	1788934
16	938185	486	314994	1237	860777	4210	1788285
17	1207797	235	800154	1811	969828	1780	1622016
18	1230673	4	517836	5064	1147755	136	2436574
19	1308655	303	957166	4208	997995	1739	2073975
20	1289932	260	670869	1939	1023226	1577	2368338

3.4 The CHAOS Challenge

The popularity of "Karaciğer Bölütleme Algoritmaları Yarışıyor!" challenge showed us the potential of the challenges on abdomen imaging. Therefore, it was decided to organize an international challenge with extended data.

Many previously organized challenges for the abdomen organ segmentation focused on the segmentation of single organs from a single modality, especially CT. On the other hand, the researches presented in Section 3.2 revealed that the aims of abdomen organ segmentation challenges should be improved due to huge progress on Deep Leaning studies. More specifically, the traditional abdomen organ-related

challenges are not powerful enough to deeply analyze state-of-the-art algorithms. In addition, there is a new trend in deep learning studies to create a single solution for multiple clinical needs. Therefore, DMs working with multi-modal and/or cross-modality are recent, but not well-studied topics (Cerrolaza et al., 2019). In order to gain awareness on these topics, CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation challenge was designed. The challenge has on-site and online sections. The on-site session of the challenge has been organized in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI), 2019, in Venice, Italy as a one-time event. After that, the online submission system was enabled to give the opportunity of participating in the challenge from all around the world.

Unlike traditional abdomen imaging challenges, CHAOS contains unpaired abdominal CT and MR data from patients with healthy organs. Five individual tasks have been proposed to deeply investigate the effectiveness of up-to-date methods in many aspects. The data and the benchmark platform will provide a continuous resource.

3.4.1 Aims and Tasks

The aim of the CHAOS challenge is to create a benchmark platform to deeply analyze state-of-the-art segmentation solutions for abdomen images. To serve this purpose, different data and tasks have been designed in the competition. Unlike single modality and single task challenges, CHAOS brought new kinds of goals to the field such as multi-organ segmentation from multiple modalities. Multi-organ based activities involve a comprehensive description of complex and adaptive abdominal anatomy. Hence, new efficient computer and machine learning models are needed in this developing field. CHAOS was designed to improve the field by addressing new DL ideas for multi-modal segmentation and cross-modality segmentation. The emphasis is on the segmentation of multiple organs from unpaired modalities: CT and MR. The overall aim of CHAOS is distributed into five different and complementary tasks, which demand the participating systems to have a higher generalization and

translation capabilities. With these tasks, CHAOS offers the participants different design options for segmentation algorithms:

Task 1: Liver Segmentation from CT and MRI aims the use of a single algorithm capable of segmenting liver from multiple modalities, CT, and MRI. In other words, the proposed system can be to handle 'cross-modality' images. According to researches, (Valindria et al., 2018) cross-modality systems will be used more preferably due to huge improvements in deep learning studies.

Task 2: Liver Segmentation from CT is the most studied and a typical segmentation task. It focuses on segmentation liver from CT image series. Although it is a relatively easy task, the performance on vein and low gradient borders of the liver is still challenging because liver segmentation algorithms making the most mistakes in those regions.

Task 3: Liver Segmentation from MRI targets the same problem with Task 2 but on different modality. The systems need to segment liver from MRI image series. Since MRI series have two sequences (detail will be presented in the following sections), it is a more challenging task than Task 2. In addition, the low resolution of the MRI images makes accurate segmentation harder.

Task 4: Segmentation of abdominal organs from CT-MRI is the most complicated task of the challenge. It combines both cross-modality and multi-organ segmentation in a single task. A single solution needs to handle 1) segmentation of liver from CT series, 2) segmentation of four abdominal organs (liver, kidneys, and spleen) from the MRI series.

Task 5: Segmentation of abdominal organs from MRI is the widened version of Task 3 with multiple organ segmentation aims from MRI scans. Here, an algorithm needs to segment the liver, both kidneys, and spleen at the same time.

The tasks were designed for the replication of the real-world needs of physicians. For example, a single system (i.e. software for clinical usage) can be designed for cross-modality (both CT and MRI data). On the other hand, participants can prefer

classical approaches, such as designing a segmentation system for only one modality and one organ. The ensemble done by hand or manually of individual models working on specific modality is not allowed. However, the ensemble of different solutions for MRI sequences (T1-DUAL and T2-SPIR) is valid in all MRI-included tasks. More details are published on the CHAOS challenge website (https://chaos.grand-challenge.org/).

3.4.2 Data Information and Details

As mentioned in previous sections, the CHAOS challenge contains abdomen scans from two different modalities: CT and MRI. In total, the whole data consist of 80 patients' images. 40 of them belong to CT data while the other 40 are in MRI data. All organs of interest (liver, kidneys, spleen) in the images are in healthy condition. The source of all images is Dokuz Eylul University Hospital, Department of Radiology, Izmir, Turkey. The technical details of the data are summarized in Table 3.7.

Table 3.7 Statistics about CHAOS CT and MRI dataset

Specification	CT	MR
Number of patients (Train + Test)	20 + 20	20 + 20
Number of sets (Train + Test)	20 + 20	60 + 60*
Spatial resolution of files	512 x 512	256 x 256
Number of files in all sets [min-max]	[78 - 294]	[26 - 50]
Average number of files in a set	160	32x3*
Total files in the whole dataset	6407	3868x3*
X space (mm) [min-max]	[0.54 - 0.79]	[0.72 - 2.03]
Y space (mm) [min-max]	[0.54 - 0.79]	[0.72 - 2.03]
Slice thickness (mm.) [min-max]	[2.0 - 3.2]	[4.4 - 8.0]

* MRI sets have 3 different pulse sequences. For each patient T1-DUAL (in) and (oppose) phases (registered) and T2-SPIR phase are acquired.

3.4.2.1 Dataset 1: Abdomen CT images

The CT data is the extension of the same dataset in our previous challenge, "Karaciğer Bölütleme Algoritmaları Yarışıyor!". Therefore, the CT data has similar characteristics explained in Section 3.3.1. The CT database includes images of 40 different patients. The scans were acquired at the portal venous phase. In this phase, first, a contrast agent is injected into the patient. Then, the scan starts 70-80 seconds after injection. As explained in Section 3.3.1, portal vein of the liver reaches maximum blood supply. Therefore, the portal veins are highly enlarged. It becomes easier to analyze portal veins due to the effect of the contrast agent. The portal phase is one of the most used abdomen CT scanning protocols in clinics for the liver and vessel segmentation. The CT data was acquired via three different CT scanner. They are Philips SecuraCT with 16 detectors, Philips Mx8000 CT with 64 detectors, and Toshiba AquilionOne with 320 detectors. Each case in the CT data has the same patient orientation and alignment with the following specifications:

- Similar range of Hounsfield values of neighbor organs,
- Different Hounsfield ranges for inferior vena cava and portal veins across data sets because of the contrast agent,
- Important variations in the shape of anatomical structures across patients,
- 15% of the data includes atypical shapes of the liver (i.e. abnormal volume or location of the liver).

To sum up, the CT dataset in the CHAOS challenge has very similar specifications with in real life utilization. Hence, the algorithms must be designed to handle these problems (shown in Figure 3.6) in advance.

3.4.2.2 Dataset 2: Abdomen MR images

MRI dataset contains 120 cases from two individual MRI sequences as follow:

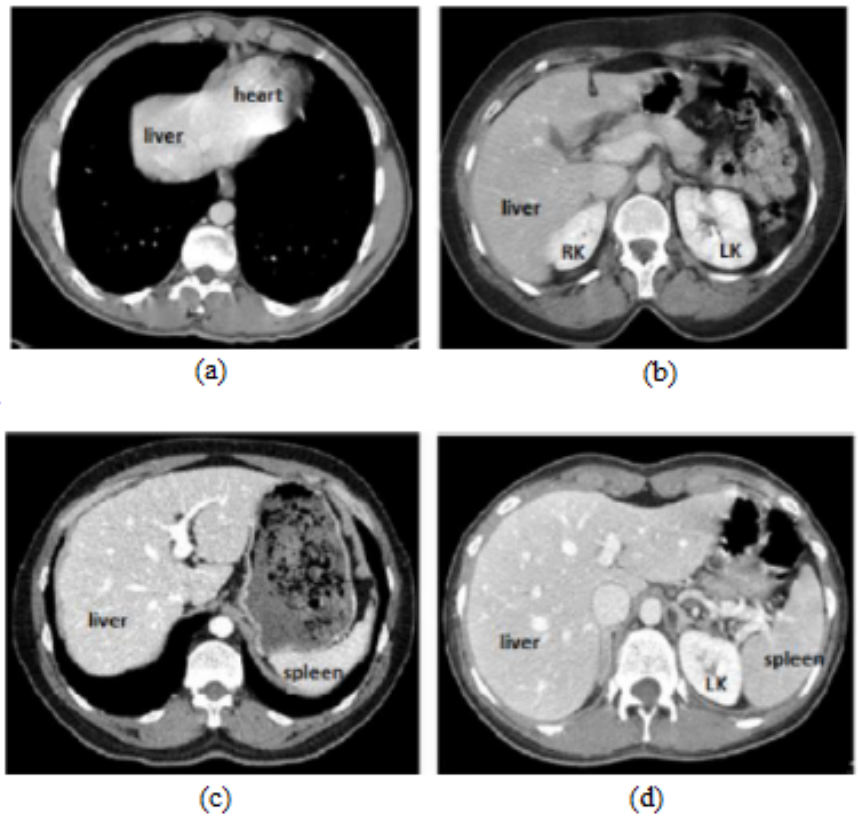

Figure 3.6 Sample images from CHAOS CT dataset. (a) very low contrast difference and unclear boundary between the heart and the liver; (b) unclear boundary due to partial volume effects between the right kidney and the liver; (c) contrast enhanced vascular tissues inside the liver parenchyma; (d) relatively less enhanced vessels compared to (c) (Kavur & Selver, 2019)

- T1-DUAL in- and oppose- phase images from 40 patients,
- T2-SPIR from 40 patients.

The scans were performed for routine clinical examination. Different gradient and radiofrequency parameter pairs were used to obtain different sequences. As same as the CT dataset, MRI data contains not tumors or lesions in the target organs. The MRI scans were obtained via a 1.5T Philips MRI scanner. The technical specifications of the dataset are presented in Table 3.7.

This dataset has two modified versions of T1 and T2 sequences that are widely used ones in daily clinical routine. The properties and differences of T1 and T2-weighted images were explained in Section 2.2.2. In this data set, images obtained by T1-DUAL

and T2-SPIR, specific versions of T1 and T2 sequences.

SPIR (Spectral Pre-Saturation Inversion Recovery) provides for a synthetic image series which utilizes a T2-weighted contrast method. The pre-saturation pulse shall be employed individually to each slice selection gradient for selective suppression of fat protons. SPIR needs delicate calibration adjustment and a very uniform magnetic field. Therefore SPIR is used for liver scanning because it is easier to examine the liver with suppression of the fat tissue in the parenchym. Since there is fat content between abdomen organs, the borders of the organs appear more clearly with darker intensity values. The veins in the liver can also be detected because they seem hyper-intense. The neighbor abdominal organs and structures are more detachable from each other. Another significant feature to the SPIR is its modest sensitivity to patient movement. This feature minimizes the artifacts which reduce the quality of the scans in abdominal examinations.

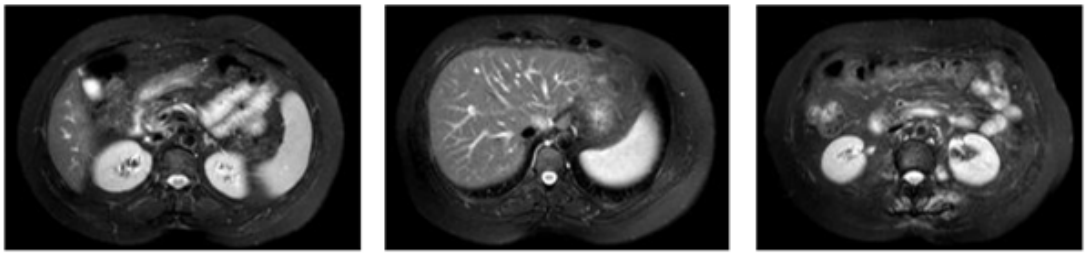

Figure 3.7 Samples of abdominal MRI images from T2-SPIR sequence (Kavur & Selver, 2019)

T1-DUAL is a series of fat suppression with in-phase and oppose or out-phase, which incorporates the disparity between water and fat protons. In-phase and oppose-phase images come from two signal acquisitions from two phases of the protons. By using this information to determine the Time of Echo, fat suppression is obtained by the difference between related water and fat signal frequencies. This series is very useful in recognizing the substance of fat in structures especially lesions. Because T1-DUAL is a T1-weighted sequence, the identification of tissues and blood with high protein content is very compelling. This series also assists in assessing the amount of liver fat. The edge of the structures appears to be dark in oppose-phase scans, because of the rapid switch in the load of water and fat which blocks the acquired signal. This T1-DUAL feature is also preffered for the algorithms

for boundary detection.

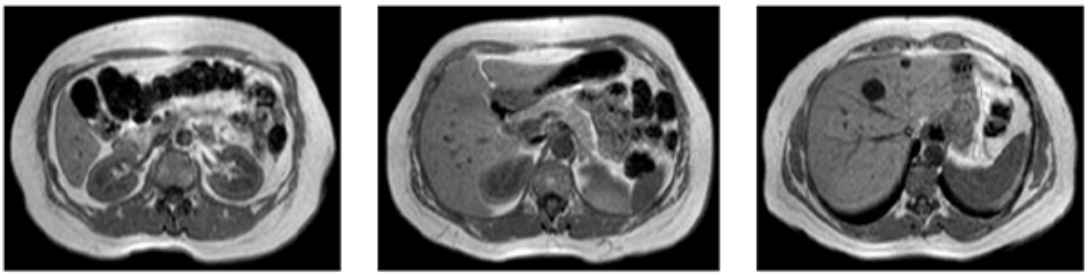

Figure 3.8 Samples of abdominal MRI images from T1-DUAL (in-phase) sequence (Kavur & Selver, 2019)

3.4.3 Annotation of the dataset

Annotation is one of the most prolonged stages in challenge design. The quality of annotations has a direct role in both the training and testing of algorithms. There are different approaches to handle this step. The most commons ones, manually (slice-by-slice), semi-automatically (with help of a segmentation tool), crowdsourcing (from a service such as Amazon Mechanical Turk). In the CHAOS challenge, the most precise but most difficult method was preferred as a manual annotation. The data was annotated by three different radiology experts who have 10, 12, and 28 years of experience, respectively.

Sometimes even experienced radiologists may have different decisions for ground truths. To achieve consistency, the experts reached consensus over critical regions. For example, inferior vena cava (IVC) was excluded if it is not completely inside of the liver as shown in Figure 3.9.

The annotation step took more time than expected due to manual segmentation. However, this hand-crafted data has very valuable references considering crucial effects of annotations over algorithm design, validation, and evaluation.

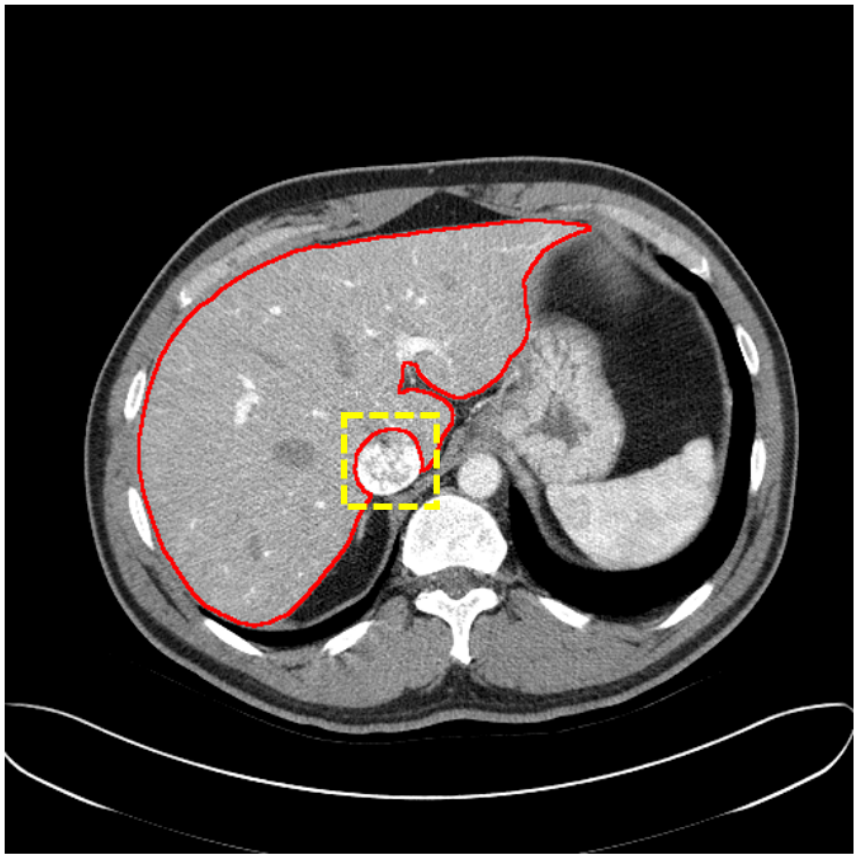

Figure 3.9 In CHAOS dataset, partial IVC regions (marked with of yellow dashes) were excluded

3.4.4 Challenge Setup and Distribution of the Data

To give adequate data containing wide variability, the training data were chosen to include both the challenges determined throughout the database (i.e. for CT scans, partial volume effects or for MRI scans, bias fields, abnormal liver forms) (Figures 3.10 and 3.11).

The images are distributed as DICOM file series as original format after anonymization of the files. All patient-related information was erased in order to follow privacy rules. The ground truths in the training data are also included as image PNG series to match the original order of the DICOM file series. One of the important aims of the challenges is to provide data for long-term academic studies. Thus, CHAOS data is available free with its DOI number via the zenodo.org webpage under CC-BY-SA 4.0 license (Kavur et al., 2019). It is expected that this data will be

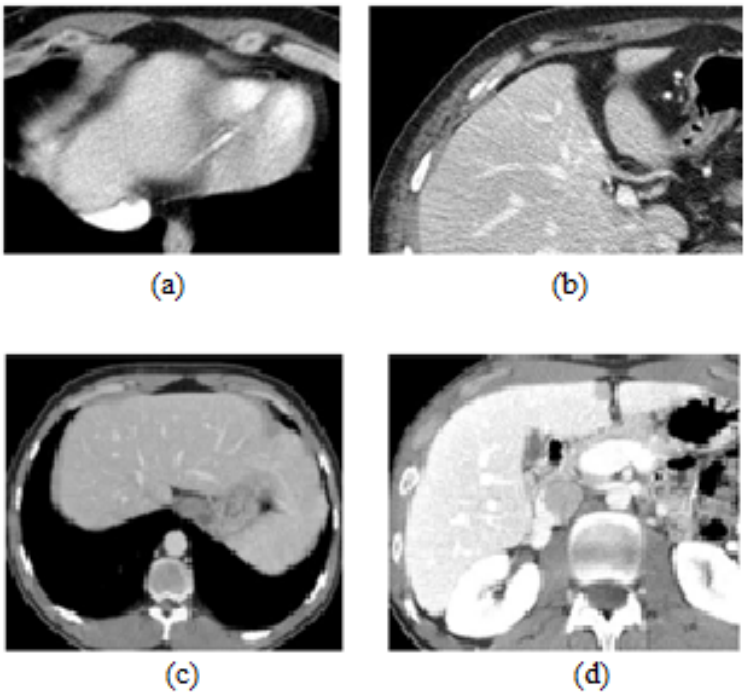

Figure 3.10 Examples of challenges from the training data of the CT database. (a) Unclear boundary between the liver and the heart. (b) Liver has three disconnected components on a single slice (c) Atypical liver shape, which causes unclear boundary with the spleen (d) Varying Hounsfield range and non-homogeneous liver parenchyma due to the injection of contrast media (Kavur & Selver, 2019)

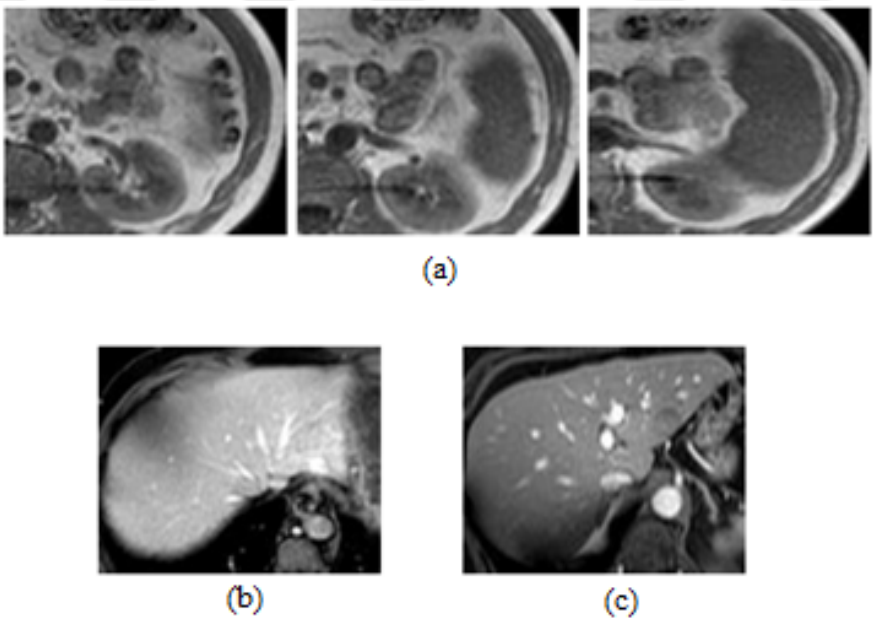

Figure 3.11 Examples of challenges from the training data of the second database (abdominal MRI) (a) sudden changes in planar view and unclear boundary (spleen-left kidney). Effect of bias field in (b) T1-DUAL, and (c) T2-SPIR (Kavur & Selver, 2019)

used not only for the CHAOS challenge but also for other scientific studies such as cross-modality works, medical image synthesis from different modalities, and so on. Data is already used for development in some prestigious studies such as Dou et al. (2020).

3.4.5 Evaluation

Evaluation of the CHAOS challenge has two main stages. First, the segmentations are evaluated by selected metrics. After that their results are converted to scores. Finally, the submissions are ranked by their scores.

3.4.5.1 Metrics

There is no standardization of metric(s) to evaluate the segmentation efficiency. The majority of segmentation related studies only use a single metric or multiple metrics with similar properties. However, there are significant findings that show these approaches may not enough for a proper and complete evaluation (Maier-Hein et al., 2018; Yeghiazaryan et al., 2015). The error margin in medical imaging is very strict in comparison with segmentation applications in different fields. Therefore using multiple and diverse metrics guarantees fair evaluation. Hence, a similar approach within "Karaciğer Bölütleme Algoritmaları Yarışıyor!" challenge followed. Four widely-used and proven segmentation metrics (Maier-Hein et al., 2018) were preferred in the CHAOS challenge. These are:

- DICE coefficient (DICE)
- Relative absolute volume difference (RAVD)
- Average symmetric surface distance (ASSD)
- Maximum symmetric surface distance (MSSD)

These metrics are capable of analyzing the segmentation in terms of overlapping, volumetric, and spatial differences. Details of all metrics were presented in Section 2.4.

In addition, we prepared two experiments for the uncover drawbacks of two most popular segmentation evaluation metrics (DICE and MSSD/Hausdorff distance) in literature (Maier-Hein et al., 2018) to demonstrate that using single metric may fail the evaluation. The figures containing different segmentation scenarios and results of them are presented in Figure 3.12, 3.13 and Table 3.8, 3.9.

