JOSIP JURAJ STROSSMAYER UNIVERSITY OF OSIJEK
FACULTY OF ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION TECHONOLGY OSIJEK

Graduate University Study Program in Computer Engineering

SEGMENTATION OF HEART CHAMBERS FROM 2D MRI IMAGES USING THE U-NET CONVOLUTIONAL NEURAL NETWORK

Master's thesis

Marko Grubeša

Osijek, 2021.

Table of contents

1. INTRODUCTION

Segmentation is one of the key problems in the field of computer vision and image processing. The goal of segmentation is to group pixels into prominent areas of the image, that is areas that correspond to individual parts of an image or object [1,2]. Such a technique gives us a more detailed understanding of the object in the picture. Segmentation is widely applied in various fields. Some examples are self-driving cars, virtual reality, human-computer interaction, and it is also greatly applied in the field of medical image processing. Medical image processing is used to obtain images of body parts for medical use to identify or study diseases [3]. The field of medical image processing is developing extremely fast, mostly due to the development of image processing techniques, which includes analysis, recognition, improvement, and segmentation an image. Medical image processing involves the use and research of three-dimensional data sets, most commonly obtained using computed tomography (CT) or magnetic resonance imaging (MRI) [4]. Magnetic resonance imaging is a medical imaging technique used in radiology to shape images of human anatomy and physiological processes in the body [5]. Using this technique, images of heart chambers are obtained on which medical image processing can be performed. In these images, various methods are implemented to remove noise and improve image quality, with deep learning methods increasingly being used.

Deep learning is a branch of machine learning, which is, in turn, a branch of the field of artificial intelligence. Deep learning provides a set of algorithms and methods that can be used to solve problems that people perform intuitively and almost automatically, -but are otherwise very challenging for computers [6]. The U-Net convolutional neural network, based on deep learning, has shown exceptional results in the field of medical image segmentation and is therefore often used for various cardiovascular image processing tasks.

1.1. The aim of the thesis

The aim of this thesis is to explore and describe the method of obtaining MRI (Magnetic Resonance Imaging) images and their characteristics. Then, describe the clinical background (representation of the heart on MRI images, parts of the heart, clinical need) and give a brief overview of the areas of previously developed methods for segmentation of the heart from MRI images. It is necessary to explain the theoretical foundations of the way convolutional neural networks work and to develop a method for pre-processing MRI images. It is necessary to develop

a system for whole heart segmentation using a U-net convolutional neural network and compare the obtained segmentation results with and without pre-processing of data and show the accuracy of the developed system.

2. CURRENT SEGMENTATION METHODS

U-Net is considered one of the standard CNN (Convolutional Neural Network) architectures for image classification tasks, when we need not only to define the whole image according to its class but also to segment image areas by class, that is to produce a mask that will separate the image into several class [7]. It is most commonly used in biomedicine to segment the heart, brain, and other parts of the body.

Sander and others [8] described automatic segmentation with local error detection in MRI images of the heart. This technique uses 3 convolutional neural networks. Each network receives a CMR (Cardiac Magnetic Resonance) image as an input, and each of them has four output channels that show the probabilities for the three heart structures and the background. Subsequently, an auxiliary CNN that analyzes the input image was used to detect and correct local segmentation errors. After the analysis, a map is created showing the areas of failed segmentation. Figure 2.1 shows the steps of the described method of automatic segmentation with local error detection.