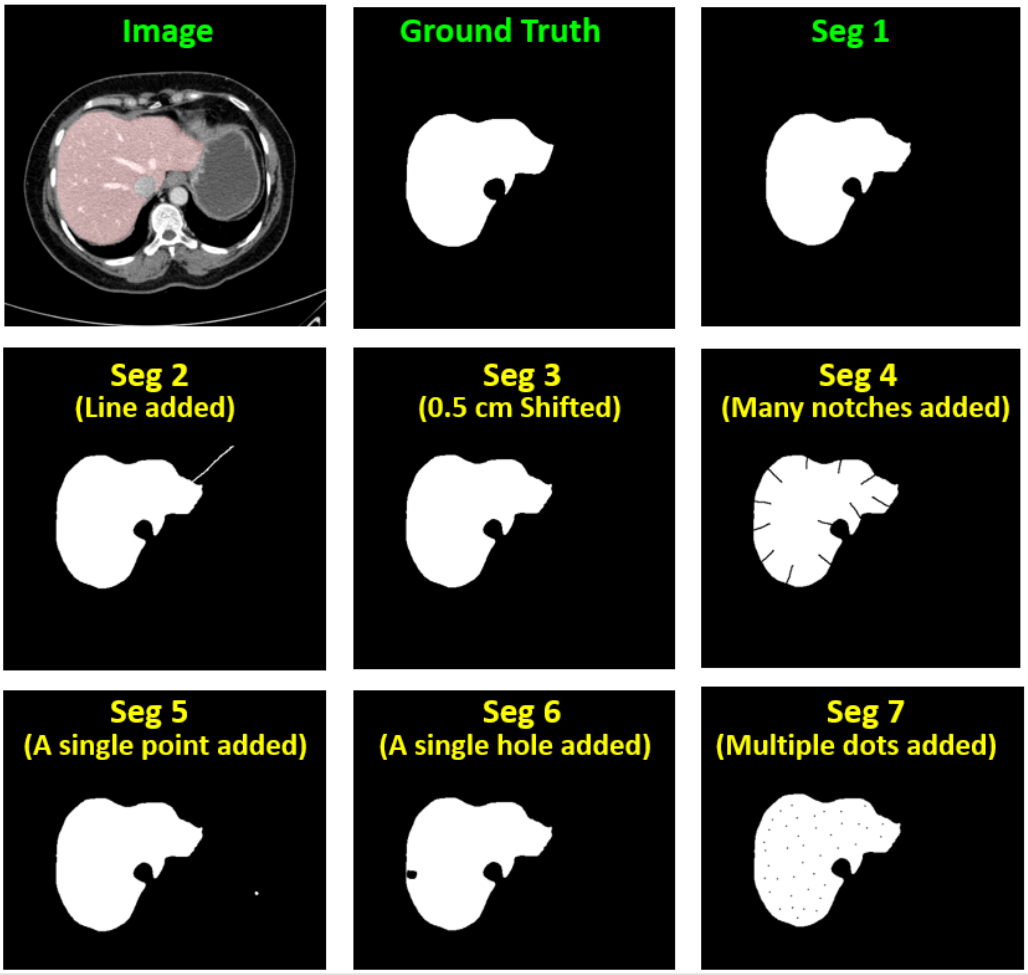

Figure 3.12 From top-left: 1) A sample slice from CHAOS CT data. 2) Its ground truth. 3) Segmentation result of an algorithm. 4-9) Syntactically manipulated version of (3) for DICE metric experiment

Table 3.8 Metrics results of segmentations in Fig.3.12. In many conditions marked bold (except Seg 3 and Seg 4), DICE metric is not sensitive for the different segmentation errors

Image	DICE	RAVD	ASSD	MSSD
Seg 1	**0.985**	0.159	2.99	53.731
Seg 2	**0.986**	0.729	1.313	53.731
Seg 3	0.972	2.412	3.731	53.731
Seg 4	0.782	55.701	16.128	53.731
Seg 5	**0.974**	2.023	3.58	53.731
Seg 6	**0.985**	6.449	3.632	53.731

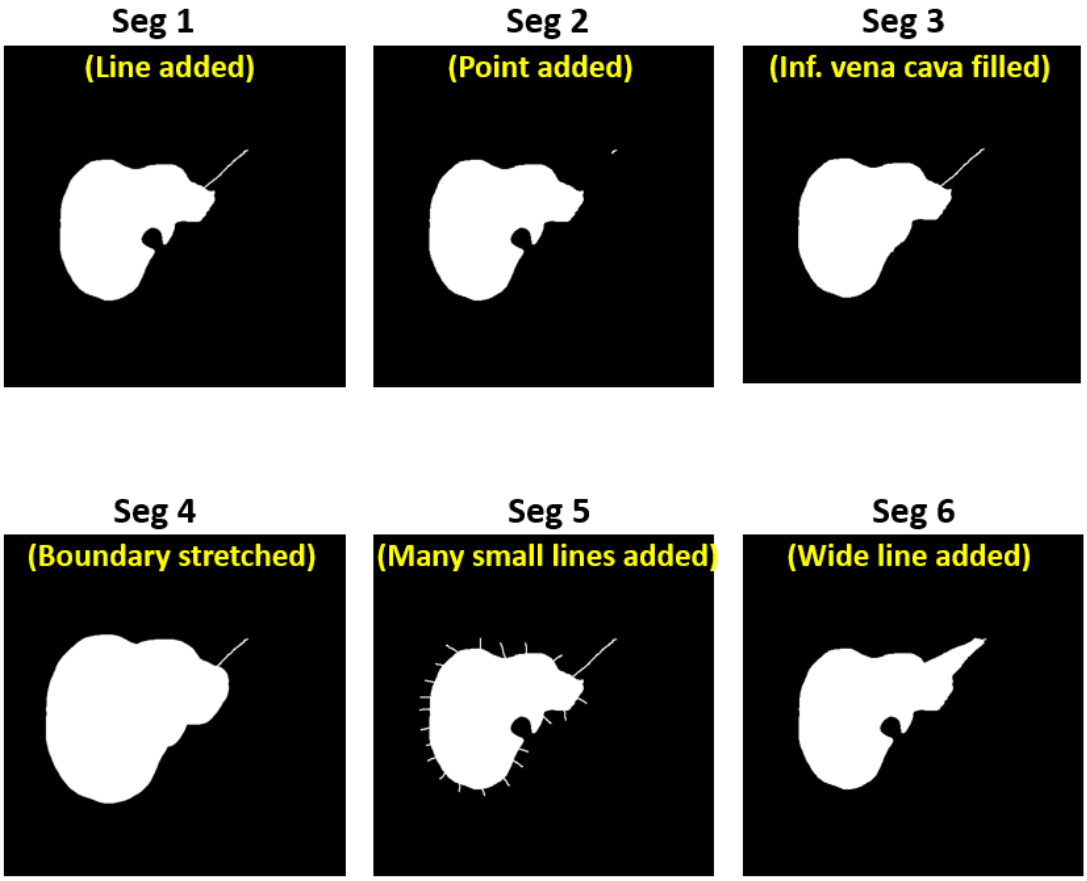

Figure 3.13 Syntactically manipulated segmentation results of (3) in Fig.3.12 for MSSD/Hausdorff distance metric experiment

Table 3.9 Metrics results of segmentations in Fig.3.13. In all cases MSSD/Hausdorff distance have same value. Thus, it is not possible to distinguish the different segmentation errors with single metric usage

Image	DICE	RAVD	ASSD	MSSD
Seg 1	0.985	0.159	2.99	**53.731**
Seg 2	0.986	0.729	1.313	**53.731**
Seg 3	0.972	2.412	3.731	**53.731**
Seg 4	0.782	55.701	16.128	**53.731**
Seg 5	0.974	2.023	3.58	**53.731**
Seg 6	0.955	6.449	3.632	**53.731**

3.4.5.2 Scoring System and Ranking

After calculating metrics, there are two widely-used approaches for ranking the submissions with multiple metrics. They are "rank then aggregate" and "aggregate than rank". In "rank then aggregate" method, the submissions are ranked using individual metrics. In other words, each submission has multiple rankings coming from individual metrics. After that, the mean of multiple rankings determines the final rank of the submission. Here, there is a possibility that multiple submissions will receive the same rank. The second approach, "aggregate than rank" uses inverse steps. First, the metrics results of the submission are converted to the same scale (score), then the mean of all scores determines the final score of the submission. After all, submissions are ranked via their final score. This is a more preferable way of ranking. The only important step is, scaling different metrics outputs to the same intervals.

In CHAOS, "aggregate than rank" was used approach as our previous challenge. The values of each metric have been converted to $[0, 100]$ range. Here, higher values represent better segmentations. The thresholds for the transformations are obtained by the intra- and inter-user similarities among ground truths coming from our radiology experts. Since the values of thresholds have a very crucial effect on ranking, using the values from real-life utilization is preferred again (Maier-Hein et al., 2018). Two

manual segmentations performed by the same expert on the same CT data set at different times resulted in liver volumes of 1491 mL and 1496 mL. The volumetric overlap is found to be 97.21%, while RAVD is 0.347%, ASSD is 0.611 (0.263 mm), and MSSD is 13.038 (5.632 mm). A similar analysis for the segmentation of the liver from MRI was performed. By using these values, thresholds were determined. The metrics and scoring system were summarized in Table 3.10.

Table 3.10 Details of metrics and threshold values in the CHAOS challenge. Δ represents longest possible distance in the 3D image

Metric name	Best value	Worst value	Threshold
DICE	1	0	DICE >0.8
RAVD	0%	100%	RAVD <5%
ASSD	0 mm	Δ	ASSD <15 mm
MSSD	0 mm	Δ	MSSD <60 mm

As our previous challenge, zero points are given for the metric results out of thresholds. Other values in the range of thresholds are scaled to $[0,100]$ range. The average scores determine the case (patient image set) score. The average of all case scores generates the final score of the submission. If a case does not have a score due to missing data, the zero point is given for this case. These zero points are also used for the final score. In other words, missing cases are penalized. The code for all metrics (in MATLAB, Python, and Julia) is available at https://github.com/emrekavur/CHAOS-evaluation.

3.4.6 Methods of Participants

CHAOS challenge gave us a unique opportunity to examine up-to-date and sophisticated methods in abdomen organ segmentation. In this section, participants' algorithms in the on-site challenge session are briefly explained to compare their solutions. In addition, there are three selected methods from the online challenge session. The reason for selecting these three submissions is that they have valuable approaches and already won other challenges.

Since peeking can dramatically impact the results as explained in Section 3.1.1, here all explained methods were validated that there was no peeking attempt. All methods have proper development stages. They are fully automatic methods that use different DM architecture in order to handle challenge tasks. After the explanations of the methods, their approaches can be compared via Table 3.11 and 3.12. All results and discussions of the CHAOS challenge are presented in Section 3.4.7.

3.4.6.1 OvGUMEMoRIAL

OvGUMEMoRIAL team participated in all tasks in the challenge. The DM that OvGUMEMoRIAL designed is based on U-Net architecture (Ronneberger et al., 2015). However, they modified it with an adaptation of attention U-Net (Abraham & Khan, 2019). Here they preferred soft attention gates from the reference design. Input images are the multi-scaled matrix for more accurate feature extraction. At each scale level, they used Tversky loss as a loss function. The most significant alteration in the DM is Parametric ReLU. Parametric ReLU has more parameters than the typical ReLU function. The additional parameter is called "coefficient of leakage" which is also trained with CNN. They used Adam optimizer in the training sessions. 120 epochs with 256 batch sizes are used for the training of the DM.

3.4.6.2 ISDUE

Team ISDUE participated in all tasks in the challenge. The design has three main blocks.

1. Prior encoder(f_{enc_p}) and prior decoder(g_{dec}) in a convolutional auto encoder network.
2. Imitating encoder(f_{enc_i}) and imitating decoder(g_{dec}) in a segmentation hourglass network.

3. In order to enhance g_{dec} a U-Net module was added to the system. The enhancement is handled by ordering decoding stage for accurate localization.

The autoencoder (1) is tuned by the DICE loss function. Optimization is handled by Adam optimizer with a learning rate of 0.001. Blocks (2) and (3) are optimized individually. The loss function is regularized DICE los. To train each model, 2400 iterations are performed in a single batch. Also, random translation and rotation operations while training is used for data augmentation.

3.4.6.3 Lachinov

Team Lachinov focused on the Tasks that have a single organ (tasks 1,2,3). The design of the DM is based on 3D U-Net (Ronneberger et al., 2015). There are also skip connections between encoder and decoder blocks. A residual network is preferred in the encoder block to improve the training process. Unlike many DMs that use batch normalization (Ioffe & Szegedy, 2015), the proposed DM uses group normalization (Wu & He, 2018). Random mirroring, 90 degrees rotation in random directions, and intensity shift are used for data augmentation.

3.4.6.4 IITKGP-KLIV

IITKGP-KLIV made submissions for all tasks. In other words, they need to carry out multi-modality segmentation using a single system. Therefore, they adapted the multi-task adversarial learning strategy to SUMNet (Nandamuri et al., 2019) that used a base network model. There are two complementary segmenters (C1 and C2) to handle adversarial learning. Also, there is a single discriminator (D) network in the model.

C1 is trained by feedback from the SUMNet encoder that offers modality-specific functionality. The C2 classifier is responsible for determining the class labels of the chosen segmentation maps. Cross entropy loss is used during training of C2 and

segmentation network. C1 and D are trained by the cross-entropy loss function. In the optimization stage, Adam optimizer is preferred. The developed model is capable of processing all modalities (CT, MRI T1-DUAL In Phase, MRI T1-DUAL Oppose Phase, and MRI T2-SPIR) in the challenge.

3.4.6.5 METU_MMLAB

METU_MMLAB participated in MRI related tasks: 1,3,5. U-Net is used as the base framework as other methods in the challenge. Besides, they integrated a Conditional Adversarial Network (CAN) in the model. Before each convolution, batch normalization is done in order to keep vanishing gradients and improve selectivity. In addition, instead of typical ReLU, parametric ReLU with a trainable leakage parameter is used to retain the negative values using.

The benefit of adding CAN to the model is to enhance accuracy around sharp edges around the organs. This brings a new loss function. The loss function is used for the regularization of parameters for spinous edges. After the segmentation is finished, 3D connected component analysis is used to eliminate small artifacts in the results.

3.4.6.6 PKDIA

PKDIA made submissions to all tasks in the challenge. The proposed model uses conditional generative adversarial networks (GAN) approach. Here the encoder is made of cascade-connected pre-trained encoder-decoder networks in the standard U-Net (Ronneberger et al., 2015) model. The encoder part of U-Net is replaced by VGG-19, a bigger network. To sum up, the difference between the proposed model and standard U-Net are: 1) 64 channels are produced by the first convolutional layer instead of 32 channels in U-Net. 2)The number of channels amplified until 512 after max-pooling steps (it is 256 in U-Net). 3) 4 cascade convolutional layers used after the second max-pooling step (2 in U-Net). 4) Adam optimizer is used during training. The loss function is determined as Fuzzy DICE score.

3.4.6.7 MedianCHAOS

MedianCHAOS focused on only Task 2. Their approach is using the ensemble of multiple networks. For this purpose, the final segmentation is calculated by the mean of five individual CNNs. These five different models are 1) DualTail-Net, 2) TernausNet (U-Net with VGG11 architecture (Iglovikov & Shvets, 2018)), 3) LinkNet34 (Shvets et al., 2018), 4) ResNet-50, 5) SE-Resnet50.

DualTail-Net consists of a single encoder, 2 connected parts in the decoder, and a central block between them. Downsampling is handled by the max-pooling operation as usual. However, the max-pooling indexes are stored for each feature map to be reused while the upsampling step. The first part in the encoder has four blocks which are the central block of the U-Net model. The second part has three blocks. The two parts in decoder worked simultaneously. The feature maps are concatenated after upsampling steps. The final layers have 1×1 convolution and the sigmoid activation function to create a segmentation map.

TernausNet, LinkNet34, ResNet-50, and SE-Resnet50 are widely used models that are explained clearly in their references. All models use the same Adam optimizer during the training stage. The loss functions of DualTail-Net and LinkNet34 networks are DICE loss while the others are average of DICE loss and binary cross-entropy. There is no pre- or post-processing in the system.

3.4.6.8 Mountain

Team Mountain participated in Tasks 3 and 5. They used a 3D model that uses U-Net in Han et al. (2019) as base architecture. This U-Net model has a different encoder part than the original U-Net in Ronneberger et al. (2015). The difference is that there is a residual block at each scale level in the encoder part. The other difference is the instance normalization (Ulyanov et al., 2017) instead of batch normalization. The reason for this choice is the robustness of instance normalization to changes of intensity in the image.

The summation of all levels in the decoder generates the output. Here, two networks (NET1 and NET2) following the model referred to above with separate channels and rates are used. NET1 is responsible for determining the organ roughly. It creates a mask of the region of interest where the organ and neighbor structures are located. Thus the spatial size can be decreased to improve performance and reduce computational cost/time. The output of NET1 is the input of NET2 which makes final segmentation. In both networks, Adam optimizer and DICE loss are preferred. For augmentation of data; rotation, deformation, and scaling are used.

3.4.6.9 CIR_MPerkonigg

CIR_MPerkonigg team targetted Task 3 in the challenge. Since there are multiple MRI sequences, the IVD-Net from Dolz et al. (2018) is adapted. IVD-Net has dense connections in the encoder part. However, these connections are not used because they do not bring any improvements according to trials. Also, residual convolutional blocks (He et al., 2016) are included.

The optimization method is selected as Adam optimizer. For regularization, the Modality Dropout (Li et al., 2016) method is adapted. Here modalities are omitted with a certain probability. That is how overfitting on specific modalities are prevented. Data augmentation is preferred to increase the number of images during training. The methods are elastic transformations, histogram shifting, affine transformations, and adding Gaussian noise.

3.4.6.10 nnU-Net

nnU-Net The nnU-Net team participated in Tasks 3 and 5 in the challenge. The developed model (Isensee et al., 2019) has been used in many challenges before and has been the first place in Medical Segmentation Decathlon (MSD) in 2018,(Simpson et al., 2019) as well as CHAOS Tasks 3 and 5.

Briefly, nnU-Net uses the ensemble of multiple networks. These networks are slightly modified variants of U-Net models with different parameters. First, all five networks are trained. Then the system selects three of them by using cross-validation on the training cases. In other words, the final results are an average of three 3D U-Nets ("3d_fullres" configuration of nnU-Net).

Table 3.11 Pre-processing, post-processing operations, and participated tasks in the CHAOS challenge

Team	**Pre-process**	**Post-process**	**Tasks**
OvGUMEMoRIAL	Training with resized images (128×128). Inference: full-sized.	Threshold by 0.5	1,2,3,4,5
ISDUE	Training with resized images (96,128,128)	Threshold by 0.5. Bicubic interpolation for refinement.	1,2,3,4,5
Lachinov	Resampling $1.4 \times 1.4 \times 2$ z-score normalization	Threshold by 0.5	1,2,3
IITKGP-KLIV	Training with resized images (256×256), whitening. Additional class for body.	Threshold by 0.5	1,2,3,4,5
METUMMLAB	Min-max normalization for CT	Threshold by 0.5. Connected component analysis.	1,3,5
PKDIA	Training with resized images: 256×256 MR, 512×512 CT.	Threshold by 0.5. Connected component analysis.	1,2,3,4,5
MedianCHAOS	LUT [-240,160] HU range, normalization.	Threshold by 0.5.	2
Mountain	Resampling $1.2 \times 1.2 \times 4.8$, zero padding. Training with resized images: $384 \times 384 \times 64$. Rigid register MR.	Threshold by 0.5. Connected component analysis for selecting/eliminating some of the model outputs.	3,5
CIRMPerkonigg	Normalization to zero mean unit variance.	Threshold by 0.5.	3
nnU-Net	Normalization to zero mean unit variance, Resampling $1.6 \times 1.6 \times 5.5$	Threshold by 0.5.	3, 5

Table 3.12 Brief comparison of participating methods in the CHAOS challenge

Team	**Details of the method**	**Training strategy**
OvGUMEMoRIAL (P. Ernst, S. Chatterjee, O. Speck, A. Nürnberger)	• Modified Attention U-Net, employing soft attention gates and multiscaled input image pyramid for better feature representation is used (Abraham & Khan, 2019). • Parametric ReLU activation is used instead of ReLU, where an extra parameter, i.e. coefficient of leakage, is learned during training.	• Tversky loss is computed for the four different scaled levels. • Adam optimizer is used, training is accomplished by 120 epochs with a batch size of 256.
ISDUE (D. D. Pham, G. Dovletov, J. Pauli)	• The proposed architecture consists of three main modules: i. Autoencoder net composed of a *prior encoder* f_{enc_p}, and decoder g_{dec}; ii. Hourglass net composed of an *imitating encoder* f_{enc_i}, and decoder g_{dec}; iii. U-Net module, i.e. h_{unet}, which is used to enhance the decoder g_{dec} by guiding the decoding process for better localization capabilities.	• The segmentation networks are optimized separately using the DICE-loss and regularized by $\mathcal{L}_{sc}$ with weight of $\lambda = 0.001$. • The autoencoder is optimized separately using DICE loss. • Adam optimizer with an initial learning rate of 0.001, and 2400 iterations are performed to train each model.
Lachinov (D. Lachinov)	• 3D U-Net, with skip connections between contracting/expanding paths and an exponentially growing number of channels across consecutive resolution levels (Lachinov, 2019). • The encoding path is constructed by a residual network for efficient training. • Group normalization (Wu & He, 2018) is adopted instead of batch (Ioffe & Szegedy, 2015) (# of groups = 4). • Pixel shuffle is used as an upsampling operator	• The network was trained with ADAM optimizer with learning rate 0.001 and decaying with a rate of 0.1 at 7th and 9th epoch. • The network is trained with batch size 6 for 10 epochs. Each epoch has 3200 iterations in it. • The loss function employed is DICE loss.

Table 3.12 continues

Team	Details of the method	Training strategy
IITKGP-KLIV (R. Sathish, R. Rajan, D. Sheet)	• To achieve multi-modality segmentation using a single framework, a multi-task adversarial learning strategy is employed to train a base segmentation network SUMNet (Nandamuri et al., 2019) with batch normalization. • Adversarial learning is performed by two auxiliary classifiers, namely C1 and C2, and a discriminator network D.	• The segmentation network and C2 are trained using cross-entropy loss while the discriminator D and auxiliary classifier C1 are trained by binary cross-entropy loss. • Adam optimizer. Input is the combination of all four modalities, i.e. CT, MRI T1 DUAL In and Oppose Phases, MRI T2 SPIR.
METU_MMLAB (S. Özkan, B. Baydar, G. B. Akar)	• A U-Net variation and a Conditional Adversarial Network (CAN) is introduced. • Batch Normalization is performed before convolution to prevent vanishing gradients and increase selectivity. • Parametric ReLU to preserve negative values using a trainable leakage parameter.	• To improve the performance around the edges, a CAN is employed during training (not as a post-process operation). • This introduces a new loss function to the system which regularizes the parameters for sharper edge responses.
PKDIA (P.-H. Conze)	• An approach based on Conditional Generative Adversarial Networks (cGANs) is proposed: the generator is built by cascaded pre-trained encoder-decoder (ED) networks extending the standard U-Net (Ronneberger et al., 2015) (VGG19, following (Conze et al., 2019)), with 64 channels generated by first convolutional layer. • After each max-pooling, channel number doubles until 512. The auto-context paradigm is adopted by cascading two EDs (Yan et al., 2019): the output of the first is used as features for the second.	• Adam optimizer with a learning rate of 10^{-5} is used. • Fuzzy DICE score is employed as a loss function. • Batch size was set to 3 for CT and 5 for MR scans.

Table 3.12 continues

Team	Details of the method	Training strategy
MedianCHAOS (V. Groza)	• Averaged ensemble of five different networks is used. The first one is *DualTail-Net* that is composed of an encoder, central block and 2 dependent decoders. • Other four networks are U-Net variants, i.e. TernausNet (U-Net with VGG11 backbone (Iglovikov & Shvets, 2018)), LinkNet34 (Shvets et al., 2018), and two with ResNet-50 and SE-Resnet50.	• The training for each network was performed with Adam. • DualTail-Net and LinkNet34 were trained with soft DICE loss and the other three networks were trained with the combined loss: 0.5*soft DICE + 0.5*BCE (binary cross-entropy).
Mountain (Shuo Han)	• 3D network adopting U-Net variant in (Han et al., 2019) is used. Two nets, i.e. NET1 and NET2, adopting (Han et al., 2019) with different channels and levels. NET1 locates organ and outputs a mask for NET2 performing finer segmentation.	• Adam optimizer is used with the initial learning rate = 1×10^{-3}, β_1 = 0.9, β_2 = 0.999, and $\epsilon = 1 \times 10^{-8}$. • DICE coefficient was used as the loss function. Batch size was set to 1.
CIRMPerkonigg (M. Perkonigg)	• For joint training with all modalities, the IVD-Net (Dolz et al., 2018) (which is an extension of U-Net (Ronneberger et al., 2015)) is used with a number of modifications: (i) dense connections between encoder path of IVD-Net are not used since no improvement is achieved. (ii) training images are split. • Moreover, residual convolutional blocks (He et al., 2016) are used.	• Modality Dropout (Li et al., 2016) is used as the regularization technique when the training is performed using multiple modalities which help to decrease over-fitting on certain modalities. • Training is done by using Adam optimizer with a learning rate of 0.001 for 75 epochs.
nnU-Net (F. Isensee, K. H. Maier-Hein)	• An internal variant of nnU-Net (Isensee et al., 2019), which is the winner of Medical Segmentation Decathlon (MSD) in 2018 (Simpson et al., 2019), is used. • Ensemble of five 3D U-Nets ("3d_fullres" configuration), which originate from cross-validation on the training cases. Ensemble of T1 in and oppose phases was used.	• T1 in and out are treated as separate training examples, resulting in a total of 60 training examples for the tasks. • Task 3 is a subset of Task 5, so training was done only once and the predictions for Task 3 were generated by isolating the liver.

3.4.7 Results

The CHAOS challenge has started as a part of the IEEE International Symposium on Biomedical Imaging (ISBI) on April 11, 2019, Venice, Italy. The training dataset was shared globally three months prior to ISBI 2019. The test dataset was provided 24 hours prior to the challenge. At the end of the session, the submitted results were analyzed and the winners were declared. After the on-site event, both training and test datasets were uploaded on zenodo.org website (Kavur et al., 2019). Then, the online submission system was opened on the challenge website.

The total submission numbers are presented in Table 3.13. According to the table, the tasks with more typical aims are more popular as predicted. In the following sections, the top results in the tasks are reviewed according to their popularity. Thus, the scores achieved for more conventional approaches (Tasks 2, 3, and 5), guided the discussions of multi-modality/organ concepts (Tasks 1 and 4).

Table 3.13 CHAOS challenge submission statistics for on-site and online sessions

	Task 1	Task 2	Task 3	Task 4	Task 5	Total
On-site (ISBI 2019)	5	14	7	4	5	35
Online	27	312	91	22	120	572

On the challenge website, there are two individual scoreboards, one for the on-site and one for online submissions. We prepared detailed analyses on these results. The majority of the results in this section are coming from all on-site results among with some remarkable results from online submissions. The reason for this choice is that we would like to guarantee the fairness of the results here. In other words, there was no 'peeking' attempt on these results. Therefore, it is possible to use them to guide researches in the medical imaging field. In the following tables and figures, each team has a unique color code to make following their results easy while reading.

All case results of submission are used to generate box plots of them in Figure 3.14 while individual scores of cases in the tasks are presented in Figure 3.15. Also all

metric results and scores are presented in Table 3.14. These tables and plots give us the opportunity of analyzing the distribution, mean, and median of the submissions which is not possible with using just average task scores.

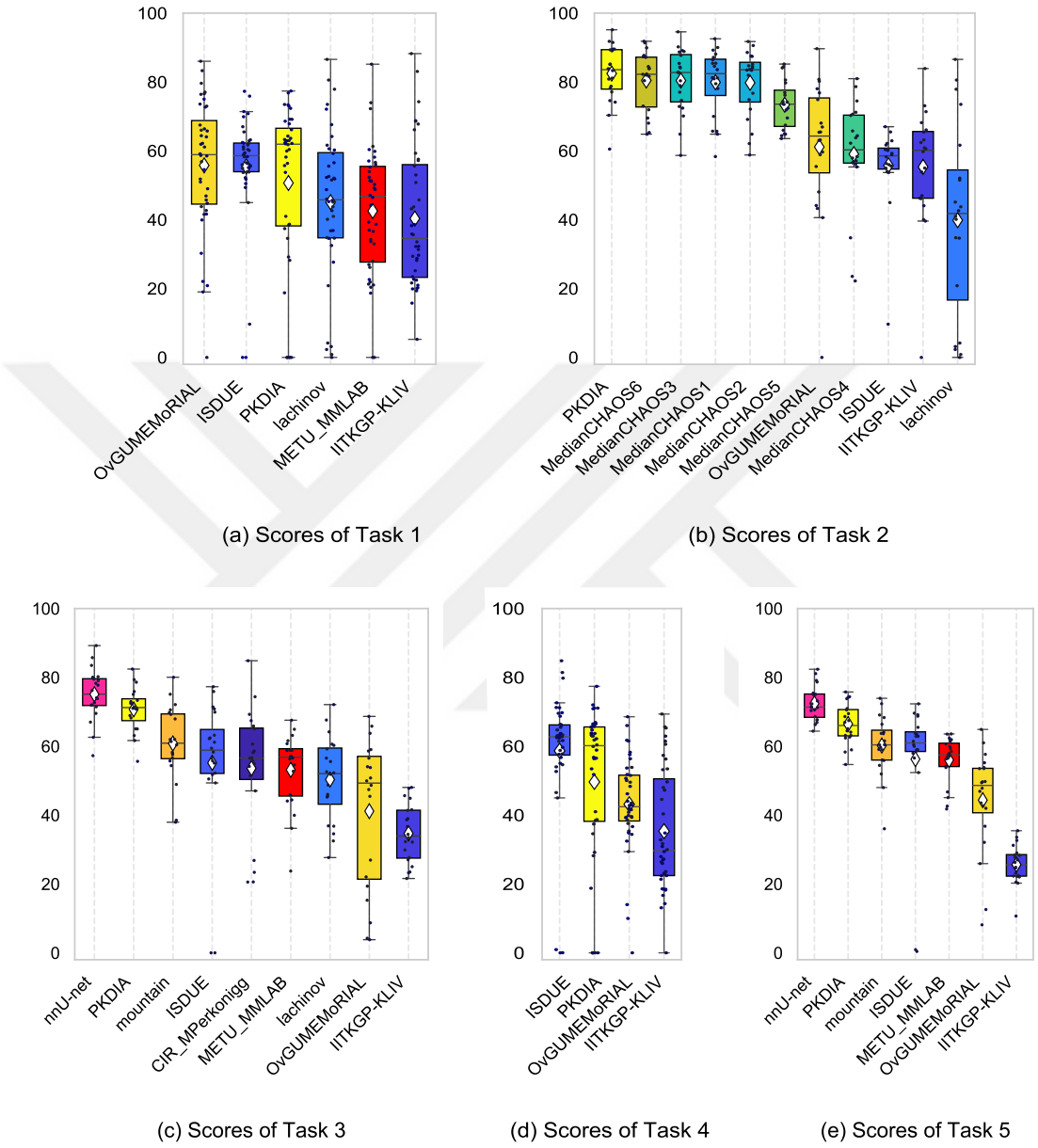

Figure 3.14 Box plot of results for each task. White diamonds represent the mean values of the scores. Solid vertical lines inside of the boxes represent medians. Separate dots show scores of each individual case (Kavur et al., 2020a)

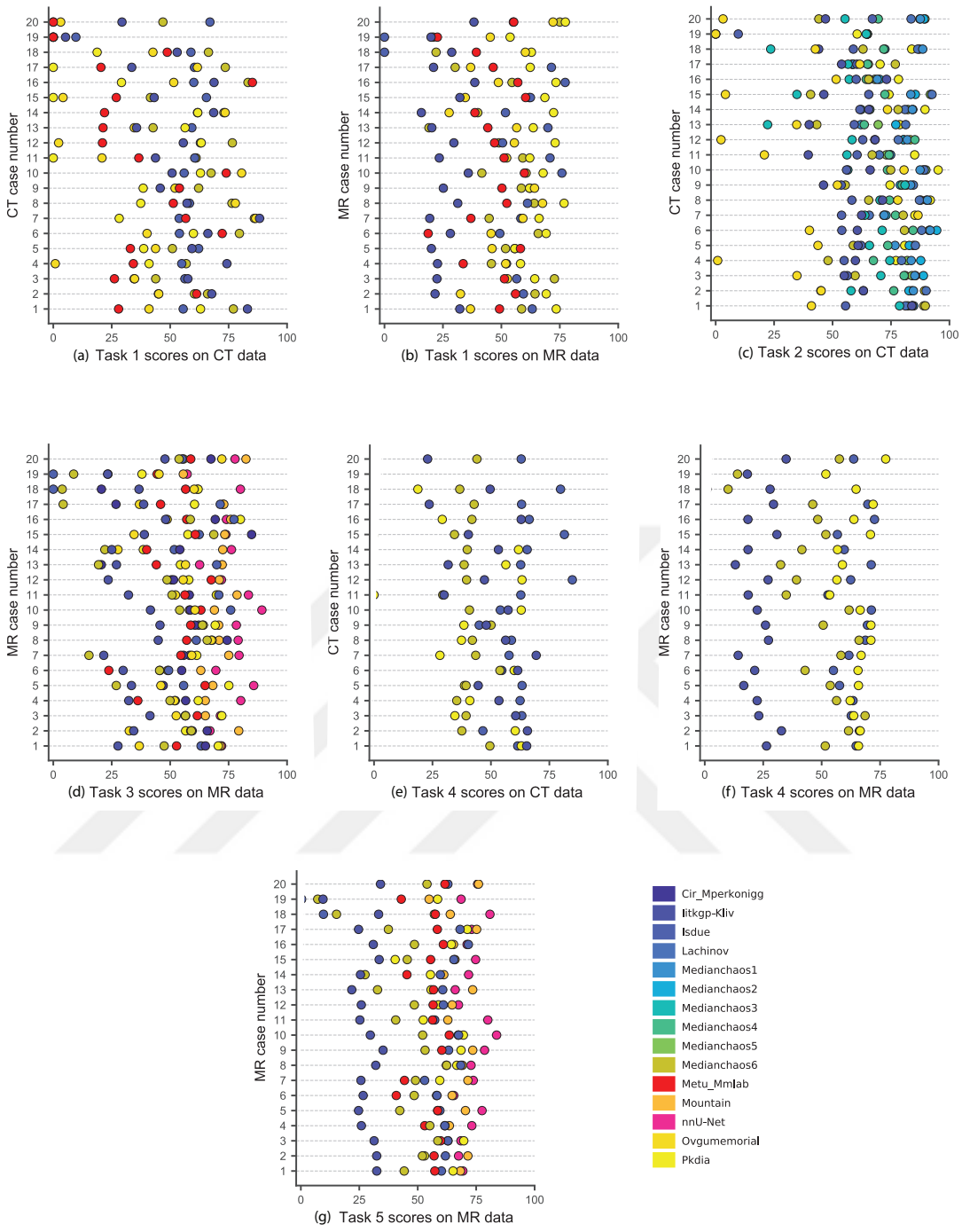

Figure 3.15 Distribution of the methods' scores over the cases in test data (Kavur et al., 2020a)

Table 3.14 Metric values and corresponding scores of submissions. The given values represent the average of all cases and all organs of the related tasks in the test data (The best results are given in bold)

	Team Name	**Mean Score**	DICE	DICE Score	RAVD	RAVD Score	ASSD	ASSD Score	MSSD	MSSD Score
Task 1	● OvGUMEMoRIAL	**55.78 ± 19.20**	0.88 ± 0.15	83.14 ± 28.16	13.84 ± 30.26	**24.67 ± 31.15**	11.86 ± 65.73	**76.31 ± 21.13**	57.45 ± 67.52	31.29 ± 26.01
	● ISDUE	55.48 ± 16.59	0.87 ± 0.16	83.75 ± 25.53	12.29 ± 15.54	17.82 ± 30.53	5.17 ± 8.65	75.10 ± 22.04	36.33 ± 21.97	44.83 ± 21.78
	● PKDIA	50.66 ± 23.95	0.85 ± 0.26	**84.15 ± 28.45**	6.65 ± 6.83	21.66 ± 30.35	9.77 ± 23.94	75.84 ± 28.76	46.56 ± 45.02	**42.28 ± 27.05**
	● Lachinov	45.10 ± 21.91	0.87 ± 0.13	77.83 ± 33.12	10.54 ± 14.36	21.59 ± 32.65	7.74 ± 14.42	63.66 ± 31.32	83.06 ± 74.13	24.30 ± 27.78
	● METU_MMLAB	42.54 ± 18.79	0.86 ± 0.09	75.94 ± 32.32	18.01 ± 22.63	14.12 ± 25.34	8.51 ± 16.73	60.36 ± 28.40	62.61 ± 51.12	24.94 ± 25.26
	● IITKGP-KLIV	40.34 ± 20.25	0.72 ± 0.31	60.64 ± 44.95	9.87 ± 16.27	24.38 ± 32.20	11.85 ± 16.87	50.48 ± 37.71	95.43 ± 53.17	7.22 ± 18.68
Task 2	● *PKDIA**	***82.46 ± 8.47***	*0.98 ± 0.00*	***97.79 ± 0.43***	*1.32 ± 1.302*	***73.6 ± 26.44***	*0.89 ± 0.36*	***94.06 ± 2.37***	*21.89 ± 13.94*	***64.38 ± 20.17***
	● MedianCHAOS6	80.45 ± 8.61	0.98 ± 0.00	97.55 ± 0.42	1.54 ± 1.22	69.19 ± 24.47	0.90 ± 0.24	94.02 ± 1.6	23.71 ± 13.66	61.02 ± 21.06
	● MedianCHAOS3	80.43 ± 9.23	0.98 ± 0	97.59 ± 0.44	1.41 ± 1.23	71.78 ± 24.65	0.9 ± 0.27	94.02 ± 1.79	27.35 ± 21.28	58.33 ± 21.74
	● MedianCHAOS1	79.91 ± 9.76	0.97 ± 0.01	97.49 ± 0.51	1.68 ± 1.45	66.8 ± 28.03	0.94 ± 0.29	93.75 ± 1.91	23.04 ± 10	61.6 ± 16.67
	● MedianCHAOS2	79.78 ± 9.68	0.97 ± 0	97.49 ± 0.47	1.5 ± 1.2	69.99 ± 23.96	0.99 ± 0.37	93.39 ± 2.48	27.96 ± 23.02	58.23 ± 20.27
	● MedianCHAOS5	73.39 ± 6.96	0.97 ± 0	97.32 ± 0.41	1.43 ± 1.12	71.44 ± 22.43	1.13 ± 0.43	92.47 ± 2.87	60.26 ± 50.11	32.34 ± 26.67
	● OvGUMEMoRIAL	61.13 ± 19.72	0.90 ± 0.21	90.18 ± 21.25	$9x10^3 \pm 4x10^3$	44.35 ± 35.63	4.89 ± 12.05	81.03 ± 20.46	55.99 ± 38.47	28.96 ± 26.73
	● MedianCHAOS4	59.05 ± 16	0.96 ± 0.02	96.19 ± 1.97	3.39 ± 3.9	50.38 ± 33.2	3.88 ± 5.76	77.4 ± 28.9	91.97 ± 57.61	12.23 ± 19.17
	● ISDUE	55.79 ± 11.91	0.91 ± 0.04	87.08 ± 20.6	13.27 ± 7.61	4.16 ± 12.93	3.25 ± 1.64	78.30 ± 10.96	27.99 ± 9.99	53.60 ± 15.76
	● IITKGP-KLIV	55.35 ± 17.58	0.92 ± 0.22	91.51 ± 21.54	8.36 ± 21.62	30.41 ± 27.12	27.55 ± 114.04	81.97 ± 21.88	102.37 ± 110.9	17.50 ± 21.79
	● Lachinov	39.86 ± 27.90	0.83 ± 0.20	68 ± 40.45	13.91 ± 20.4	22.67 ± 33.54	11.47 ± 22.34	53.28 ± 33.71	93.70 ± 79.40	15.47 ± 24.15

* Corrected submission of PKDIA right after the on-site session (i.e. During the challenge, they have submitted the same results, but in reversed orientation. Therefore, the winner of Task 2 at conference session is the MedianCHAOS6).

Table 3.14 continues

	Team Name	**Mean Score**	DICE	DICE Score	RAVD	RAVD Score	ASSD	ASSD Score	MSSD	MSSD Score
Task 3	nnU-Net	**75.10 ± 7.61**	0.95 ± 0.01	**95.42 ± 1.32**	2.85 ± 1.55	**47.92 ± 25.36**	1.32 ± 0.83	**91.19 ± 5.55**	20.85 ± 10.63	**65.87 ± 15.73**
	PKDIA	70.71 ± 6.40	0.94 ± 0.01	94.47 ± 1.38	3.53 ± 2.14	41.8 ± 24.85	1.56 ± 0.68	89.58 ± 4.54	26.06 ± 8.20	56.99 ± 12.73
	Mountain	60.82 ± 10.94	0.92 ± 0.02	91.89 ± 1.99	5.49 ± 2.77	25.97 ± 27.95	2.77 ± 1.32	81.55 ± 8.82	35.21 ± 14.81	43.88 ± 17.60
	ISDUE	55.17 ± 20.57	0.85 ± 0.19	82.08 ± 28.11	11.8 ± 15.69	24.65 ± 27.58	6.13 ± 10.49	73.50 ± 25.91	40.50 ± 24.45	40.45 ± 20.90
	CIR_MPerkonigg	53.60 ± 17.92	0.91 ± 0.07	84.35 ± 19.83	10.69 ± 20.44	31.38 ± 25.51	3.52 ± 3.05	77.42 ± 18.06	82.16 ± 50	21.27 ± 23.61
	METU_MMLAB	53.15 ± 10.92	0.89 ± 0.03	81.06 ± 18.76	12.64 ± 6.74	10.94 ± 15.27	3.48 ± 1.97	77.03 ± 12.37	35.74 ± 14.98	43.57 ± 17.88
	Lachinov	50.34 ± 12.22	0.90 ± 0.05	82.74 ± 18.74	8.85 ± 6.15	21.04 ± 21.51	5.87 ± 5.07	68.85 ± 19.21	77.74 ± 43.7	28.72 ± 15.36
	OvGUMEMoRIAL	41.15 ± 21.61	0.81 ± 0.15	64.94 ± 37.25	49.89 ± 71.57	10.12 ± 14.66	5.78 ± 4.59	64.54 ± 24.43	54.47 ± 24.16	25.01 ± 20.13
	IITKGP-KLIV	34.69 ± 8.49	0.63 ± 0.07	46.45 ± 1.44	6.09 ± 6.05	43.89 ± 27.02	13.11 ± 3.65	40.66 ± 9.35	85.24 ± 23.37	7.77 ± 12.81
Task 4	ISDUE	**58.69 ± 18.65**	0.85 ± 0.21	81.36 ± 28.89	14.04 ± 18.36	14.08 ± 27.3	9.81 ± 51.65	78.87 ± 25.82	37.12 ± 60.17	55.95 ± 28.05
	PKDIA	49.63 ± 23.25	0.88 ± 0.21	**85.46 ± 25.52**	8.43 ± 7.77	**18.97 ± 29.67**	6.37 ± 18.96	**82.09 ± 23.96**	33.17 ± 38.93	**56.64 ± 29.11**
	OvGUMEMoRIAL	43.15 ± 13.88	0.85 ± 0.16	79.10 ± 29.51	$5x10^3 \pm 5x10^4$	12.07 ± 23.83	5.22 ± 12.43	73.00 ± 21.83	74.09 ± 52.44	22.16 ± 26.82
	IITKGP-KLIV	35.33 ± 17.79	0.63 ± 0.36	50.14 ± 46.58	13.51 ± 20.33	15.17 ± 27.32	16.69 ± 19.87	40.46 ± 38.26	130.3 ± 67.59	8.39 ± 22.29
Task 5	nnU-Net	**72.44 ± 5.05**	0.95 ± 0.02	**94.6 ± 1.59**	5.07 ± 2.57	**37.17 ± 20.83**	1.05 ± 0.55	**92.98 ± 3.69**	14.87 ± 5.88	**75.52 ± 8.83**
	PKDIA	66.46 ± 5.81	0.93 ± 0.02	92.97 ± 1.78	6.91 ± 3.27	28.65 ± 18.05	1.43 ± 0.59	90.44 ± 3.96	20.1 ± 5.90	66.71 ± 9.38
	Mountain	60.2 ± 8.69	0.90 ± 0.03	85.81 ± 10.18	8.04 ± 3.97	21.53 ± 15.50	2.27 ± 0.92	84.85 ± 6.11	25.57 ± 8.42	58.66 ± 10.81
	ISDUE	56.25 ± 19.63	0.83 ± 0.23	79.52 ± 28.07	18.33 ± 27.58	12.51 ± 15.14	5.82 ± 11.72	77.88 ± 26.93	32.88 ± 33.38	57.05 ± 21.46
	METU_MMLAB	56.01 ± 6.79	0.89 ± 0.03	80.22 ± 12.37	12.44 ± 4.99	15.63 ± 13.93	3.21 ± 1.39	79.19 ± 8.01	32.70 ± 9.65	49.29 ± 12.69
	OvGUMEMoRIAL	44.34 ± 14.92	0.79 ± 0.15	64.37 ± 32.19	76.64 ± 122.44	9.45 ± 11.98	4.56 ± 3.15	71.11 ± 18.22	42.93 ± 17.86	39.48 ± 16.67
	IITKGP-KLIV	25.63 ± 5.64	0.56 ± 0.06	41.91 ± 11.16	13.38 ± 11.2	11.74 ± 11.08	18.7 ± 6.11	35.92 ± 8.71	114.51 ± 45.63	11.65 ± 13.00

3.4.7.1 CT Liver Segmentation (Task 2)

This task contains one of the most researched segmentation methods for the liver. Hence, it offers a proper chance to assess the effectiveness of the methods with the previous strategies. The contrast agent in the portal veins may create difficulty for segmenters. However, Table 3.14 and Figure 3.14b shows that the models reached the highest scores as expected in this task.

In this task, two submissions have outstanding segmentation accuracy. Team MedianCHAOS was the winner of the challenge with its 6th submission (MedianCHAOS6). Their score is 80.45 ± 8.61. Method of MedianCHAOS uses the ensemble strategy of multiple sub-networks whose performance is illustrated in Figure 3.15c. Following this, PKDIA placed first in the online session with 82.46±8.47 points. When we examine the metrics in more detail, DICE scores of both submission is very high (PKDIA:97.79 ± 0.43, MedianCHAOS:97.55 ± 0.42). ASSD scores show similar performance with 0.89 ± 0.36mm for PKDIA and 0.90 ± 0.24mm for MedianCHAOS. However, the results of RAVD and MSSD metrics have poor values than the other two metrics. On the other hand, this situation can be considered the same for all submissions.

If we compare results of this task with outcomes from our previous challenge (Kavur et al., 2020b), again DMs have outperformed semi-automatic approaches such as active contours, robust static segmenter, and watershed. Thus, the results of CHAOS Task 2 approve our findings in our first challenge. The performance of the top methods reached inter-expert level in terms of DICE and ASSD metrics. On the other hand, there is still a need for improvements for RAVD and MSSD which are related to maximum error margins. Despite huge developments, DMs may have still local problems such as near inferior vena cava shown in Figure 3.16.

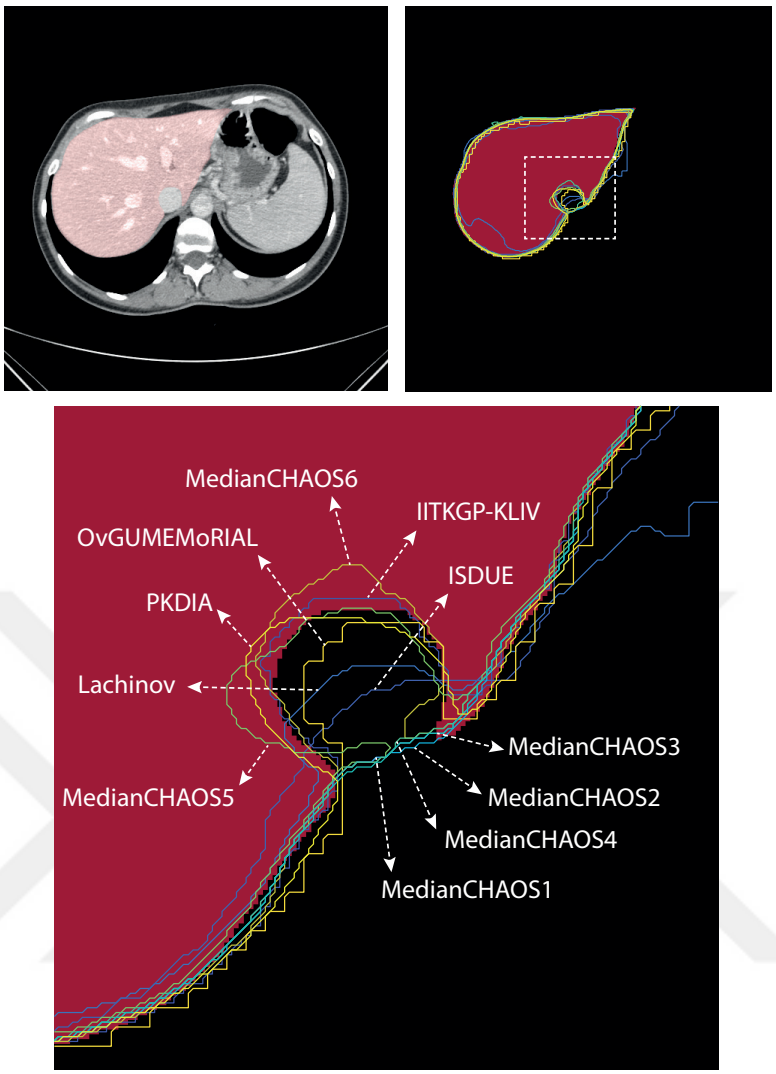

Figure 3.16 Top left: Example image from CHAOS CT set, case 35, slice 95. Top right: borders of segmentation results on ground truth mask. Bottom: zoomed onto inferior vena cava (IVC) region (marked with dashed lines on the top right image) respectively. Since the contrast between liver tissue and IVC is relatively lower due to timing error during the CT scan, algorithms mostly mistakes here. On the other hand, many of them are quite successful at the other regions of the liver

3.4.7.2 MRI Liver Segmentation (Task 3)

Despite the dominance of CT scans in abdominal imaging, the drawbacks of MRI are being eliminated day-by-day. Still, segmentation from MRI scans is a relatively harder problem than Task 2. Not having standardized values for tissues compared to CT (pre-defined Hounsfield units) makes segmentation from MRI scans more difficult. Another disadvantage of MRI against CT is the lower resolution with higher slice

thickness. Because of these problems, the scores in Task 3 are lower than Task 2 even though they both target the same organ: liver. Details of the scores are shown in Figures 3.14c and 3.15d.

Team PKDIA was the winner of the task with a 70.71 ± 6.40 score. In addition, this score has the least standard deviation over all individual case scores. Therefore, it is possible to say that, the method of PKDIA has precise outputs over different cases. In other words, this method can be considered as 'robust' for this task. For example, team CIR_MPerkonigg has higher scores than PKDIA in particular cases, but overall its performance is less successful.

Model based on nnU-Net is the winner of the online session of the challenge with 75.10 ± 7.61 score. If we compare two winners against all metrics, DICE (PKDIA:0.94 ± 0.01 and nnU-Net:0.95 ± 0.01) and ASSD (PKDIA:1.56 ± 0.68mm and nnU-Net:1.32 ± 0.83mm) scores are remarkable. On the other hand, there is a similar situation for RAVD and MSSD as expected. In addition to the toughness of these metrics, low resolution and higher spacing of the MR data cause higher spatial errors.

The results revealed that the gap between interactive models and DMs is getting closer to segmentation from MRI scans. It can be said that there is a significant potential for DMs in this type of task. Further developments on DL studies as well as MRI technology will make reaching accurate segmentation results possible in the future.

3.4.7.3 CT-MR Liver Segmentation (Task 1)

This task combines all difficulties in Tasks 2 and 3 with its cross-modality data. A single solution must handle all problems in various modalities to obtain useful segmentations. Figure 3.14a presents mean score distributions of the task. Also, scores for CT and MRI can be analyzed from Figure 3.15a-b separately. The first inference from the figures is that DMs trained on single modalities achieve

significantly better performance than DMs trained on cross-modality data. Given the difficulties of the task, these results are not surprising.

The on-site session of this task was won by team OvGUMEMoRIAL with 55.78±19.20 score. In contrast with their promising DICE score (0.88 ± 0.15), the other metrics have relatively fair performance. Since OvGUMEMoRIAL participated in all tasks in the challenge, we can compare their results in single modality tasks (Tasks 2 and 3) with Task 1. The interesting fact is that their score (55.78 ± 19.20) in this task higher than their MRI segmentation score in Task 3 (41.15 ± 21.61). This disproves that cross-modality studies always perform poorly than single modality studies. A further important deduction comes from team PKDIA. Despite their remarkable results in Task 1 and Task 2, their performance dropped critically. This fact shows that additional solutions are needed to develop in cross-modality problems even if successful performances on individual modalities.

Even though the ranking of the challenge is handled via mean case scores, it is important to analyze the distribution of the individual case scores. In this way, we can focus on generalization capabilities and using them in real-life applications. For instance, there is an noticeable case in Figure 3.14a. The winner submission of OvGUMEMoRIAL has fair performances than the second winner method, ISDUE if we focus on the standard deviation of the case scores. Considering scattering of the scores shown in Figures 3.15a and 3.15b, the results of ISDUE are more precise (but less accurate) than results of OvGUMEMoRIAL especially on CT data. Another ranking method based on the standard deviation of the cases can generate different scoreboards in the task.

Since cross-modality segmentation is a newly developing area, it is not possible to compare these findings with other studies yet. However, all analyses reveal that DMs have significant performances to solve problems in this field despite having less accurate results than single modality data. Additional studies in the feature may solve these problems.

3.4.7.4 Multi-Modal MR Abdominal Organ Segmentation (Task 5)

Task 5 is the extended version of Task 3 over four abdominal organs; liver, kidneys, and spleen. Task 5 explores how DMs help on develop more comprehensive anatomical models that lead to tasks involving multiple organs. DMs can accurately reflect the complex and versatile abdominal anatomy by integrating inter-organ relationships through the internal hierarchical method of extracting features. As previously explained tasks, two successful teams share the first place on-site and online sessions. The on-site winner was PKDIA with a score of 66.46 ± 0.81 and the online winner was nnU-Net with a score of 72.44 ± 5.05. The DICE results seem to be almost similar for nnU-Net and PKDIA if the ratings of individual metrics are compared to Task 3. It is an important finding since instead of only one, all four organs are segmented in this task. Another model, Mountain, has almost exactly the same average score in Task 3 and Task 5 that is worthy of mention as well.

With respect to RAVD, the drop in performance is evidently higher than DICE. The decreased performance of DICE and RAVD is partially offset by better MSSD and ASSD performance. It should be remembered that the liver in Task 3 and the other organs in Task 5 can usually be considered to be comparatively simpler to examine than other abdominal structures. However, the important finding here is that single and multiple organ segmentation does not significantly change the performance of DMs. Even the current versions of DMs have a promising performance for the problem of segmentation of multiple structures. General comparison of all DMs in this task are presented in Figure 3.17.

3.4.7.5 CT-MR Abdominal Organ Segmentation (Task 4)

This task was designed to push DMs to their boundaries to analyze the last sophisticated developments in the Deep Learning field. Since it combines all difficulties of multiple organ segmentation and cross-modality learning, this is the most difficult task in the CHAOS challenge. Hence, it is no surprise that the scores are

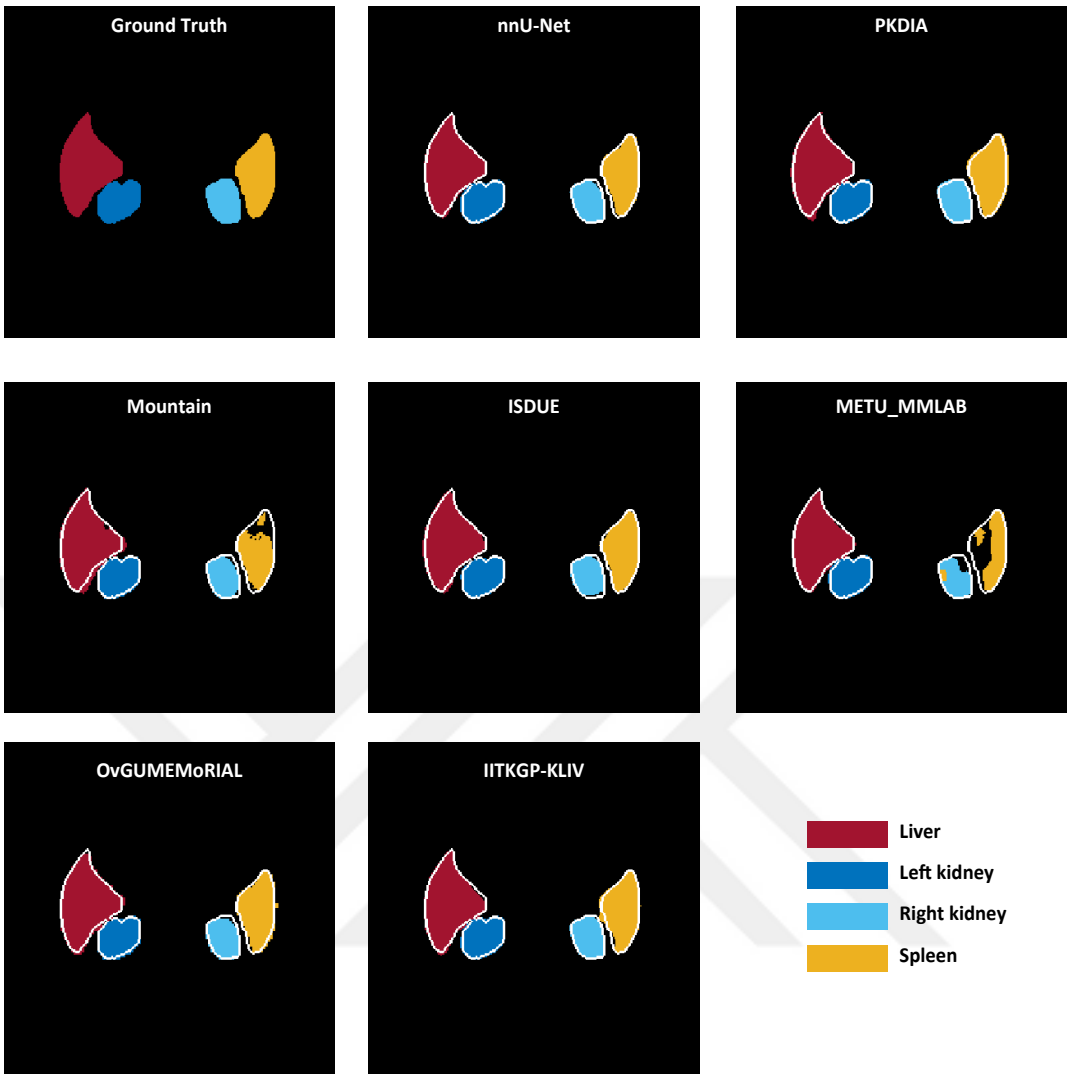

Figure 3.17 Illustration of ground truth and all results in the task 5. (The image was taken from CHAOS MR set, case 40, slice 15. White lines on the results represent borders of ground truths)

the lowest (Figure 3.15e-f). On the other hand, any high score is also worthwhile considering the toughness of the task.

The task was won by team ISDUE with 58.69 ± 18.65 score at the on-site session. The findings indicate that the performance of their model was spreaded CT and MR data consistently. Two convolutionary encoders in their model can be thought to boost performance on cross-modality data. These encoders can compress anatomical details. The second most successful submission came from team PKDIA with a score of 49.63 ± 23.25. The performance on CT data can be considered unsatisfactory with respect to the success of MRI sets in conjunction with the situation on Task 1. This shows

that the training procedure of their CNN might not be effective. Transfer learning and a pre-trained weights approach are perhaps not successful in a variety of ways in the encoder part of the system. With the average ranking, OvGUMEMoRIAL has reached the third place and has balanced score distributions between CT and MR data. In terms of generalization, their approach can be considered as efficient because of not having outlier results.

In combination with the results of Tasks 1 and 5, CNNs are shown better segmentation efficiency on single-modality tasks by their current strategies and architectures. This could be regarded as a normal outcome, as the effectiveness of CNNs depends heavily on the coherence and homogeneity of the data. The use of multiple modalities produces a significant variation in the results. On the other hand, the findings also show that, if suitable models are developed, CNNs have great potential for cross-modality studies.

To sum up, the results of the CHAOS challenge are presented with important analyses and reviews. The unregistered multi-modality (CT-MR), multi-sequence (T1-DUAL in / oppose, and T2-SPIR), public, and novel data set were created for five challenging tasks. A substantial range of well known and state-of-the-art segmentation methods was evaluated.

Except for one, the alteration of the U-Net as a primary or supportive system has been employed by all teams. Nevertheless, while there is the same basic CNN structure, the high variance between the scores also relies on several parameters in DMs. While many common algorithmic characteristics can be extracted from well-performed models, it is not easy to interpret and/or explain why a specific model works better or not. As stated before, these analyzes are practically beyond the bounds of possibility on a heterogeneous number of models produced by various participants. In addition, choices in the ranking method may have significant consequences for the reporting results (Maier-Hein et al., 2018).

In conclusion, the major findings display some significant results:

1. DMs have surpassed definitely semi-automatic techniques for the segmentation of liver from CT scans. Inter-expert alternation in DICE and volumetry has been achieved, but more improvements in distance metrics are still required which are vital in evaluating surgical error allowances.

2. Task 2 (liver segmentation from CT scans), has received more than 300 submissions since the beginning of the challenge. Quantitative and qualitative analyzes indicate that CNNs have almost accomplished clinically acceptable results. Given the excellent outcomes for single modality segmentation studies, it can be inferred that some minimal optimizations, particularly in the post-processing stage, can create clinically acceptable segmentation.

   Therefore, it can be considered that Task 2 solutions have become saturated. It may not be beneficial for the effort to establish minor improvements. We recommend that researchers should concentrate more on applying their models for real-world applications rather than trying to achieve minimum score improvements. Reducing computational costs, improving generalization, and creating easy to implement solutions are remaining problems of DMs for real-world utilization.

3. Taking into account the segmentation of the liver from MRI, the participating DMs were closely good for DICE but lacked efficiency for distance-based metrics such as ASSD, MSSD.

4. The performance of DMs is increased compared with liver segmentation if we consider all four abdominal organs. Nevertheless, it is difficult to know if this change can be applied to the segmentation of multi structures because the liver can be considered the most complex one to segment.

5. Segmentations from cross-modality is yet a difficult task for DMs. Further studies are necessary to implement them in clinical usage.

6. The most challenging problem remains the segmentation of multiple structures from cross-modality. Including the spatial and/or topological features or adding shape models in the loss functions can be used in DM designs.

7. Despite using the same base DM (U-Net), the decisions on design stages can tremendously change the segmentation accuracy. Therefore it is not still possible to suggest a single DM solution for all problems. One exception can be considered as using ensembles of multiple models. 3 of 5 tasks winner methods used an ensemble of multi DMs. **Therefore, ensembles can be regarded as a solution for the drawbacks of segmentation problem.**

CHAPTER FOUR
FUSION OF DIFFERENT METHODS FOR SEGMENTATION OF THE LIVER

4.1 Introduction

Outcomes of CHAOS showed one more time the weakness of the DMs for creating reproducible and general solutions. Even using the same base model in the design stage, the segmentation performances are distributed widely. Even the methods regarded as state-of-the-art, are actually prepared for specific data with performing many iterative training sessions. DMs need to be carefully designed, tuned, and trained with consuming extensive time. That is why there are many proposed DM based solutions in the literature but very few of them have real-life implementations. In other words, unfortunately, the success of theoretically the best DMs over limited data may not mean any sense for solving real-world problems in the medical image processing field.

An example of the generalization problem in DMs can be considered the design of a race car that can reach very high speeds only on specific tracks. However, such a car is not suitable for general use. On the other hand, a car that cannot reach such high speeds but can operate in all conditions is a more valuable vehicle for the general human population. The main goal of this thesis is the same as this example: Creating a reachable and generalized solution with a reasonable sacrifice of performance. Therefore, we created a model to use the potential of DMs without making huge design efforts. We decided to include four well-studied DMs for medical image segmentation with their default (vanilla-style) designs and parameters. In other words, multiple DMs were collected from their source and they were not modified. Thus, the need for high expertise on DL is not required. It eliminates any parameters from being modified, the structure changed or a new training strategy planned. We tried previously proposed ensemble methods in the literature (Kuncheva, 2014). After that, we developed a unique ensemble strategy for this problem to reach more

accurate segmentation results.

This section is structured as follows: the first two datasets that were used for design and testing are presented. Then the previous ensemble methods in the literature, as well as our solution, are explained. Finally, the results of our proposed ensemble strategies are discussed in Section 4.6 in the last chapter of the thesis.

4.2 Datasets

The target of the study is selected as segmentation of the liver from CT scans. This is the same target as Task 2 in the CHAOS challenge. The reason for this choice is that the studies in this area are the most popular ones. Therefore, it will be possible to make further analyses and comparisons with other methods due to the high competition in this field.

We carried out the study on two publicly accessible datasets that were released at different times. They have specific properties that are used to show the accuracy of the proposed algorithms. The first set is the CT part of the CHAOS data. Since the details are explained in Section 3.4.2, there is no additional description for this dataset in this section. The second dataset is called 3DIRCADB1. The details of this data are explained in the following section. Comparison of both sets are summarized in Table 4.1.

4.2.1 3DIRCADB1 Data

3DIRCADB1 (3D Image Reconstruction for Comparison of Algorithm Database) (IRCAD, 2009) contains CT abdominal scans of 20 patients. In comparison to CHAOS CT data, in 75% of cases, hepatic cancers are found in the liver. Clinical professionals annotated the data for reference segmentations. All structures included in the liver were used as a segmentation target, except for tumors. 3DIRCADB1 was split 50%-50% for training and testing.

Table 4.1 Specifications of CHAOS CT and 3DIRCADB1 datasets

	3DIRCAD	CHAOS
Number of 3D image sets (train and test)	20 (10 + 10)	40 (20 + 20)
Spatial resolution of files	512 x 512	512 x 512
Number of files (slices) in all cases [min–max]	[74 – 260]	[78 – 294]
Average number of files in the cases	141	160
Total number of files in the dataset	2823	6407
X space (mm) [min–max]	[0.56 – 0.87]	[0.54 – 0.79]
Y space (mm) [min–max]	[0.56 – 0.87]	[0.54 – 0.79]
Slice thickness (mm) [min–max]	[1.60 – 4.00]	[2.00 – 3.20]

4.3 Ensemble Members

It is possible to use any segmentation method in the community process. For example, the usage of fully automatic and semi-automatic approaches is possible. Besides, combining both semi-automatic and full automatic methods would reach the highest success because the ensemble of them will tolerate errors of other methods. However, we did not choose to use semi-automatic approaches. The reasons for these decisions are 1) after the tremendous developments in the Deep Learning field, Deep Learning-based segmenters clearly outperformed interactive segmentation methods, 2) using semi-automatic methods will significantly reduce reproducibility and generalizability of the proposed method. However, our target is to improve these features. Hence, only fully automatic methods, specifically DMs, were used as ensemble members.

Even though using more diverse ensemble members can improve the accuracy, there are not so many possible DMs for medical image segmentation in the literature. The

base segmentation frameworks were selected from the most popular and well-studied DMs: U-Net, DeepMedic, V-Net, Dense V-Networks. The details of these models were explained in Section 2.5.2. All of these DMs were directly downloaded from their original source. They were used with their default parameters (vanilla style) on both datasets. Then, the inference stage was handled by individual DMs on test data. Each of them generated probability maps with the same size as the input image. These probability maps have distribution in the range of $[0, 1]$ that indicates the location of the target structure(liver). The values closer to 1, show that the probability of finding a voxel belongs to the liver is higher than finding a voxel from background class. An example set of probability maps coming from four models in this thesis is presented in Figure 4.1. Also the ensemble strategy is illustrated in Figure 4.2.

4.4 Ensemble Methods

There are various ensemble strategies for various data and tasks (Rokach, 2010; Kuncheva, 2014). The main factor for selecting an ensemble method is the size and diversity of data as well as the segmentation target of interest. Some of the ensemble methods require training for defined parameters while some of them do not need training. There are also more complex ensemble methods like AdaBoost that progress the fusion method while training of various independent segmenters.

The biggest drawback of developing an ensemble method is small data size. Where big data is available, it is possible to train and test non-intersecting subsets that ensure fine-tuning and avoid overfitting. However, this kind of approach is very hard to implement on small data. Despite many contributions for open datasets in the medical imaging field, the size of available datasets is very small with respect to other fields such as pattern recognition. Therefore, it is very challenging to develop an advanced and generalizable ensemble method in medical image segmentation studies. If the training is necessary for the method, the number of parameters should be limited. Due to the relative scarcity of data sets containing medical images, attention was paid to choosing the ensemble methods to be used in this study from methods that can work

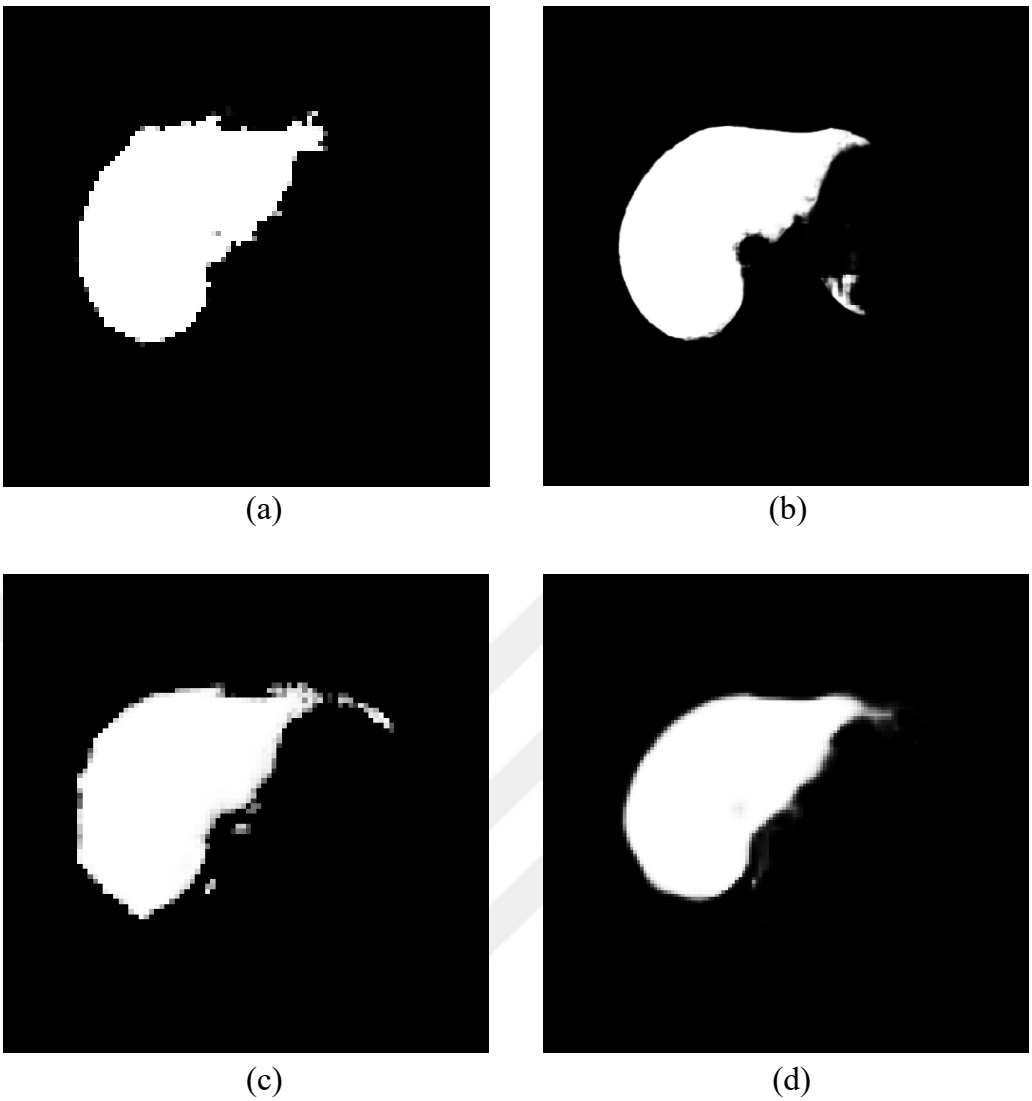

Figure 4.1 Probability maps for an example segmentation coming from (a) U-Net, (b) DeepMedic, (c) V-Net, and (d) Dense V-Networks

simpler but effectively.

In addition to the limitations on the data size, there are other reasons for keeping ensemble methods as simple as possible. First, heavily tuned advanced ensembles come with overfitting problems. In other words, the proposed method would fit the specific data. Overfitting problems would significantly reduce the reproducibility of the results. Secondly, implementations of simple methods are straightforward and do not need advanced expertise. Again, this fact will improve the reproducibility of the ensemble strategies by different scientists.

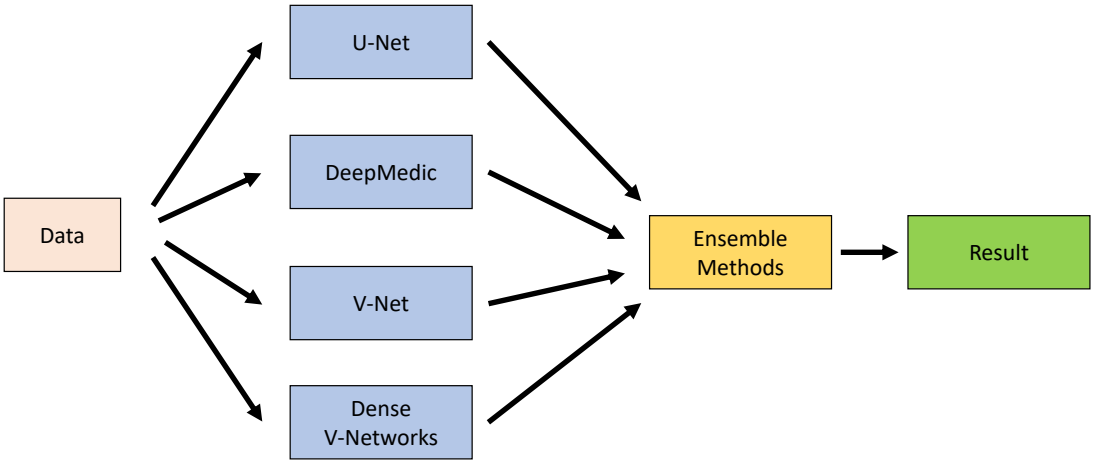

Figure 4.2 The proposed ensemble strategy for segmentation of whole liver

The formula of simple ensembles is defined as:

$$\mu_{liver}(\mathbf{x}) = F(p_1(\mathbf{x}), ...p_L(\mathbf{x})) \tag{4.1}$$

where $\mu_{liver}(\mathbf{x})$ is the support for the hypothesis that a given voxel $\mathbf{x}$ belongs to class liver, $p_1, \ldots, p_L$ indicate probability maps created by the various segmenters for every voxel in the data. Such values determine how likely the voxel is from the class (here background and liver). L is the number of individual classifiers, and F is the combination function.

Five ensemble strategies are used in this thesis:

1. Majority Voting
2. Average combiner
3. Product combiner
4. Min/Max combiner
5. Logit Combiner

Methods 1-4 are the most popular ensemble approaches that have already been studied in detail (Kuncheva, 2014). However, to the best of our knowledge, there is no

such a study to examine them from different perspectives for problem of liver segmentation. In this thesis, adaptation of ensemble methods in the medical image segmentation domain is handled in comprehensive way. The last method, "Logit Combiner", was specifically developed for the medical image segmentation problem in this thesis. The details of the ensemble methods are presented in the next section.

4.4.1 Majority Voting

Majority Voting is one of the most common strategy for classifier ensemble (Grofman et al., 1983). A majority vote is getting more than half of the votes to make a decision. Assume that, decisions are discrete values such that $\mu_{\text{liver}}(\mathbf{x}) = [d_1(\mathbf{x}), \ldots, d_L(\mathbf{x})], d_i(\mathbf{x}) \in \{0, 1\}$. Value $d_i(\mathbf{x}) = 0$ indicates that segmenter i labels voxel $\mathbf{x}$ as background, and $d_i(\mathbf{x}) = 1$ indicates that segmenter i proposes a label "liver" for this voxel. Then the majority vote combiner labels x as "liver" if:

$$\sum_{i=1}^{L} d_i(\mathbf{x}) \geq 0.5L \tag{4.2}$$

This is called 'simple majority voting' if number of class is 2. In this case, a decision is determined by any vote of more than 50% support. On the other hand, steps of a generalized majority vote for any number of class are summarized below:

1. Find class labels of each individual classifiers
2. Calculate number of votes for each class
3. Assign the labels according to votes for each class
4. Repeat 1-3 if there are more than single objects of interest (such as multiple organ segmentation)

Despite the meaning of the majority is more than half, there are different majority criteria that use the agreement of different portions such as 60%, 70%. A general formula for any majority criteria is expressed below:

$$\begin{cases} w_{c1}, & if \sum_{i=1}^{L} d_{i,c} \geq \alpha.L \\ w_{c2} & otherwise \end{cases} \tag{4.3}$$

where the values of d sum up to 1 for each segmenter, w_{c1} is class, w_{c2} is another class, L is number of ensemble members, and α is the threshold between $0 < \alpha \leq 1$. If $\alpha = 0.5 + \epsilon$, the formula becomes simple majority. If $\alpha = 1$, the formula represents unanimity voting. In other words, the decision is taken by agreement of all ensemble members.

The accuracy of majority voting can be calculated by the probability of correct decision. If the classifiers are independent, the probability of the method is:

$$P_{maj} = \sum_{m=\lfloor L/2 \rfloor + 1}^{L} \binom{L}{m} p^m (1-p)^{L-m} \tag{4.4}$$

Here, we assume that L is the number of ensemble members, p is the probability of true labels coming from each ensemble member, m is the number of ensemble members that make the correct decision. m must be greater at least half of the ensemble members ($\lfloor L/2 \rfloor + 1$) in order to create a correct decision.

If we assume that the probability of correct decision coming from each classifier is higher than 0.5 ($p > 0.5$), the Equation 4.4 shows that the accuracy of majority voting is increased by number of independent results:

$$\text{If } L \to \infty, \text{ then } P_{maj} \to 1 \tag{4.5}$$

The equation above shows that the number of ensemble members has an important impact on the decision of majority voting.

In the case of two segmentation classes (such as background and liver in this thesis), the majority voting method uses binary masks as discrete input data. Therefore majority voting can be implemented on probability maps of CNNs after thresholding them. CNNs' probability maps are thresholded by 0.5 (or defined

criteria) to create segmentation masks. Then, the final decision of each voxel is determined by the decision of multiple DMs.

4.4.2 Average combiner

Average combiner is produced from the generalized mean combiner (Eq.4.6) which uses continuous data, specifically probability maps, for classifier ensemble. That means, unlike majority voting, the probability maps of individual models are not thresholded and original values of probability maps from CNNs ($[p_1, ...p_L]$) can be used.

The generalized mean combiner is defined by (Dubois & Prade, 1985) as:

$$P_j(\alpha) = \left(\frac{1}{L} \sum_{i=1}^{L} (p_{i,j})^{\alpha} \right)^{\frac{1}{\alpha}} \tag{4.6}$$

where i is index of ensemble member, j is class number. If $\alpha = 1$, the formula is called simple average combiner. The probability map of average combiner (P_ave_j) is calculated by:

$$P\_ave_j = \frac{1}{L} \sum_{i=1}^{L} p_{i,j} \tag{4.7}$$

The combined probability map is thresholded by 0.5. For each voxel, if $P\_ave_j > 0.5$ the class foreground (P_f) is assigned to that voxel. Otherwise, the voxel belongs to the class background (P_b).

4.4.3 Product combiner

Product combiner also uses native probability maps of individual DMs. The formula can be derived from Eq. 4.6 if $\alpha \to 0$:

$$\lim_{\alpha\to 0} P_j(\alpha) = \lim_{\alpha\to 0}\left(\frac{1}{L}\sum_{i=1}^{L}(p_{i,j})^{\alpha}\right)^{\frac{1}{\alpha}}$$

$$\ln\left(\lim_{\alpha\to 0} P_j(\alpha)\right) = \ln\left(\lim_{\alpha\to 0}\left(\frac{1}{L}\sum_{i=1}^{L}(p_{i,j})^{\alpha}\right)^{\frac{1}{\alpha}}\right)$$

$$\lim_{\alpha\to 0}\left(\ln\left(P_j(\alpha)\right)\right) = \lim_{\alpha\to 0}\left(\ln\frac{1}{L}\sum_{i=1}^{L}(p_{i,j})^{\alpha}\right)^{\frac{1}{\alpha}} \tag{4.8}$$

$$= \lim_{\alpha\to 0}\left(\frac{\ln\left(\sum_{i=1}^{L}\frac{1}{L}\left(p_{i,j}\right)^{\alpha}\right)}{\alpha}\right)$$

If we apply L'Hôpital's rule to Equation 4.8, we obtain:

$$\frac{d}{d\alpha}\left(\ln\sum_{i=1}^{L}\frac{1}{L}p_{i,j}^{\alpha}\right) = \frac{\sum_{i=1}^{L}\frac{1}{L}p_{i,j}^{\alpha}\ln p_{i,j}^{a}}{\sum_{i=1}^{L}\frac{1}{L}p_{i,j}^{\alpha}}$$

$$\lim_{\alpha\to 0}\frac{\sum_{i=1}^{L}\left(\frac{1}{L}p_{i,j}^{\alpha}\ln p_{i,j}\right)}{\sum_{i=1}^{L}\left(\frac{1}{L}p_{i,j}^{\alpha}\right)} = \frac{\sum_{i=1}^{L}\left(\frac{1}{L}\ln p_{i,j}\right)}{\frac{L}{L}}$$

$$= \frac{1}{L}\sum_{i=1}^{L}\left(\ln p_{i,j}\right) \tag{4.9}$$

$$= \ln\left(\prod_{i=1}^{L} p_{i,j}\right)^{\frac{1}{L}}$$

Applying *exp* function on both side of the Equation 4.9 gives the final formula of product combiner shown in Equation 4.10.

$$P_j = \left(\prod_{i=1}^{L} (p_{i,j}) \right)^{\frac{1}{L}} \tag{4.10}$$

Instead of using product, the formula is converted to summation via logarithm to reduce computer memory needs. Hence logarithm of the Equation 4.10 is used.

$$P_j = \frac{1}{L} \sum_{i=1}^{L} \log(p_i) \tag{4.11}$$

Then, prior probability of class P_{j0} is subtracted to normalize process. This probability can be calculated as the proportion of class voxels in all images. After these calculations, support for class foreground (P_f) and background (P_b) are calculated by:

$$P_f = -\log(p_{f0}) + \sum_{i=1}^{L} \log(p_{fi}) \tag{4.12}$$

$$P_b = -\log(1 - P_{b0}) + \sum_{i=1}^{L} 1 - \log(p_{bi}) \tag{4.13}$$

Finally, the classes are assigned by values. The voxel values of $P_f > P_b$ are decided as foreground and the vise versa.

4.4.4 Minimum and Maximum combiners

Minimum and maximum combiners also use native probability maps of individual DMs. The formulas of P_{max} and P_{min} can be derived from Eq. 4.6 if $\alpha \to +\infty$ and $\alpha \to -\infty$.

Suppose that $p_{k,j}$ is the biggest element in $[p_{1,j}, p_{2,j}, ..., p_{L,j}]$. Then:

$$\lim_{\alpha\to\infty}\left(\ln\left(P_j(\alpha)\right)\right) = \lim_{\alpha\to\infty}\left(\frac{\ln\left(\sum_{i=1}^{L}\frac{1}{L}\left(p_{i,j}\right)^{\alpha}\right)}{\alpha}\right)$$

$$= \lim_{\alpha\to\infty}\left(\ln\left(p_{k,j}\right) + \frac{\ln\left(\sum_{i=1}^{L}\frac{1}{L}\left(\frac{p_{i,j}}{p_{k,j}}\right)^{\alpha}\right)}{\alpha}\right) \tag{4.14}$$

$$= \ln\left(p_{k,j}\right) + \lim_{\alpha\to\infty}\left(\frac{\ln\left(\sum_{i=1}^{L}\frac{1}{L}\left(\frac{p_{i,j}}{p_{k,j}}\right)^{\alpha}\right)}{\alpha}\right)$$

If we apply *exp* on both side in Eq.4.14, we obtain:

$$\lim_{\alpha\to\infty}\left(\ln\left(P_j(\alpha)\right)\right) = \ln\left(p_{k,j}\right)$$

$$\lim_{\alpha\to\infty}\left(P_j(\alpha)\right) = p_{k,j} \tag{4.15}$$

$$P_{max} = max\{p_{1,j}, p_{2,j}, ..., p_{L,j}\}$$

Similarly if $\alpha \to -\infty$, in Eq.4.14 and 4.15 then:

$$P_{min} = \frac{1}{max\{p_{1,j}, p_{2,j}, ..., p_{L,j}\}} \tag{4.16}$$

$$P_{min} = min\{p_{1,j}, p_{2,j}, ..., p_{L,j}\}$$

In two-class segmentation applications (background and one foreground), the minimum and the maximum combination rules are identical (Kuncheva, 2014).

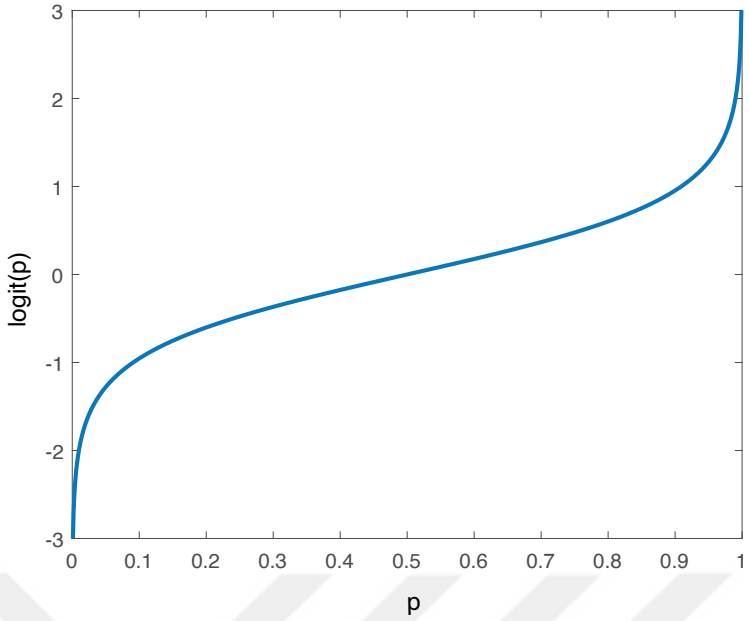

Figure 4.3 Logit function between $[0, 1]$

Therefore one of them is used in this study. In this combiner, we calculate P_f and P_b by

$$P_f = min\{p_{1,j}, p_{2,j}, ..., p_{L,j}\} \tag{4.17}$$

and

$$P_b = min\{1 - p_{1,j}, 1 - p_{2,j}, ..., 1 - p_{L,j}\} \tag{4.18}$$

Again, the voxel values of $P_f > P_b$ are decided as foreground and the vise versa.

4.4.5 Logit Combiner

Unlike the other four combiners, logit combiner was created specifically segmentation of liver problems during this thesis. In literature, several attempts have been made to adapt linear regression methods to map the output between a target range. Joseph Berkson used the logarithm of odds in 1944 and named this function logit, the "logistic unit" abbreviation. The plot of the logit function is shown in Figure 4.3

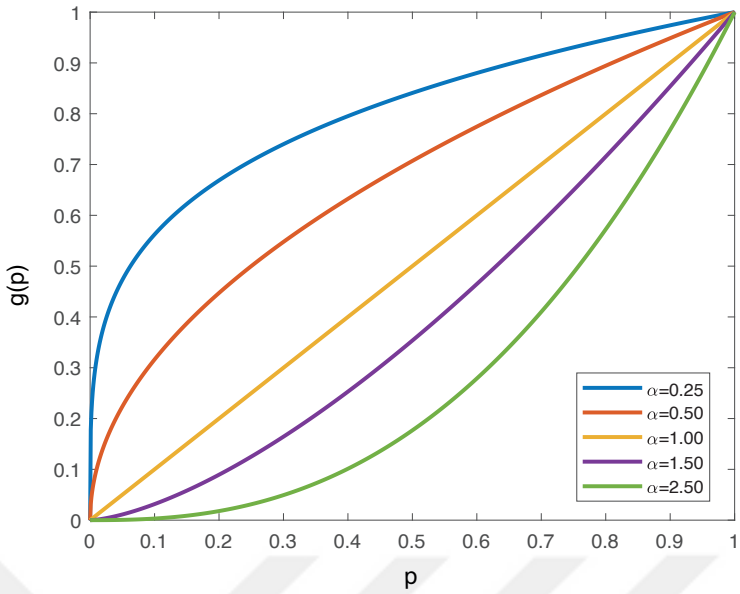

Figure 4.4 g(p) between $[0,1]$ with different α values

One of the common usages of logit is in Deep Learning studies. The last layer of CNNs, which are responsible for segmentation tasks, can base on logit function to make predictions from $(-\infty,+\infty)$ range to $(0,1)$. The idea of this mapping was adapted to our combiner approach as explained in Equation 4.19.

$$\begin{aligned} logit(p) &= log(p) - log(1-p) \\ &= log\left(\frac{p}{1-p}\right) \end{aligned} \tag{4.19}$$

$$g(p) = \left(log\left(\frac{p}{1-p}\right)\right)^{\alpha} \tag{4.20}$$

if $0 < \alpha < 1$

$$g(p) = \frac{p^{\alpha}}{p^{\alpha} + (1-p)^{\alpha}} \tag{4.21}$$

The α parameter determines the transformation of the function. α slightly but effectively changes the transformation of distributions as shown in Figure 4.4.

As it was mentioned at the beginning of this chapter, it is not possible to make tuning of multiple parameters in complex combiner designs due to the small size of available data. Therefore, it was paid attention to the formula to be concise and to contain only one trainable parameter: α.

α is used for fine-tuning during the ensemble in order to use the full potential of probability maps coming from different DMs. The whole training sets are used for validation of α values from $[0,3]$ interval with 0.1 steps. In other words, 30 different α values are tried and their final scores overall training cases are calculated. According to the scores, the $\alpha = 0.5$ is determined as the most effective value.

4.5 Evaluation

After the implementation of all ensemble methods, finally, the result of both individual DMs and ensemble strategies are examined in detail. The analyzes do not only contain metrics values but also statistical comparisons between individual DMs and ensemble for each metrics in both databases.

The same evaluation strategy in the CHAOS challenge was used. Since it was determined that these metrics performed effective and successful evaluation during the challenge, the same ones were used:

- DICE coefficient (DICE)
- Relative absolute volume difference (RAVD)
- Average symmetric surface distance (ASSD)
- Maximum symmetric surface distance (MSSD)

For both the individual DMs and the ensembles methods, all metrics were calculated. In addition, the statistical significance of the results was examined to test

our hypothesis: ensembles are better than individual segmenters. The statistical significance test protocol is explained below:

Suppose that we are looking for statistical significance of two methods, A and B. x and y are the vectors that have elements from $x \in A$ and $y \in B$. x and y are called paired samples if they represent the same element. For example, x represents a metric score of a method while y represents the same metric score from another method. If x and y represent same metrics but for different groups of objects, they are called non-paired samples.

The statistical significance is calculated for paired and non-paired samples by:

- *Paired samples:* Here we want to reveal that how x and y are significantly different. First, the Lilliefors test is run in order to check the normality of the difference of $x - y$. If it is normal, which means normality cannot be rejected at the 0.05 level, the paired t-test is applied to correlate means. If it is not normal, the Wilcoxon signed-rank test is applied.
- *Non-paired samples:* The normality of x and y is inspected first. If x and y are both normal, the 2-sample t-test is applied. Otherwise, the Wilcoxon rank-sum test (Mann–Whitney U test) is run.

The results of metrics and their significance against each other reveal the potential of ensemble systems. These findings are presented in Section 4.6 with various tables and illustrations.

4.6 Results

The ensemble members explained in Section 4.4, which are four different DMs, were downloaded from their original source. After that, they were trained with native parameters in source codes. Then, the generated probability maps by their outputs were used for both further analyses and inputs for ensemble methods.

After completing all experiments, the evaluation strategy mentioned in Section 4.5 was applied. According to these metrics, the performance of both individual DMs and ensemble methods tested in two steps:

1. Testing methods on train data to investigate overfitting.
2. Testing methods on test data to investigate segmentation performances.

DMs and ensemble methods are trained and tuned on train data explained in Section 4.4. In the first evaluation step, these methods were inferred on the train data. It is obvious that the results will be very high, but these can be used to examine the overfitting problem of algorithms. The second evaluation stage is a typical way to analyze the segmentation performance of methods. All individual DMs and ensemble methods were performed on test data. Metrics of the results were calculated by the same metrics with the CHAOS challenge. A full set of results is provided in Tables 4.2 – 4.5 showing the mean results for both CHAOS and 3DIRCADB1 datasets.

Table 4.2 Metric results of the individual segmenters and the ensemble methods on CHAOS train data to examine overfitting. The circle marker indicates results where the overfitting (calculated by the difference of training and testing performances) was not found to be significant

	DICE	RAVD	ASSD	MSSD
U-Net	0.935	∘14.800	3.903	54.650
Deepmedic	0.984	1.115	1.709	67.078
V-Net	0.948	∘3.824	1.656	42.972
Dense V-networks	0.932	3.039	2.289	∘78.118
Majority Vote	0.976	2.401	0.746	11.043
Average	0.981	1.003	0.637	11.621
Product	0.975	∘3.493	0.888	12.581
Min-Max	0.978	1.208	0.811	11.559
Logit	0.979	0.956	0.611	10.906

Table 4.3 Metric results of the individual segmenters and the ensemble methods on CHAOS test data to examine segmentation accuracy. The best value in each column is bold

	DICE	RAVD	ASSD	MSSD
U-Net	0.811	54.842	14.253	104.515
Deepmedic	0.951	**3.058**	7.174	141.473
V-Net	0.879	17.434	6.146	104.189
Dense V-networks	0.886	7.702	4.492	113.139
Majority Vote	0.952	4.235	1.719	28.517
Average	0.953	3.839	1.956	30.676
Product	0.946	6.867	2.121	32.696
Min-Max	0.937	6.094	2.311	35.052
Logit	**0.962**	4.215	**1.701**	**27.499**

Table 4.4 Metric results of the individual segmenters and the ensemble methods on 3DIRCADB1 training data to examine overfitting. The circle marker indicates results where the overfitting (calculated by the difference of training and testing performances) was not found to be significant

	DICE	RAVD	ASSD	MSSD
U-Net	0.903	19.336	∘7.414	∘69.760
Deepmedic	0.987	0.265	0.440	83.998
V-Net	0.964	2.795	1.027	17.048
Dense V-networks	0.970	1.162	0.946	∘53.298
Majority Vote	0.978	2.524	0.644	10.372
Average	0.982	1.491	0.568	17.951
Product	0.982	∘1.589	0.616	19.238
Min-Max	0.980	1.139	0.625	18.236
Logit	0.983	1.375	0.509	11.056

Table 4.5 Metric results of the individual segmenters and the ensemble methods on 3DIRCADB1 test data to examine segmentation accuracy. The best value in each column is bold

	DICE	RAVD	ASSD	MSSD
U-Net	0.672	74.923	66.869	172.513
Deepmedic	0.903	10.231	5.581	143.340
V-Net	0.826	19.004	9.547	95.222
Dense V-networks	0.900	8.881	9.953	113.306
Majority Vote	0.889	14.275	3.350	72.675
Average	0.920	6.736	3.338	75.057
Product	0.917	6.847	3.655	73.921
Min-Max	0.905	9.296	4.298	75.903
Logit	**0.932**	**6.091**	**3.144**	**71.972**

4.6.1 Ensemble segmenters show less overfitting than individual DMs

Overfitting problem of DMs can be seen comparing Tables 4.2 and 4.4 with Tables 4.3 and 4.5. The average training results are superior to the average testing results, not only for individual DMs but also for individual ensembles as expected. Overfitting in deep models also directly affects the results of the ensemble methods. Additional analyses have found that 8 of 64 differences at the 0.05 level are not statistically important. These 8 insignificant differences are pointed with a circle marker in Tables 4.2 and 4.4.

The results in Tables 4.2 – 4.5 reveal that ensembles does not have better metric values, but also they are unaffected by overfitting problem. In order to illustrate this, two tables (Table 4.6 and 4.7) were prepared for both datasets. These tables show overfitting magnitude calculated by training value minus testing value. Positive results for DICE mean better training values. Negative values for the remaining metrics point that the training value is better since the lower values of these measures are superior.

Table 4.6 Overfitting magnitude for the CHAOS dataset. Large overfitting corresponds to blue color and small overfitting, to red color. Each column (metric) is scaled individually

	DICE	RAVD	ASSD	MSSD
U-Net	0.1238	-40.0423	-10.3499	-49.8649
Deepmedic	0.0329	-1.9436	-5.4651	-74.3951
V-Net	0.0695	-13.6101	-4.4899	-61.2170
Dense V-networks	0.0463	-4.6631	-2.2030	-35.0205
Majority	0.0237	-1.8338	-0.9726	-17.4741
Average	0.0268	-2.8366	-1.3186	-19.0548
Product	0.0281	-3.3736	-1.2328	-20.1146
Min-Max	0.0381	-4.8866	-1.4997	-23.4931
Logit	0.0283	-3.2790	-1.1152	-19.9933

Table 4.7 Overfitting magnitude for the 3DIRCADB1 dataset. Large overfitting corresponds to blue color and small overfitting, to red color. Each column (metric) is scaled individually

	DICE	RAVD	ASSD	MSSD
U-Net	0.2320	-55.7723	-59.4755	-101.4027
Deepmedic	0.0830	-10.1691	-4.3534	-102.9930
V-Net	0.1406	-16.2087	-7.9385	-77.7407
Dense V-networks	0.0714	-7.6161	-8.0785	-53.1059
Majority	0.0896	-11.9168	-2.7082	-42.3905
Average	0.0626	-5.2867	-2.4555	-55.8543
Product	0.0614	-2.9921	-2.4300	-52.8161
Min-Max	0.0738	-6.7472	-3.0785	-57.2058
Logit	0.0616	-5.0802	-2.3179	-52.5097

Besides, the color-coded values for all metric are presented in Tables 4.6 and 4.7. Red colors show minor overfitting while blue colors show major overfitting. According to Tables 4.6 and 4.7, the blue color indicates at the tops, the individual DMs are more vulnerable to overfitting than the ensembles.

4.6.2 Ensemble segmenters offer better results than individual DMs

The superiority of ensembles can be clearly observed by Tables 4.3 and 4.5 in the most cases according to four different metric results. In addition, the significance of these results are analyzed in Tables 4.8 – 4.11. These 8 different tables show the statistical significance between individual DMs and ensemble methods for four metrics and two datasets. The significance level is determined as 0.05. The values in the tables again confirm the preferability of ensembles against individual methods.

After showing that fusion/ensemble methods are more successful, the question of which method is more preferable was also examined. Two glyph plots in Figures 4.5 and 4.6 show the comparison of ensemble methods. The cumulative test performance of all metrics was used to create these two plots. The DICE values were reversed to make all metrics results have the same distribution (smaller values indicate better results). All scores were mapped between 0.0 and 1.0. They were marked on the edges of the plots. The smallest area represents the most successful ensemble method in glyph plots. In Figures 4.5 and 4.6, the most preferable ensemble is the Logit ensemble. It is followed by majority voting in Figure 4.5 and average combiner in Figure 4.6. However, majority voting has a bad performance in Figure 4.6 with its large area. The plots show that the average based methods such as Logit combiner and simple average combiner perform the best segmentation.

Besides with glyph plots, the values in Tables 4.8 – 4.11 support the fact that Logit and average combiners are the most successful ones. These analyses indicate that the designed Logit combiner can be recommended for the ensemble of DMs with vanilla-style parameters.

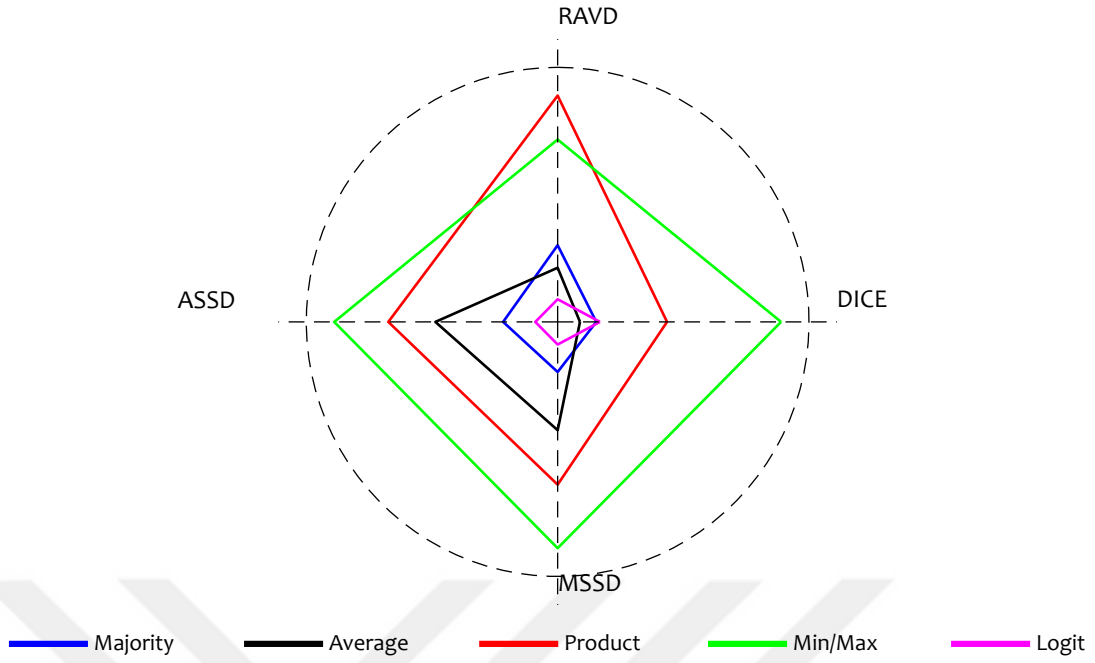

Figure 4.5 Glyph plot of the four ensemble methods for the CHAOS dataset. The spokes are the four metrics. Small-area ensembles are preferable

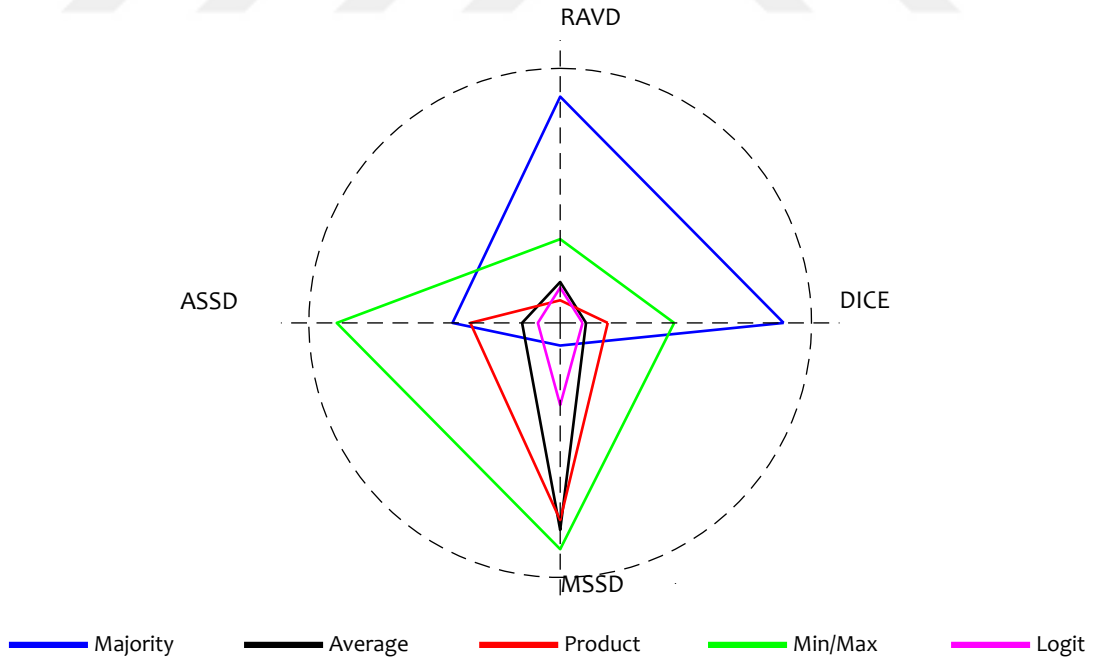

Figure 4.6 Glyph plot of the four ensemble methods for the 3DIRCADB1 dataset. The spokes are the four metrics. Small-area ensembles are preferable

Table 4.8 DICE: Statistical comparison between individual DMs and ensembles. Bullet means that the ensemble wins; circle means that the DM wins; line means that no statistical difference.

CHAOS dataset

	Majority	Average	Product	Min/Max	Logit
U-Net	•	•	•	•	•
Deepmedic	–	–	–	○	–
V-Net	•	•	•	•	•
Dense V-Networks	•	•	•	•	•

3DIRCADB1 dataset

	Majority	Average	Product	Min/Max	Logit
U-Net	•	•	•	•	•
Deepmedic	–	–	–	–	–
V-Net	•	•	•	•	•
Dense V-Networks	–	–	–	–	•

Table 4.9 RAVD: Statistical comparison between individual DMs and ensembles. Bullet means that the ensemble wins; circle means that the DM wins; line means that no statistical difference

CHAOS dataset

	Majority	Average	Product	Min/Max	Logit
U-Net	•	•	•	•	•
Deepmedic	○	–	○	–	–
V-Net	–	•	–	–	•
Dense V-Networks	–	•	–	–	–

3DIRCADB1 dataset

	Majority	Average	Product	Min/Max	Logit
U-Net	•	•	•	•	•
Deepmedic	–	–	–	–	•
V-Net	–	•	•	•	•
Dense V-Networks	–	–	–	–	–

Table 4.10 ASSD: Statistical comparison between individual DMs and ensembles. Bullet means that the ensemble wins; circle means that the DM wins; line means that no statistical difference

CHAOS dataset

	Majority	Average	Product	Min/Max	Logit
U-Net	•	•	•	•	•
Deepmedic	•	•	•	•	•
V-Net	•	•	•	•	•
Dense V-Networks	•	•	•	•	•

3DIRCADB1 dataset

	Majority	Average	Product	Min/Max	Logit
U-Net	•	•	•	•	•
Deepmedic	–	•	•	–	•
V-Net	•	•	•	•	•
Dense V-Networks	–	•	–	–	•

Table 4.11 MSSD: Statistical comparison between individual DMs and ensembles. Bullet means that the ensemble wins; circle means that the DM wins; line means that no statistical difference

CHAOS dataset

	Majority	Average	Product	Min/Max	Logit
U-Net	•	•	•	•	•
Deepmedic	•	•	•	•	•
V-Net	•	•	•	•	•
Dense V-Networks	•	•	•	•	•

3DIRCADB1 dataset

	Majority	Average	Product	Min/Max	Logit
U-Net	•	•	•	•	•
Deepmedic	•	•	•	•	•
V-Net	•	–	–	–	•
Dense V-Networks	•	•	•	•	•

In conclusion, the capabilities of DMs in their native versions were analyzed as well as the potential of their ensembles. Despite the remarkable success of DMs, their vanilla style versions are far away from the heavily optimized ones. It was revealed that DMs need to be revised comprehensively for specific problems and data. In addition, DMs are prone to overfitting. Small variations in parameters and/or minor differences in design can significantly alter the results produced by DMs. Therefore, DMs can produce remarkable outputs as a result of the intense efforts of the experts. On the other hand, our experimental findings verified and also demonstrated that overfitting can be reduced by using even simple ensemble methods. We showed this onto two publicly accessible datasets.

Fascinated by the progress of DMs in medical segmentation, we have explored the ability of ensemble of segmenters based on DMs consisting of state-of-the-art publicly accessible DMs. In general, the segmentation accuracy of ensemble methods has been tested with the various metrics suggested in the literature. Our finding is that the performance of individual vanilla style DMs can be boosted by using ensemble approaches without making huge efforts.

On the other hand, it has been observed that the accuracy of the ensemble methods is sensitive to the accuracy of ensemble members. To eliminate this, normally ensembles are based on the same members with different training data. However, it is not feasible in the medical image analysis field due to the lack of data. The ensemble studies that use different models can be optimized using calibration. Therefore a calibration parameter can improve segmentation accuracy. The performance of the Logit combiner proves this situation. Even a single parameter (α) can improve segmentation accuracy. Logit combiner reached the best accuracy out of five ensemble rules. This indicates that the development of more effective ensemble methods may be a positive way forward.

CHAPTER FIVE
SEGMENTATION OF ABDOMINAL AORTIC PATHS AND LIVER VEINS FROM CT ANGIOGRAPHY

5.1 Introduction

The importance of the liver was explained in Section 2.1.1 as well as its segmentation from medical images in Section 4.1. Segmentation of the whole liver is the primary stage for further analyses. One of the most demanded examination in clinics is extracting vein structures of the liver via segmentation. As mentioned in Section 2.4, the vein system in the liver is too complex with so many branches. Therefore these structures need to be examined by invasive methods such as angiography. These methods provide liver venous systems with angiographic control. Angiography has a significant role in diagnostics of hypertension in the portal vein, lesions, abdominal trauma, and hepatic venous occlusion. Also, angiography of the liver may show a hypervascular tumor(s). However invasive angiographic methods can be tough.

The alternative to direct hepatic angiography is computed tomography angiogram (CT angiogram) (Winter & Auer, 2012). In this method, the contrast agent is injected into the patient with the synchronization of an abdomen CT scan. The images acquired with a pre-defined delay to catch the flow of contrast agent in veins. Even small defects can be sensitively detected with CT angiogram. Besides, blockages in veins can be identified.

In addition with liver veins, segmentation of abdomen aorta is another important operation in clinics. The problems such as abdominal aortic aneurysm (AAA) (shown in Fig. 5.1) may restrict surgical operations in the liver. Therefore segmenting vessel tree of the liver as well as the abdomen aorta can be used together in radiology department.

Identifying of veins in the liver also is a necessary process to classify the liver into

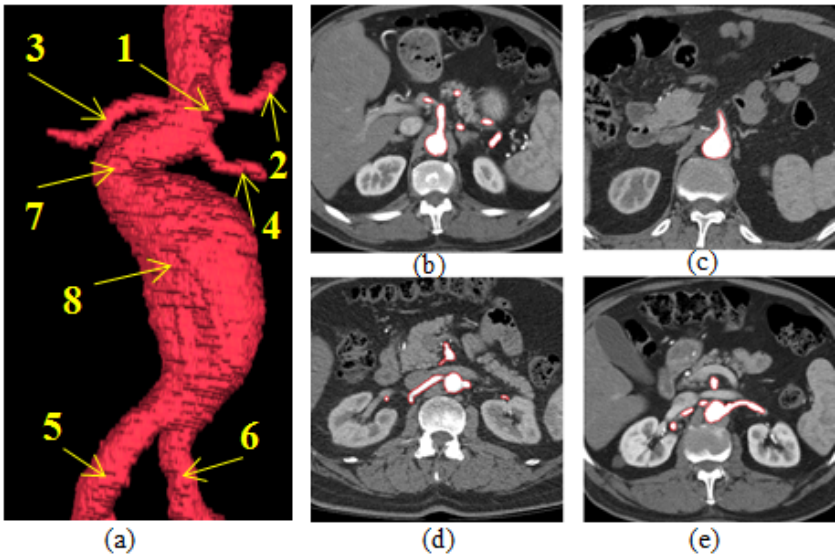

Figure 5.1 (a) 3D aorta with AAA and outgoing celiac vessels (1, 2), renal vessels (3, 4), iliac arteries (5, 6), aneurysm neck (7) and aneurysm (8). 2D cross sectional CTA images for well-organized acquisition showing departure of vessels at celiac region (b) vessel 1, (c) vessel 2, and renal arteries (d) vessel 3, (e) vessel 4 (Selver & Kavur, 2015)

Couinaud lobes shown in Figure 5.2. Couinaud classification splits the liver into eight segments that perform individually. In other words, there are independent blood flow and drainage in these segments. Each segment has a section of the hepatic artery, portal vein, and bile duct in the middle (Shepherd & Turmelle, 2017).

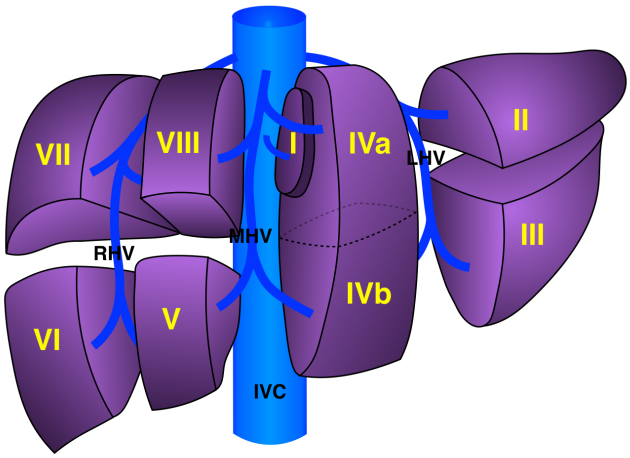

Figure 5.2 Overview of Couinaud lobes of the liver (Jones, 2019)

The right hepatic venous separates the right lobe into the front (anterior) and back (posterior) parts. The center hepatic vein separates the organ into the left-right parts.

This left-right surface places between the inferior vena cava and the gallbladder. The portal vein separates the liver in the vertical direction (up-down)

To sum up, the segmentation of veins of the liver and abdomen aorta are used for many clinical examinations. It is the preliminary stage to decide Couinaud lobes as well as detecting so potential abnormalities in the liver venous system and abdomen vessels. In this chapter, first, a novel method was developed to segment abdomen aorta via pairwise geodesic distance fields. The method was published (Selver & Kavur, 2015) and all details are available in the article. Then, the ensemble methods were examined whether the same methods with whole liver segmentatios could be used for liver vein segmentation problem. Due to the difficulty of this problem, fewer methods have been proposed in the literature comparing to the segmentation of the whole liver.

5.2 Segmentation of Abdominal Aortic Paths Using Pairwise Geodesic Distance Fields

Segmentation of arteries exiting from aorta (i.e. celiac, renal, iliac) is used to prepare liver transplant surgery. Also it is important to decide the graft 's location prior to aortic aneurysm surgery. With help of minor user input such as adding seed points at the edges of the target vessels can be used to extract aorta and neighbor vessels. The algorithms can detect the proper path between these seed points. In the proposed algorithm, Geodesic Distance (GD) was used to connect seed points.

Geodesic distance (GD) can be described as the route between two image nodes based on specified limitations. These limitations force the route to remain in a subset of the image, which is called Geodesic Mask (GM). If S is a connected set containing pixels α at (x1, y1) and β at (x2, y2) (i.e. S is GM), then GD between α and β, $D_S(\alpha,\beta)$, can be calculated as:

$$D_S(\alpha,\beta,) = min\{L(P)|p_1 = \alpha, p_l = \beta \; and \; P \subseteq S\} \tag{5.1}$$

where, $L(P)$ is the length of a path P=(p_1, p_2, ... p_l) between all achievable paths

connecting α and β and included in S. Using GD, value of a Propagation Function (PF) at pixel ψ is defined as:

$$PF_S(\psi) = max\{\ D_S(\psi, \sigma)\ \ |\ \sigma \epsilon S\ \}\ \forall \psi \epsilon S \tag{5.2}$$

This expression matches to the minimum of distance of the geodesic paths starting from ψ and included in S. Based on these statements, a geodesic path can be found for two pixels α and β in a connected region S using Soille (2003):

1. Calculate PF_S and select pixels satisfying $max_{y \epsilon S}\ PF_S(y)$. Here, the assumption is that there are only two pixels (i.e. α, β) giving this condition (Otherwise, procedure must be adopted).
2. Compute GD Functions (GDF),$GDF_S(\alpha)$ and $GDF_S(\beta)$, using a proper algorithm, such as Euclidean GD Soille (1994).
3. Compute the sum of $GDF_S(\alpha)$ (Fig. 2.a) and $GDF_S(\beta)$ (Fig. 2.b), that is equal to $GDF_S\ (\alpha, \beta)$ (i.e. PGDF) (Fig. 2.c).
4. Along the searched path, $GDF_S(\alpha, \beta)$ is minimal. Therefore, application of a regional minimum search algorithm to $GDF_S(\alpha,\beta)$ gives the geodesic path linking α to β.

This method was proposed in Soille (1994) for 2D data. However it was not optimized for 3D data before. The practicality of this approach to the detection of the lumen route has still not been analyzed.

5.2.1 Insertion of Vessel Nodes

Adding seed points within the renal and iliac vessels involves the position of small vessels of interest. As a commonly used technique in clinic, Multiplanar Reconstruction (MPR) is the easiest way of restoration that is constructed by piling the initial image slices. Since whole volumetric data is used, it is possible to achieve

any necessary plane, thereby allowing easy 2D viewing of the vessels. The algorithm can then be started by inserting nodes in the vascular segment (branches) that the user wishes to integrate via the MPR interface.

5.2.2 Generation of the Geodesic Mask (GM)

Generation of GM has a key role in precision and efficiency. Through removing other anatomical structures, GM must be built as close as possible to the aorta. All large vessels will be operated at the same time. Therefore, rooted arteries and their connections to the aorta must be limited for the development of a GM to maintain its connectedness.

Due to the acquisition is performed with injection of the contrast agent, the thresholding can be used a suitable GM generation technique. In the optimal case where the processing timing and modality parameters are well organized, the volume data histogram consists of one single peak for an aorta.

By reason of inhomogeneous scattering of contrast matter, in routine clinical procedure vessels may not be properly enhanced, often resulting in smaller hills of volume histograms. In fact, a broader variety of Hounsfield values must also be preserved to avoid specific vascular details being lost. This involves a challenging task which is resolved through the automatic version of the hill-based approach proposed by Papamarkos & Gatos (1994) and by regulating the connectivity of specified points to further change the threshold range. Their bi-level thresholding method has the following steps:

1. Hill clusters by iterative cell scan, depending on the location of the histogram peaks, when the gray level in each cell is halved in each iteration

2. Histogram fitting via real rational functions over a altered linear rational approximation method Papamarkos (1989) which provide both computational cost and less error,

3. The multilevel threshold amounts are calculated as the global minimum values, which are detected by golden search method Press et al. (1986), of rational functions (found in Step 2).

4. Two level threshold values are calculated between different threshold values by choosing clustered hills composed of nodes added by the physician.

5. Perform a 3D connected component analysis to control if user added nodes are inside of the result. If not extend two-level threshold at the route determined by searching gray values at the vicinity of the non-included node(s).

5.2.3 Calculation of Pairwise Geodesic Distance Functions (PGDFs) and Geodesics in 3D

3D-PGDFs are determined between the seed-point pairs introduced by the user after the GM calculation shown in Fig. 5.3. As in the 2D case, 3D-PGDF is chosen in pairs rather than generating PGDF at the same time by all seed points. As stated before, the large number of GDFs that construct PGDF via cumulative sum (PGDF= $\sum_{i=1}^{M} GDF_S(x_i)$) increases as a result of simulations that show small vessels by PGDF decreases.

Approximated equidistant geodesics are calculated by the following steps after building of 3D-PGDF, :

1. Compute *PGDF*(i.e. $GDF_S(\alpha,\beta)$) using user inserted seed pairs α at (x_1,y_1,z_1) (i.e. $GDF_S(\alpha)$) and β at (x_2,y_2,z_2) (i.e. $nDF_S(\beta)$) inside GM.

2. Calculate modulus of the PGDF by N (i.e. PGDF mod N). Save quotients of modulus operation which constitute an integer matrix having the same size as PGDF. The thickness and the amonut of geodesics in GM is determined.

3. The set of voxels having the value of the smallest quotient gives the distorted geodesic (DG) connecting α and β. Each of the sets of quotient values forms other geodesics.

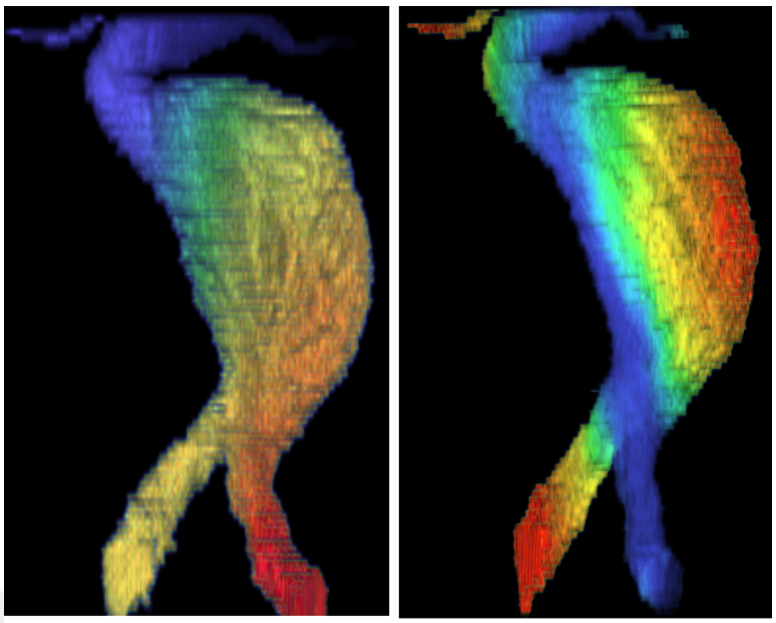

Figure 5.3 3D-CGDF which is suitable for gradient search as it is monotonically increasing from seed 1 to seed 2, (b) 3D-PGDF Selection of the geodesics represented inside blue region provide a path between seed 1 and seed 2. Path becomes more robust at aneurysm neck if turquoise region is also used (Selver & Kavur, 2015)

The number of geodesics for a PGDF depends on the selection of modulus parameter, N. If PGDF constitutes values between $PGDF_{min}$ and $PGDF_{max}$, then the number of geodesics inside GM is equal to

$$q = (PGDF_{max} - PGDF_{min})\ mod\ N\ where\ q \in Z^{+} \tag{5.3}$$

5.2.4 Path Extraction Using Enhanced PGDFs

The path between seed pairs are obtained by using a thinning algorithm to the DG after obtaining enhanced PGDF. A fast image thinning method in ITK library was used for 3D thinning (Ibáñez et al., 2003). The application and result of the proposed algorithm are presented in Section 5.1. Besides these results were published in our article (Selver & Kavur, 2015).

5.2.5 Results

In 12 CTA datasets from 12 different patients in 4 different modalities, the proposed lumen path extraction procedure was tested. The first data set has ideal image characteristics. The images was acquired by a CT scanner with a 16-row detector (Philips Mx8000) with a slice thickness (ST) of 3.2 mm. Images were selected by a PACS specialist at the Dokuz Eylül University of Medicine School, Izmir, Turkey.

The remaining data sets were obtained from acquisitions which represent the daily clinical practice challenges and are addressed in this study. 6 data sets have been acquired with a 320-row CT scanner (Toshiba Aquilion One) with 3.0 mm ST. 4 data sets have been acquired with a 64-row CT scanner (Toshiba Aquilion) with 5.0 mm ST. These data sets were chosen from regular clinical acquisitions from the PACS of Gúlhane Faculty of Medicine in Ankara, Turkey. The efficiency of the proposed method for ST values of 3.0 and 5.0 mm defines the clinical usefulness of the process, as these are the most typical ST values used in clinical practice. A single dataset obtained from the 320-row detector CT scanner with 0.8 mm ST. This data set is analyzed to check the output of the proposed system at low ST values. A total of 2758 DICOM image slices were analyzed, all of which have 512 x 512 pixels.

The ground truth data was built slice-by-slice manual aorta delineation by an experienced radiologist who has been practicing abdominal aortic aneurysm graft injection surgery at Gülhane Medicine Faculty, Ankara, Turkey for more than 10 years. Another 2 experts conducted the insertion of user identified seed points. Using the developed MPR interface, 6 points were inserted into the departing vessels for each data set from aorta to liver, spleen, kidneys (right / left), and iliac arteries (right / left). The insertion of seed points procedure was repeated 5 times at different dates to analyze the dependency of the proposed method to seed points. Thus, experiments for various sets of user implanted seeds is replicated 5 times for each dataset.

In order to evaluate performance of the proposed method three metrics were used:

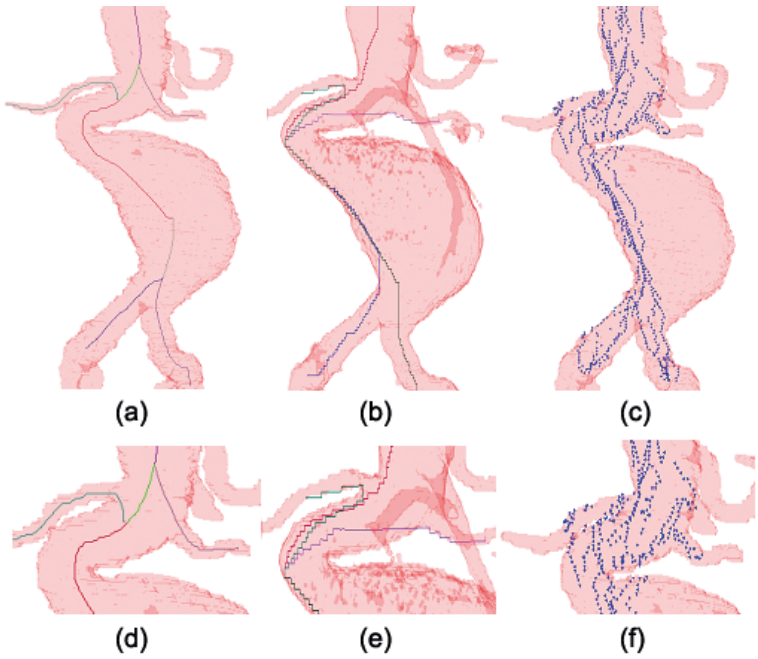

Figure 5.4 Results of the compared algorithms for the same seed points (a), (d) lumen path found by subvoxel MSFM, (b) (e) lumen path found by 3-D FM, (c), (f) seed points generated by the proposed method (Selver & Kavur, 2015)

percentage of correct classification (CC), sensitivity (SE), specificity (SP). To compare performance two well studied methods were applied to the same dataset under same conditions. These methods are 3D fast marching (3D-FM) (Peyre, 2004) and subvoxel multi-stencil fast marching (MSFM) Kroon (2009). All results are presented in Table 5.1 and Fig. 5.4.

Table 5.1 Quantitative Comparison of the Accuracy for 6 Seeds

	3D-PGDF			3D-FM			Subvoxel MSFM		
Dataset	CC	SE	SP	CC	SE	SP	CC	SE	SP
1	99.80	85.09	99.97	99.78	81.55	99.98	99.79	87.80	99.99
2-7	99.85	70.62	99.98	99.73	65.23	100	99.85	73.36	100.00
8-11	99.78	72.90	99.98	99.80	69.88	99.99	99.83	75.97	100.00
12	99.75	72.36	99.89	99.78	68.32	99.96	99.83	72.48	100.00
All	99.81	72.75	99.97	99.75	68.39	99.99	99.83	75.76	100.00

The results reveal that, while the suggested 3D-PGDF based approach outperforms 3D-FM, subvoxel MSFM performs the best in all metrics. These three approaches

indicate very similar consistency at both CC and SP speeds. However, when considering SE their performance shows significant differences. Sensitivity, which tests the proportion of positive currently known as such, determines how well the region's that algorithm executes segmentation tasks within the aortic region. The better results for Subvoxel MSFM are attributed to its ability to extract the center line whereas the suggested technique and 3D-FM merely remove a direction within the lumen. Nonetheless, in an typical SE, 3D-PGDF performs 4.17% over 3D-FM.

Despite having the best output metrics, subvoxel MSFM not only takes a considerably long period to complete the cycle, but also demonstrates high timing variation (i.e. see last column of Table 5.2. 3D-FM is the fastest algorithm of all and takes approximately 3.5 times less time compared to subvoxel MSFM. Our technique takes nearly 1.7 times as much time as 3D-FM but reveals less variation at the moment. As a result , we can conclude that using the proposed algorithm for path generation and seeding on a performance-computational complexity basis shows a performance in-between the benchmark techniques.

In consideration of these important vessels' limited scale and high curvature, seed points plays a major role, as most algorithms are initialized by these seeds. The proposed approach works in-between these systems as opposed to effective comparison approaches. The proposed algorithm, in particular, is quicker than the high-precision approach (i.e., subvoxel MSFM) and more precise than the fast-gradient based technique (i.e. 3D-FM).

Therefore, we conclude that the suggested approach provides an alternative to 3D-FM and subvoxel MSFM subvoxel, having high accuracy in a reasonable time than 3d-FM techniques. Moreover, the standard deviation from the experiment indicates that, as opposed to the 3D-FM and subvoxel MSFM techniques, the technique suggested is less influenced by the location of the original seeds introduced by the specialist.

Table 5.2 Comparison of the Computational Time (seconds) for 6 Seeds

	3D-PGDF		3D-FM		Subvoxel MSFM	
Dataset	Mean	Std.	Mean	Std.	Mean	Std.
1	248.82	6.40	143.01	9.95	708.60	41.12
2	243.60	24.80	140.32	38.75	560.30	85.59
3	205.32	37.20	118.52	62.40	479.75	60.63
4	211.87	25.40	116.02	59.89	444.92	66.36
5	218.90	37.24	120.90	57.89	463.62	81.14
6	273.18	19.80	157.23	33.17	473.57	33.02
7	345.46	27.60	256.62	46.66	526.51	42.41
8	179.22	24.86	103.93	40.04	327.38	49.21
9	148.78	24.45	97.56	24.97	295.21	67.25
10	126.58	31.14	67.93	35.61	298.02	59.79
11	191.40	21.07	110.34	35.44	266.95	45.18
12	203.58	3.60	117.75	6.14	657.90	33.17
All	216.39	23.63	129.18	37.58	458.56	55.41

5.3 Segmentation of Liver Veins via Fusion of Different Methods

5.3.1 Liver Vein Dataset

The problem of having fewer data sets in the medical imaging area compared to other fields has been mentioned before. In addition, the number of datasets containing the vessels in the liver is even less. As a result of our qualitative analyzes on a small number of datasets, we observed that the vessels in many published datasets were not annotated precisely. Since the vessels are complex and small in some areas, their correct marking is more difficult than other structures. Therefore, in many datasets, the vessels are annotated by some semi-automatic tools. However, the quality and accuracy of annotations are directly proportional to the capacities of these tools. Even in some popular databases in literature, clinically flawed markings such as missing parts and miss-/over-segmentation have been seen. In this study, we are introducing a

specific dataset that includes handcrafted annotation of liver vessels.

To overcome problems about annotations, a unique dataset that contains liver vessels have been created for the studies in our department. This dataset is called Vessel Extraction and Extrication for Liver Analysis (VEELA). It contains the same images in the CHAOS CT set with annotations of vessels inside the liver. Both portal and hepatic veins were annotated **manually**. According to our knowledge, there is not such an open dataset with handcrafted segmentation masks. Although the annotations of the vessels have taken more time than planned, the resulting ground truth mask will guarantee the precise segmentation evaluation. After the worldwide popularity of the CHAOS challenge and its dataset, a new challenge involving the segmentation of the liver vascular system is planned with this set in the future. Therefore, this set has not been published openly yet. It is only available for internal studies in our department at the moment. The technical specifications of this set are as same as CHAOS CT set presented in Table 3.7. Besides, a sample case is also presented in Figure 5.5

5.3.2 Ensemble Members

The structure of vessels has important differences than the whole liver as can be seen from Figure 5.5. These are:

1. Liver has a concave shape while vessel trees are convex.
2. Size of the liver is relatively bigger than vessels.
3. Vessels have both bigger and smaller regions while the liver has homogeneous size.

These features of the vessels above make segmentation of them much harder. Therefore, a unique strategy needed to be developed. If the vascular tree is examined from a single direction like whole liver segmentation procedures, it is determined that small vessels in some areas have just 3-4 voxel size. Due to this smallness, the segmentation of these extreme structures can be very difficult. However, if these small

structures can be examined from other angles, it can be seen that they have much larger areas in the cross-sections.

A different method has been developed to solve the problems arising from the angle of examination. Here, in addition to original data, four rotated versions of the data applied to the same DM were used to create ensemble members. Our proposed ensemble procedure is using the results of DMs, trained independently with each other with rotated data. In other words, a homogeneous ensemble approach was preferred unlike heterogeneous ensemble strategy in whole liver segmentation problem. Our strategy is illustrated in Figure 5.6.

The reason for using a single DM is that the performance of ensembles is

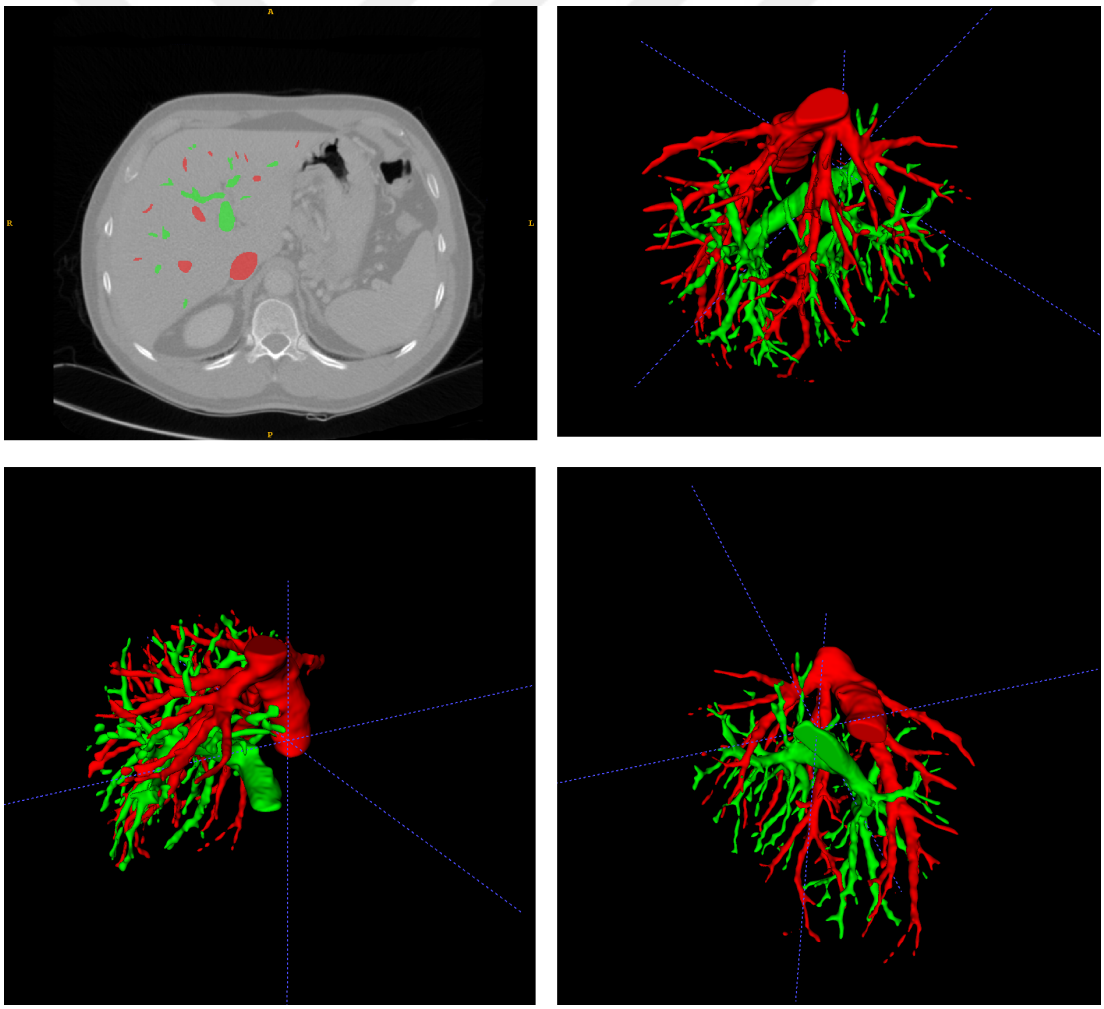

Figure 5.5 A sample case from VEELA dataset. Hepatic artery(red) and portal vein(green) on axial CT image and 3D visualizations in different angles

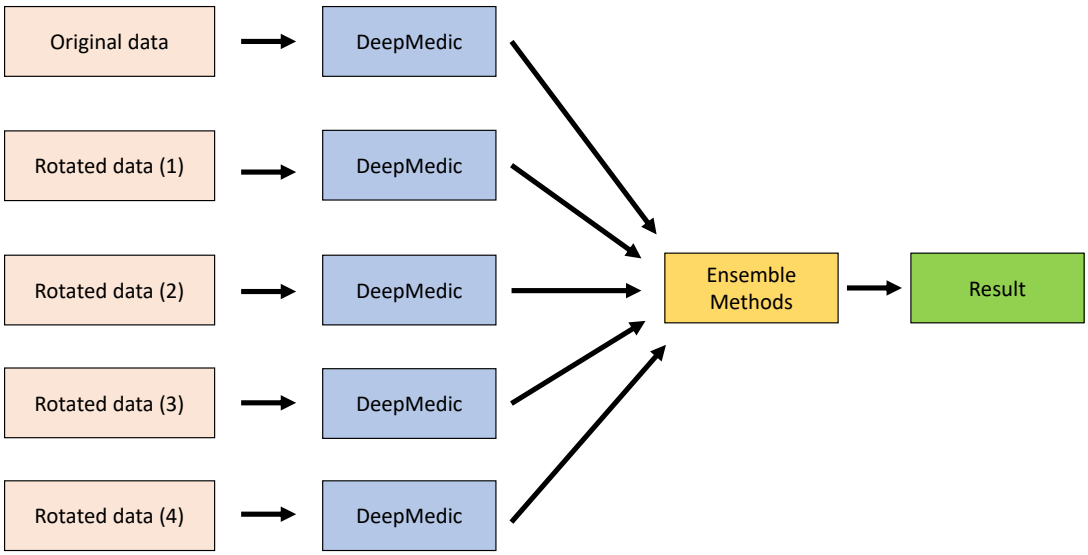

Figure 5.6 The proposed ensemble strategy for segmentation of liver veins

Table 5.3 Parameters of rotations applied on the dataset

Rotation	Plane	Degree
Rot1	X-Z	45°
Rot2	X-Z	135°
Rot3	Y-Z	45°
Rot4	Y-Z	135°

correlated with the performance of ensemble members. Since the segmentation of liver vessels is a harder problem than segmentation of the whole liver, the most successful DM was used to improve the average performance of ensemble members. Therefore the base CNN was selected as DeepMedic (Kamnitsas et al., 2016, 2017). The main reason is that robustness of DeepMedic against difficult areas is higher than other models according to the results in Section 4.6. To summarize, $E = \{D_{Axial}, D_{Rot1}, D_{Rot2}, D_{Rot3}, D_{Rot4}\}$ are the ensemble members as shown in Figure 5.6. The details of the rotations are presented in Table 5.3. The affine matrix for transformations are presented in Equations 5.4-5.7. Also, the rotation planes are illustrated in Figure 5.7.

$$A_{Rot1} = \begin{bmatrix} 0.707 & -0.500 & 0.500 & 0.000 \\ 0.500 & 0.856 & 0.146 & 0.000 \\ -0.500 & 0.146 & 0.854 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix} \tag{5.4}$$

$$A_{Rot2} = \begin{bmatrix} -0.707 & -0.500 & 0.500 & 0.000 \\ 0.500 & 0.146 & 0.854 & 0.000 \\ -0.500 & 0.854 & 0.146 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix} \tag{5.5}$$

$$A_{Rot3} = \begin{bmatrix} 0.854 & -0.500 & 0.146 & 0.000 \\ 0.500 & 0.707 & -0.500; & 0.000 \\ 0.146 & 0.500 & 0.854 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix} \tag{5.6}$$

$$A_{Rot4} = \begin{bmatrix} 0.146 & -0.500 & 0.853 & 0.000 \\ 0.500 & -0.707 & -0.500 & 0.000 \\ 0.854 & 0.500 & 0.146 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix} \tag{5.7}$$

5.3.3 Ensemble Methods

Again care was taken to the simplicity of the ensemble methods. Therefore, the same ensemble methods with whole liver segmentation in Section 4.4 were performed. They are Majority Voting, Average combiner, Product combiner, Min/Max combiner, and Logit Combiner. Since the explanations of these methods are given in detail in Section 4.4, new explanations are not included in this section.

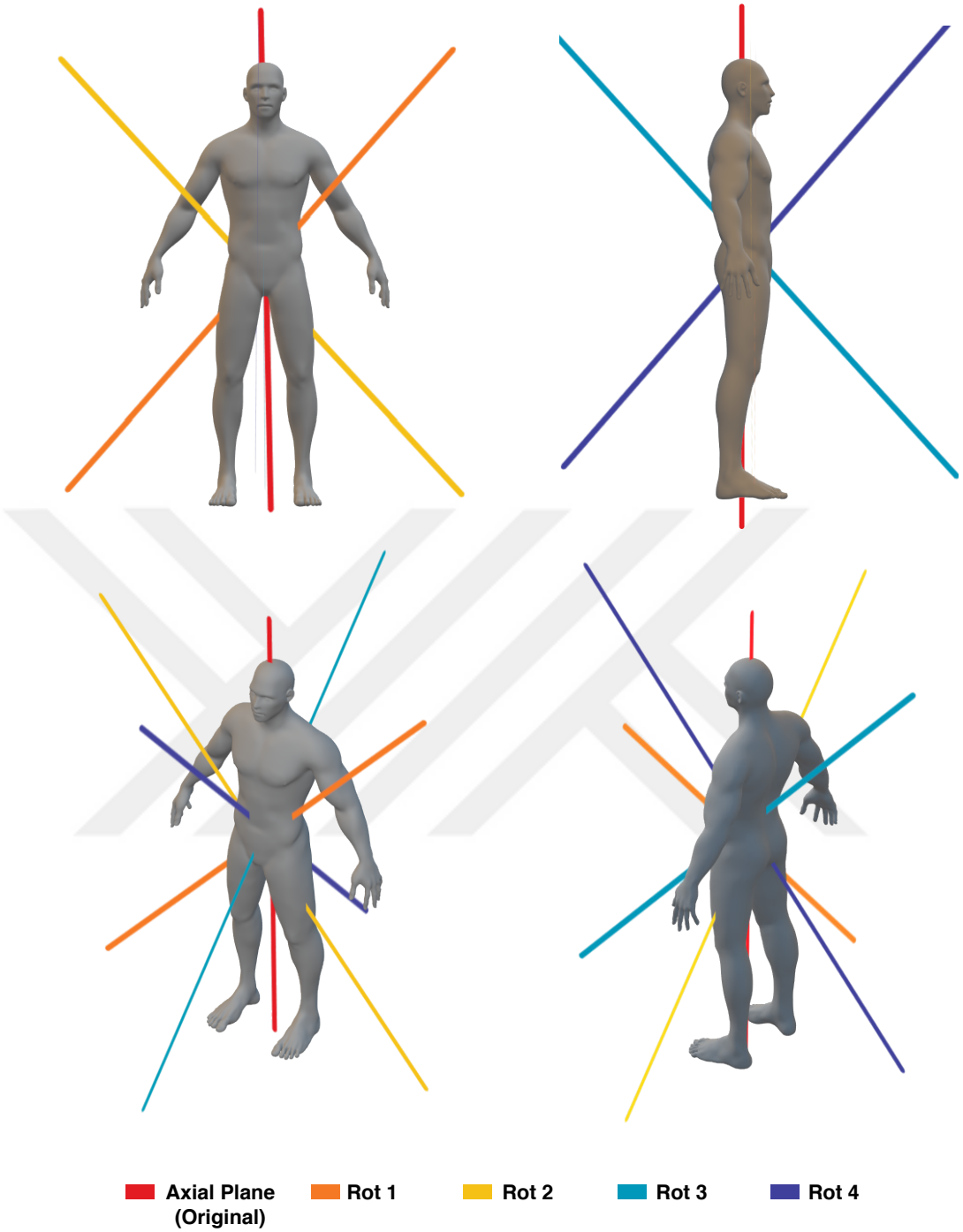

Figure 5.7 Illustration of different rotations from different views

5.3.4 Evaluation

The same evaluation strategy with CHAOS challenge and "Fusion of different methods for segmentation of the liver" sections was used for evaluation. Thus, consistency and cohesion are provided in the evaluation of all the studies presented in this thesis.

In order to keep the evaluation process compact, hepatic arteries and portal veins are considered as a single class. In other words, both individual CNNs and ensemble methods have only two class targets: all veins in the liver and background.

In addition to the evaluation of segmentation, statistical significances are calculated for paired and non-paired samples as explained in Section 4.5. The results of metrics and their significance against each other reveal the potential of ensemble systems. These findings are presented in Section 5.3.5 with various tables and illustrations.

5.3.5 Results

The positive results of using ensemble methods for segmentation of liver encouraged us to apply this strategy to a more complicated and less popular problem: segmentation of liver veins from CT angiography. The same analyses with the previous section (4.6) have been performed. The segmentation accuracies have been evaluated as well as their statistical significance. Also, the overfitting problem was examined in the previous section. The following tables, figures, and discussions were presented by the analyses on the VEELA dataset explained in Section 5.3.1.

Comparison of Table 5.4 with Table 5.5 shows the overfitting problem of individual segmentations still exists for vein segmentation. Again overfitting in DMs directly affects the results of the ensemble methods. The consistency in the results of the liver and its vein segmentation shows that the overfitting problem does not depend on the type of segmentation target. Overfitting magnitude, presented in Table 5.6,

again supports the findings in whole liver segmentation.

Table 5.4 Metric results on VEELA training data for the individual segmenters and the ensemble methods to examine overfitting. The circle marker indicates results where the overfitting was not found to be significant

	DICE	RAVD	ASSD	MSSD
DeepMedic Axial	∘ 0.834	8.702	3.170	91.687
DeepMedic Rot 1	0.796	10.379	4.796	96.298
DeepMedic Rot 2	0.795	10.812	∘ 4.099	95.593
DeepMedic Rot 3	0.792	10.533	4.262	97.865
DeepMedic Rot 4	0.795	10.351	4.335	96.245
Majority Vote	0.913	7.999	1.640	75.806
Average	0.924	6.777	1.200	71.038
Product	0.911	7.540	1.565	73.020
Min-Max	0.924	7.971	1.417	74.950
Logit	0.926	6.481	1.278	69.909

Table 5.5 Metric results on VEELA test data for the individual segmenters and the ensemble methods to examine segmentation accuracy. The best value in each column is marked bold

	DICE	RAVD	ASSD	MSSD
DeepMedic Axial	0.714	16.902	5.370	177.687
DeepMedic Rot 1	0.664	25.274	3.069	152.791
DeepMedic Rot 2	0.664	25.585	3.615	161.270
DeepMedic Rot 3	0.705	21.623	2.375	179.027
DeepMedic Rot 4	0.672	25.496	3.109	162.213
Majority Vote	0.820	39.657	1.926	104.285
Average	0.826	13.081	**1.831**	102.131
Product	0.857	16.097	1.897	116.950
Min-Max	0.830	21.561	1.782	117.281
Logit	**0.876**	**11.470**	1.877	**99.150**

Table 5.6 Overfitting magnitude for the VEELA dataset. Large overfitting corresponds to blue color and small overfitting, to red color. Each column (metric) is scaled individually

	DICE	RAVD	ASSD	MSSD
DeepMedic Axial	0.3170	-38.5777	-3.0324	-151.5513
DeepMedic Rot 1	0.2853	-21.6107	-1.4302	-107.7594
DeepMedic Rot 2	0.2665	-20.2772	-1.0357	-110.5563
DeepMedic Rot 3	0.2786	-20.4994	-1.3302	-107.3145
DeepMedic Rot 4	0.2831	-23.1188	-2.3688	-102.1495
Majority	0.2380	-19.6891	-1.2913	-112.7926
Average	0.2646	-22.5956	-0.9401	-118.2879
Product	0.2277	-18.6504	-1.0803	-114.9142
Min-Max	0.2546	-24.4913	-1.0453	-125.8247
Logit	0.2170	-14.8510	-0.6385	-98.9364

On the other hand, there is a significant difference in Table 5.4 and Table 5.5 with Table 4.2 and Table 4.3. Here, the segmentation accuracy of DMs can be considered as insufficient. In other words, the vanilla style of DMs may not have the capability of vein segmentation. Performances are poor even testing with the train data presented in Table 5.4. Therefore, if a single DM will be used for this problem, it is necessary to make huge efforts for tuning the system and developing tailored designs. On the other hand ensemble, segmenters offer better results than individual DMs again. This superiority of ensembles has been analyzed in terms of their significance. They are presented in Tables 5.7 – 5.10.

Tables 5.7 – 5.10 show that the statistical significance of results are less than whole liver segmentation problem. However, it is still the same that, ensembles outperform

Table 5.7 DICE: Statistical comparison between individual DMs and ensembles. Bullet means that the ensemble wins; circle means that the DM wins; line means that no statistical difference

	Majority	Average	Product	Min/Max	Logit
DeepMedic Axial	–	○	–	●	●
DeepMedic Rot 1	●	–	–	–	–
DeepMedic Rot 2	●	–	–	–	●
DeepMedic Rot 3	–	–	–	–	–
DeepMedic Rot 4	–	–	–	–	–

Table 5.8 RAVD: Statistical comparison between individual DMs and ensembles. Bullet means that the ensemble wins; circle means that the DM wins; line means that no statistical difference

	Majority	Average	Product	Min/Max	Logit
DeepMedic Axial	–	–	○	–	●
DeepMedic Rot 1	–	○	–	●	–
DeepMedic Rot 2	–	○	–	○	–
DeepMedic Rot 3	–	○	–	–	–
DeepMedic Rot 4	–	○	–	●	●

Table 5.9 ASSD: Statistical comparison between individual DMs and ensembles. Bullet means that the ensemble wins; circle means that the DM wins; line means that no statistical difference

	Majority	Average	Product	Min/Max	Logit
DeepMedic Axial	–	●	–	●	●
DeepMedic Rot 1	–	–	–	–	–
DeepMedic Rot 2	–	–	●	–	●
DeepMedic Rot 3	○	–	–	–	●
DeepMedic Rot 4	–	●	–	–	–

Table 5.10 MSSD: Statistical comparison between individual DMs and ensembles. Bullet means that the ensemble wins; circle means that the DM wins; line means that no statistical difference

	Majority	Average	Product	Min/Max	Logit
DeepMedic Axial	–	•	–	–	•
DeepMedic Rot 1	–	–	•	•	–
DeepMedic Rot 2	–	–	•	–	•
DeepMedic Rot 3	–	–	–	–	–
DeepMedic Rot 4	•	–	•	–	•

the individual methods. Average combiner showed the least significant results for different metrics. It is followed by the average combiner. Although the lower significance of Logit combiner in this problem comparing with whole liver segmentation, it is still the most significant fusion strategy according to the Tables 5.7 – 5.10.

To determine which ensemble method is the most preferable, again glyph plot of them has been generated. The glyph plot in Figure 5.8 shows the Logit combiner has the most accurate performance over other methods. The balanced error distribution of Logit combiner on four metrics also supports its robustness. The plots again indicate that the average based methods such as logit combiner and simple average combiner have higher segmentation accuracy.

To sum up, the effectiveness of the DMs and their ensembles have been examined by a much more tough problem in this section. Segmentation of liver veins from CT angiography is still pushing the DMs to their boundaries. Since very few solutions have been proposed in the literature for this problem, there is no definite anticipation of whether this problem can be solved by DMs alone. However, even if there is a solution by single DM, it is an undeniable fact that it will require much more complex designs, longer development processes, and more processing power. On the other

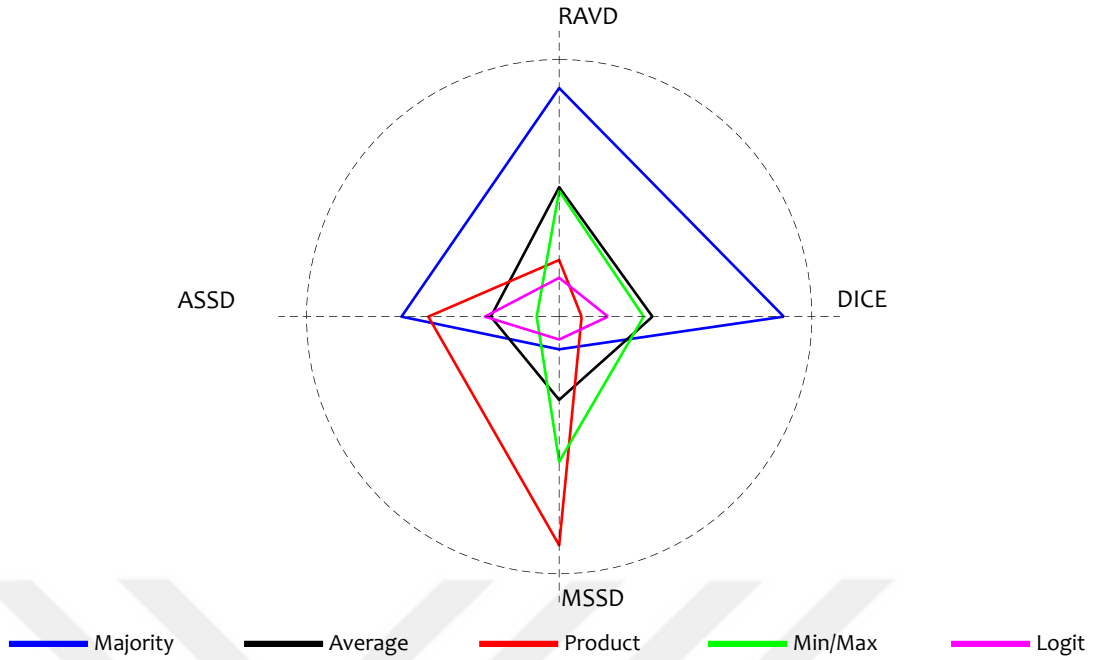

Figure 5.8 Glyph plot of the four ensemble methods for the VEELA dataset. The spokes are the four metrics. Small-area ensembles are preferable

hand, our analyses revealed that the power of ensemble strategies can help to reduce these drawbacks. Ensembles can be adapted for the segmentation of liver vessels problem to reduce overfitting and to boost segmentation accuracy. Considering their lower performance for the segmentation of liver veins comparing with the whole liver, further strategies for developing novel fusion methods can be studied.

According to the success of Logit combiner against other ones, the further ensemble method designed to handle this problem should include calibration. However, there is a certain fact that this problem has not been sufficiently studied yet. Understanding the potentials of various methods plays an important role in solving the problem of liver vein segmentation. To enrich the level of understanding, further studies such as organizing a new challenge for liver vein segmentation is planned. The overall conclusion of this section is that there is still the need for further studies to achieve better segmentation results for liver vein segmentation. Our findings show that ensemble strategies will play an important role in these studies.

CHAPTER SIX
CONCLUSIONS

In this thesis, segmentation in medical image analysis is discussed from different perspectives. The studies focused on creating new challenges to enrich level of information in abdomen image segmentation field and adapting classifier ensemble methods to segmentation of the liver and its vein. Also, a new ensemble method, that outperformed former methods, has been developed. The contributions of the studies in this thesis can be summed as:

- Instead of focusing only on the development of more successful segmentation methods, we went down to the root of the problem and examined all the newest methods in the present as well as in the past. For these purposes, considerable time was spent on preparing new data and not only for creating benchmarking systems but also other scientific works in the literature.

- The first medical image segmentation challenge was held in Turkey. Then, CHAOS, the most popular challenge in its field (with more than 1500 participants and 550 submissions), was organized. The CHAOS challenge data (Kavur et al., 2019) has reached more than 3,226 single views, 3,159 unique downloads, and 14,066 total downloads when this thesis was written.

  The CHAOS challenge was designed not only to examine current deep learning developments but also to measure their capabilities in potential matters such as segmentation in cross-modality data in the future. In this way, it is aimed at the information presented in this thesis will guide many further studies.

- The analyses coming from our challenges were used to exploit underestimated problems of them analysis. According to our knowledge, it is the first time peeking problem was widely discussed in a challenge study. The feedback that we had from our article reviewers acclaims the importance of this problem. It is planned to expand the sensitivity to this problem as well as developing new analytical tools to prevent it.

- Unlike so many studies that only present just the pros of DMs, the provided snapshot of the current state-of-the-art abdomen organ segmentation methods revealed the cons of DMs. Although it is an indisputable fact that future segmentation methods will be based on DMs, ensemble methods are also expected to be integrated into these solutions. As the studies in this thesis show, integration methods can solve the problems of DMs such as reproducibility and eliminate the obstacles between their academic studies and real-life applications.

  In this way, it may be possible to solve real-life problems with DMs, which is the main purpose of science. It is hoped that the studies in this thesis will facilitate the application of DMs to real-life problems through methods of fusion of the results.

- In addition to using DMs as independent input blocks and combining them with simple rules can be used to eliminate problems of DMs, also there is no need to have an individual design for specific problems. The general message for this analysis is that there is an alternative way to reach promising segmentation results via DMs without having high programming skills.

- A new ensemble method (Logit combiner) has been designed and implemented on medical image segmentation problem. Applying a transformation to the output of DMs increased segmentation accuracy as well as reduced overfitting. Although it is not possible to use complex but successful ensemble methods such as bagging, boosting due to the limited amount of the data, training a few parameters can be used as indicated in this thesis.

In future, it is planned to apply integration methods to more challenging problems rather than healthy liver segmentation in light of the results in this thesis. The studies on the segmentation of liver veins are the first attempt for this purpose in this thesis. It is expected to integrate fusion approaches to further studies on different organs such as the brain and different modalities such as fMRI. To enrich the knowledge in this field, a novel challenge will be held.

REFERENCES

Abraham, N., & Khan, N. M. (2019). A novel focal tversky loss function with improved attention U-Net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging*, 683–687.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, 1–6.

Alpaydin, E. (2014). *Introduction to Machine Learning*. Cambridge: MIT Press.

Ayache, N., & Duncan, J. (2016). 20th anniversary of the medical image analysis journal (MedIA). *Medical Image Analysis*, *33*, 1–3.

Bal, E., Klang, E., Amitai, M., & Greenspan, H. (2018). Automatic liver volume segmentation and fibrosis classification. *Medical Imaging 2018: Computer-Aided Diagnosis*, *10575*, 34–39.

Bilic, P., Christ, P. F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C., Han, X., Heng, P., Hesser, J., Kadoury, S., Konopczynski, T. K., Le, M., Li, C., Li, X., Lipková, J., Lowengrub, J. S., Meine, H., Moltz, J. H., Pal, C., Piraud, M., Qi, X., Qi, J., Rempfler, M., Roth, K., Schenk, A., Sekuboyina, A., Zhou, P., Hülsemeyer, C., Beetz, M., Ettlinger, F., Grün, F., Kaissis, G., Lohöfer, F., Braren, R., Holch, J., Hofmann, F., Sommer, W. H., Heinemann, V., Jacobs, C., Mamani, G. E. H., van Ginneken, B., Chartrand, G., Tang, A., Drozdzal, M., Ben-Cohen, A., Klang, E., Amitai, M. M., Konen, E., Greenspan, H., Moreau, J., Hostettler, A., Soler, L., Vivanti, R., Szeskin, A., Lev-Cohain, N., Sosna, J., Joskowicz, L., & Menze, B. H. (2019). The liver tumor segmentation benchmark (LiTS). *ArXiv*, *abs/1901.04056*, 1–43.

Cerrolaza, J. J., Picazo, M. L., Humbert, L., Sato, Y., Rueckert, D., Ángel González Ballester, M., & Linguraru, M. G. (2019). Computational anatomy for multi-organ analysis in medical imaging: A review. *Medical Image Analysis*, *56*, 44–67.

Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., & Rueckert, D. (2020). Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, *7*, 25–26.

Chlebus, G., Schenk, A., Moltz, J. H., van Ginneken, B., Hahn, H. K., & Meine, H. (2018). Automatic liver tumor segmentation in ct with fully convolutional neural networks and object-based postprocessing. *Nature Scientific Reports*, *8*(1), 2045–2322.

Choi, J., Lee, K., Jeong, W.-K., Chun, S. Y., & Park, P. (2019). *PAIP 2019 challenge*. Retrieved April 1, 2020, from https://paip2019.grand-challenge.org/.

Christ, P. F., Elshaer, M. E. A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., Rempfler, M., Armbruster, M., Hofmann, F., D'Anastasi, M., Sommer, W. H., Ahmadi, S.-A., & Menze, B. H. (2016). Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3D conditional random fields. In *Medical Image Computing and Computer-Assisted Intervention*, Cham: Springer International Publishing, 415–423.

Codella, N. C. F., Nguyen, Q., Pankanti, S., Gutman, D. A., Helba, B., Halpern, A. C., & Smith, J. R. (2017). Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, *61*(5), 1–5.

Conze, P.-H., Pons, C., Burdin, V., Sheehan, F. T., & Brochard, S. (2019). Deep convolutional encoder-decoders for deltoid segmentation using healthy versus pathological learning transferability. In *IEEE International Symposium on Biomedical Imaging*, 36–39.

Cootes, T., Taylor, C., Cooper, D., & Graham, J. (1995). Active shape models-their training and application. *Computer Vision and Image Understanding*, *61*(1), 38–59.

Dede, M. A., Aptoula, E., & Genc, Y. (2019). Deep network ensembles for aerial scene classification. *IEEE Geoscience and Remote Sensing Letters*, *16*(5), 732–735.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A

Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Deng, X., & Du, G. (2008). 3D segmentation in the clinic: a grand challenge II-liver tumor segmentation. In *The Medical Image Computing and Computer Assisted Intervention Society (MICCAI) workshop*. Retrieved June 12, 2020, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.502.8168&rep=rep1&type=pdf.

Dolz, J., Desrosiers, C., & Ayed, I. B. (2018). IVD-Net: intervertebral disc localization and segmentation in mri with a multi-modal unet. In *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*, Switzerland: Springer, 130–143.

Dou, Q., Liu, Q., Heng, P. A., & Glocker, B. (2020). Unpaired multi-modal segmentation via knowledge distillation. *IEEE Transactions on Medical Imaging*, 1–11.

Dubois, D., & Prade, H. (1985). A review of fuzzy set aggregation connectives. *Information Sciences*, *36*(1), 85–121.

Fischer, F., Alper Selver, M., Hillen, W., & Guzelis, C. (2010). Integrating segmentation methods from different tools into a visualization program using an object-based plug-in interface. *IEEE Transactions on Information Technology in Biomedicine*, *14*(4), 923–934.

Gee, J. C., Reivich, M., & Bajcsy, R. (1993). Elastically deforming 3D atlas to match anatomical brain images. *Journal of Computer Assisted Tomography*, *17*(2), 225–236.

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S. P., Clarkson, M. J., & Barratt, D. C. (2018). Automatic Multi-Organ Segmentation on Abdominal CT with Dense V-Networks. *IEEE Transactions on Medical Imaging*, *37*(8), 1822–1834.

Gilroy, A. M. (2013). *Anatomy: An Essential Textbook*. Leipzig: Thieme.

Goldenberg, R., Kimmel, R., Rivlin, E., & Rudzsky, M. (2001). Fast geodesic active contours. *IEEE Transactions on Image Processing*, *10*(10), 1467–1475.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.

Google Inc (2020). *Kaggle*. Retrieved April 1, 2020, from https://www.kaggle.com/.

Greenspan, H., van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, *35*(5), 1153–1159.

Grofman, B., Owen, G., & Feld, S. (1983). Thirteen theorems in search of the truth. *Theory and Decisions*, *15*, 261–278.

Guinney, J., Wang, T., Laajala, T. D., Winner, K. K., Bare, J. C., Neto, E. C., Khan, S. A., Peddinti, G., Airola, A., & Pahikkala, T. (2017). Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. *The Lancet Oncology*, *18*(1), 132–142.

Guo, Y., & Ashour, A. S. (2019). Neutrosophic sets in dermoscopic medical image segmentation. In *Neutrosophic Set in Medical Image Analysis*, 229 – 243. New York: Academic Press.

Han, S., He, Y., Carass, A., Ying, S. H., & Prince, J. L. (2019). Cerebellum parcellation with convolutional neural networks. In *Medical Imaging 2019: Image Processing*, vol. 10949, *International Society for Optics and Photonics*, 1–13.

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

Heimann, T., van Ginneken, B., & Styner, M. (2009). Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets. *IEEE Transactions on Medical Imaging*, *28*(8), 1251–1265.

Hesselink, J. R. (2020). *Basic principles of MR imaging*. Retrieved June 1, 2020, from http://spinwarp.ucsd.edu/neuroweb/Text/br-100.htm.

Hilbe, J. M. (2009). *Logistic Regression Models*. Florida: Chapman and Hall.

Hirokawa, Y., Isoda, H., Maetani, Y. S., Arizono, S., Shimada, K., & Togashi, K. (2008). MRI artifact reduction and quality improvement in the upper abdomen with propeller and prospective acquisition correction (PACE) technique. *American Journal of Roentgenology*, *191*(4), 1154–1158.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., & Weinberger, K. Q. (2017). Snapshot ensembles: Train 1, get M for free. *arXiv*, *abs/1704.00109*, 1–14.

Ibáñez, L., Schroeder, W., Ng, L., Cates, J., Consortium, T. I. S. C., & Hamming, R. (2003). *The ITK Software Guide*. Retrieved April 24, 2020, from https://itk.org/ItkSoftwareGuide.pdf.

Iglovikov, V., & Shvets, A. (2018). Ternausnet: U-Net with VGG11 encoder pre-trained on Imagenet for image segmentation. *arXiv*, *1801.05746*, 1–5.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, 448–456.

IRCAD (2009). *3D image reconstruction for comparison of algorithm database*. Retrieved April 1, 2020, from https://www.ircad.fr/research/3d-ircadb-01/.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., & Maier-Hein, K. H. (2019). nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. In *Informatik aktuell*, 22–23.

Jimenez-del-Toro, O., Müller, H., Krenn, M., Gruenberg, K., Taha, A. A., Winterstein, M., Eggel, I., Foncubierta-Rodríguez, A., Goksel, O., Jakab, A., Kontokotsios, G., Langs, G., Menze, B. H., Salas Fernandez, T., Schaer, R., Walleyo, A., Weber,

M., Dicente Cid, Y., Gass, T., Heinrich, M., Jia, F., Kahl, F., Kechichian, R., Mai, D., Spanier, A. B., Vincent, G., Wang, C., Wyeth, D., & Hanbury, A. (2016). Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. *IEEE Transactions on Medical Imaging*, *35*(11), 2459–2475.

Jones, J. (2019). *Couinaud classification of hepatic segments: Radiology reference article*. Retrieved April 24, 2020, from https://radiopaedia.org/articles/couinaud-classification-of-hepatic-segments?lang=gb.

Ju, C., Bibaut, A., & van der Laan, M. (2018). The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, *45*(15), 2800–2818.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., & Glocker, B. (2018). Ensembles of multiple models and architectures for robust brain tumour segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Cham: Springer, 450–462.

Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A. V., Criminisi, A., Rueckert, D., & Glocker, B. (2016). DeepMedic for brain tumor segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10154, Cham: Springer, 138–149.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., & Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, *36*, 61–78.

Kavur, A. E., Gezer, N. S., Barış, M., Conze, P.-H., Groza, V., Pham, D. D., Chatterjee, S., Ernst, P., Özkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., Rajan, R., Aslan, S., Sheet, D., Dovletov, G., Speck,

O., Nürnberger, A., Maier-Hein, K. H., Akar, G. B., Ünal, G., Dicle, O., & Selver, M. A. (2020a). CHAOS Challenge-Combined (CT-MR) Healthy Abdominal Organ Segmentation. *arXiv pre-print*, 1–19.

Kavur, A. E., Gezer, N. S., Barış, M., Şahin, Y., Şavaş Özkan, Baydar, B., Yüksel, U., Çağlar Kılıkçıer, Şahin Olut, Akar, G. B., Ünal, G., Dicle, O., & Selver, M. A. (2020b). Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors. *Diagnostic and Interventional Radiology*, *26*, 11–21.

Kavur, A. E., & Selver, M. A. (2019). *CHAOS-Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge* . Retrieved July 8, 2020, from https://chaos.grand-challenge.org/.

Kavur, A. E., Selver, M. A., Dicle, O., Barış, M., & Gezer, N. S. (2019). *CHAOS-Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data*. Retrieved April 1, 2020, from http://doi.org/10.5281/zenodo.3362844.

Kistler, M., Bonaretti, S., Pfahrer, M., Niklaus, R., & Büchler, P. (2013). The virtual skeleton database: an open access repository for biomedical research and collaboration. *Journal of medical Internet research*, *15*(11), 245–251.

Kozubek, M. (2016). Challenges and benchmarks in bioimage analysis. In *Focus on Bio-Image Informatics*, Cham: Springer, 231–262.

Kroon, D.-J. (2009). *Accurate Fast Marching*. Retrieved April 2, 2020, from http://www.mathworks.com/matlabcentral/fileexchange/24531-accurate-fast-marching.

Kumar, V., Abbas, A., Fausto, N., & Aster, J. (2010). *Robbins and Cotran Pathologic Basis of Disease*. Saunders: Elsevier.

Kuncheva, L. I. (2014). *Combining Pattern Classifiers: Methods and Algorithms*, vol. 9781118315. Hoboken: Wiley-Interscience.

Lachinov, D. (2019). Segmentation of thoracic organs using pixel shuffle. In *Proceedings of the 2019 challenge on segmentation of thoracic organs at*

risk in CT images. Retrieved May 12, 2020, from http://ceur-ws.org/Vol-2349/SegTHOR2019_paper_10.pdf.

Li, F., Neverova, N., Wolf, C., & Taylor, G. (2016). Modout: Learning to fuse modalities via stochastic regularization. *Journal of Computational Vision and Imaging Systems*, *2*(1), 25– 28.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C., & Heng, P. (2018). H-DenseUNet: Hybrid densely connected U-Net for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging*, *37*(12), 2663–2674.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

Lu, F., Wu, F., Hu, P., Peng, Z., & Kong, D. (2017). Automatic 3D liver location and segmentation via convolutional neural network and graph cut. *International Journal of Computer Assisted Radiology and Surgery*, *12*, 171–182.

Ma, S., & Chu, F. (2019). Ensemble deep learning-based fault diagnosis of rotor bearing systems. *Computers in Industry*, *105*, 143–152.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., Maier, O., Maier-Hein, K., Menze, B. H., Müller, H., Neher, P. F., Niessen, W., Rajpoot, N., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A. A., van der Sommen, F., Wang, C. W., Weber, M. A., Zheng, G., Jannin, P., & Kopp-Schneider, A. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, *9*(1), 5217–5230.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A. L., Arbel, T., Eisenmann, M., Hanbuary, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken,

B., Kopp-Schneider, A., & Landman, B. (2019). Bias: Transparent reporting of biomedical image analysis challenges. *arXiv preprint arXiv:1910.04071*, 1–36.

Maji, D., Santara, A., Mitra, P., & Sheet, D. (2016). Ensemble of deep convolutional neural networks for learning to detect retinal vessels in fundus images. *arXiv*, *abs/1603.04833*, 1–4.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, C., Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M. S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H., Shotton, J., Silva, C. A., Sousa, N., Subbanna, N. K., Szekely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Unal, G., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., & Van Leemput, K. (2015). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, *34*(10), 1993–2024.

Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings 4th International Conference on 3D Vision*, 565–571.

Moghbel, M., Mashohor, S., Mahmud, R., & Saripan, M. I. (2018). Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography. *Artificial Intelligence Review*, *50*(4), 497–537.

Molina, D. K., & DiMaio, V. J. M. (2012). Normal organ weights in men: Part II – The brain, lungs, liver, spleen, and kidneys. *The American Journal of Forensic Medicine and Pathology*, *33*(4), 368–372.

Nandamuri, S., China, D., Mitra, P., & Sheet, D. (2019). Sumnet: Fully convolutional

model for fast segmentation of anatomical structures in ultrasound volumes. *arXiv*, *1901.06920*, 1–9.

Netter, F. H. (2018). *Atlas of Human Anatomy*. Netter Basic Science. Wiesbaden: Elsevier.

Nielsen, M. A. (2019). *Neural networks and deep learning*. Retrieved April 1, 2020, from http://neuralnetworksanddeeplearning.com/.

Ortiz, A., Munilla, J., Górriz, J. M., & Ramírez, J. (2016). Ensembles of deep learning architectures for the early diagnosis of the alzheimer's disease. *International Journal of Neural Systems*, *26*(07), 165–176.

Oza, N. C., & Tumer, K. (2008). Classifier ensembles: Select real-world applications. *Information Fusion*, *9*(1), 4 – 20.

Papamarkos, N. (1989). A program for the optimum approximation of real rational functions via linear programming. *Advances in Engineering Software*, *11*(1), 37–48.

Papamarkos, N., & Gatos, B. (1994). A New Approach for Multilevel Threshold Selection. *CVGIP: Graphical Models and Image Processing*, *56*(5), 357–370.

Pelleg, D., & Moore, A. (1999). Accelerating exact k-means algorithms with geometric reasoning. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 277–281.

Peyre, G. (2004). *Toolbox Fast Marching*. Retrieved June 12, 2020, from http://www.mathworks.com/matlabcentral/fileexchange/6110-toolbox-fast-marching.

Press, W. H., Flannery, B. P., & Teukolsky, S. A. (1986). *Numerical recipes. The art of scientific computing*. Cambridge: University Press.

Prevedello, L. M., Halabi, S. S., Shih, G., Wu, C. C., Kohli, M. D., Chokshi, F. H., Erickson, B. J., Kalpathy-Cramer, J., Andriole, K. P., & Flanders, A. E. (2019). Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. *Radiology: Artificial Intelligence*, *1*(1), 26–33.

Razinkov, E., Saveleva, I., & Matas, J. (2018). Alfa: Agglomerative late fusion algorithm for object detection. In *2018 24th International Conference on Pattern Recognition (ICPR)*, 2594–2599.

Reinke, A., Eisenmann, M., Onogur, S., Stankovic, M., Scholz, P., Full, P. M., Bogunovic, H., Landman, B. A., Maier, O., Menze, B., et al. (2018a). How to exploit weaknesses in biomedical challenge design and organization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 388–395.

Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., et al. (2018b). Is the winner really the best? a critical analysis of common research practice in biomedical image analysis competitions. *arXiv*, *1806.02051*, 1–13.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1), 1–39.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, Cham: Springer, 234–241.

Selver, M. A., & Kavur, A. E. (2015). Implementation and use of 3D pairwise geodesic distance fields for seeding abdominal aortic vessels. *International Journal of Computer Assisted Radiology and Surgery*, *11*, 803–816.

Selvi, E., Selver, M. A., Kavur, A. E., Guzelis, C., & Dicle, O. (2015). Segmentation of abdominal organs from MR images using multi-level hierarchical classification. *Journal of the Faculty of Engineering and Architecture of Gazi University*, *30*(3), 533–546.

Shapiro, L. G., & Stockman, G. C. (2001). *Computer vision*. New Jersey: Prentice Hall.

Shepherd, R. W., & Turmelle, Y. P. (2017). Portal hypertension in children. In *Blumgart's Surgery of the Liver, Biliary Tract and Pancreas, 2-Volume Set*, Philadelphia: Elsevier.

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

Shvets, A. A., Rakhlin, A., Kalinin, A. A., & Iglovikov, V. I. (2018). Automatic instrument segmentation in robot-assisted surgery using deep learning. In *17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 624–628.

Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B. A., Litjens, G., & Menze, B. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv*, *1902.09063*, 1–12.

Soille, P. (1994). Generalized geodesy via geodesic time. *Pattern Recognition Letters*, *15*(12), 1235–1240.

Soille, P. (2003). *Morphological Image Analysis: Principles and Applications*. New York: Springer.

Song, Y., & Yan, H. (2017). Image segmentation techniques overview. In *2017 Asia Modelling Symposium (AMS)*, 103–107.

Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A., & van Ginneken, B. (2004). Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, *23*(4), 501–509.

Starmans, M. P., van der Voort, S. R., Castillo Tovar, J. M., Veenland, J. F., Klein, S., & Niessen, W. J. (2020). Radiomics: Data mining using quantitative medical image features. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, 429 – 456. New York: Academic Press.

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BioMed Central medical imaging*, *15*, 29–38.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2017). Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6924–6932.

Valindria, V. V., Pawlowski, N., Rajchl, M., Lavdas, I., Aboagye, E. O., Rockall, A. G., Rueckert, D., & Glocker, B. (2018). Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 547–556.

van Ginneken, B., Heimann, T., & Styner, M. (2007). *3D segmentation in the clinic: A grand challenge*. Retrieved June 12, 2020, from http://sliver07.org/p7.pdf.

van Ginneken, B., & Kerkstra, S. (2015). *Grand challenges in biomedical image analysis*. Retrieved April 1, 2020, from http://grand-challenge.org/.

Vorontsov, E., Cerny, M., Régnier, P., Di Jorio, L., Pal, C. J., Lapointe, R., Vandenbroucke-Menu, F., Turcotte, S., Kadoury, S., & Tang, A. (2019). Deep learning for automated segmentation of liver lesions at CT in patients with colorectal cancer liver metastases. *Radiology: Artificial Intelligence*, *1*(2), 64–71.

Warfield, S. K., Zou, K. H., & Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, *23*(7), 903–921.

Weight, C., Papanikolopoulos, N., Kalapara, A., & Heller, N. (2019). *KiTS19 challenge*. Retrieved April 1, 2020, from https://kits19.grand-challenge.org/.

Wikimedia Commons (2020). *Computed tomography*. Retrieved April 1, 2020, from https://commons.wikimedia.org/wiki/File:UPMCEast_CTscan.jpg.

Winter, J., & Auer, R. A. C. (2012). Metastatic malignant liver tumors: Colorectal cancer. In *Blumgart's Surgery of the Liver, Pancreas and Biliary Tract*, 1290–1304. Philadelphia: W.B. Saunders.

Wu, Y., & He, K. (2018). Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.

Yan, Y., Conze, P.-H., Decencière, E., Lamard, M., Quellec, G., Cochener, B., & Coatrieux, G. (2019). Cascaded multi-scale convolutional encoder-decoders for breast mass segmentation in high-resolution mammograms. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 6738–6741.

Yang, X., Yang, J. D., Hwang, H. P., Yu, H. C., Ahn, S., Kim, B.-W., & You, H. (2018). Segmentation of liver and vessels from ct images and classification of liver segments for preoperative liver surgical planning in living donor liver transplantation. *Computer Methods and Programs in Biomedicine*, *158*, 41–52.

Yeghiazaryan, V., Voiculescu, I., Yeghiazaryan, V., & Voiculescu, I. (2015). An overview of current evaluation methods used in medical image segmentation. Technical report, Department of Computer Science, Oxford, UK.

Zhang, C., Lim, P., Qin, A. K., & Tan, K. C. (2017). Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2306–2318.

Zhang, C., & Ma, Y. (2012). *Ensemble Machine Learning: Methods and Applications*. New York: Springer.

Zheng, J., Cao, X., Zhang, B., Zhen, X., & Su, X. (2019). Deep ensemble machine for video classification. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(2), 553–565.

Zhou, S., Greenspan, H., & Shen, D. (2017). *Deep learning for medical image analysis*. Wiesbaden: Elsevier.