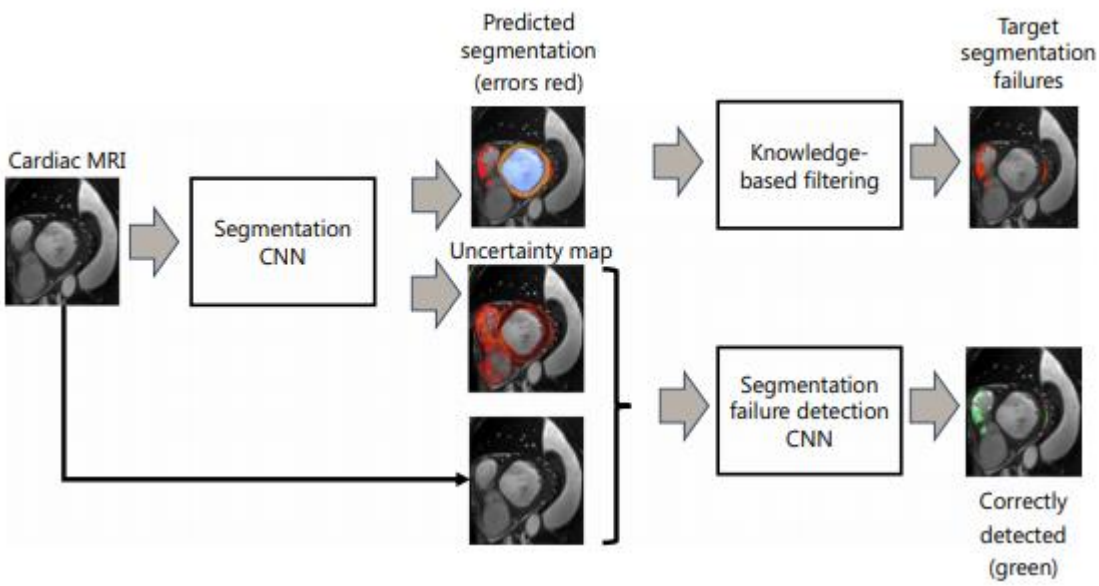


Figure 2.1 Automatic segmentation steps with local error detection [8]

Shewaye [9] presented a segmentation method using the region growth technique. The idea is to first select an initial pixel and then merge all adjacent pixels with that pixel if they meet predetermined uniformity criteria. Each time a pixel is added to a region, the mean value of the

region changes taking the newly added pixel into account. The region growth algorithm follows the following steps: select the initial pixel, calculate the measure of similarity of each adjacent pixel with the initial pixel, if the similarity norm is met add a pixel to the region, mark the added pixel with a specific color for that region and change the region value. When the growth for that region stops, select a new pixel, and repeat the above steps.

Payer et al. [10] described how it would be desirable to find the coordinates of all cardiac substructures so that one could focus only on the localized region. The idea is to use two separate convolutional neural networks. The first grid finds the coordinates of the whole heart, while the second grid focuses on the positions from the data set with a focus on anatomically suitable positions and provides the desired segmentation. Wang et al [11] describe a method that uses the U-net architecture. This method can simultaneously learn to detect the area of interest of the heart and classify pixels into different substructures without losing resolution. U-net gives a rough prediction of the pixels, and after a successful prediction, the reconstruction into the original dimension takes place. The second architecture, which is based on the SRCNN (Super Resolution Convolutional Neural Network), takes the original data and results obtained from the first network as input, thus enabling direct prediction on images with the original resolution.

Wai et al [12] described aortic segmentation. Segmentation of the aortic lumen with MRI images is essential for accurate mechanical characterization of the aorta. One of the challenges for this task is the rarity of aortic annotations, since only a few frames are marked. To address the problem, a non-rigid image registration method was applied to propagate the labels from the annotated frames to the unlabeled neighboring ones in the cardiac cycle. Using this technique effectively generates pseudo annotated images that can then be used for training. This method of partially supervised learning achieves a dice score of 0.96 for the ascending aorta and 0.95 for the descending aorta. The test was performed on 100 samples. This approach is also based on FCN (Fully Convolutional Network) and RNN (Recurrent Neural Network) and with them segmentation can be performed directly without the need to assess areas of interest. Figure 2.2 shows the detection of the aorta. Figures 2.2 a) and 2.2 b) show the ascending aorta and Figures 2.2 c) and 2.2 d) the descending aorta.

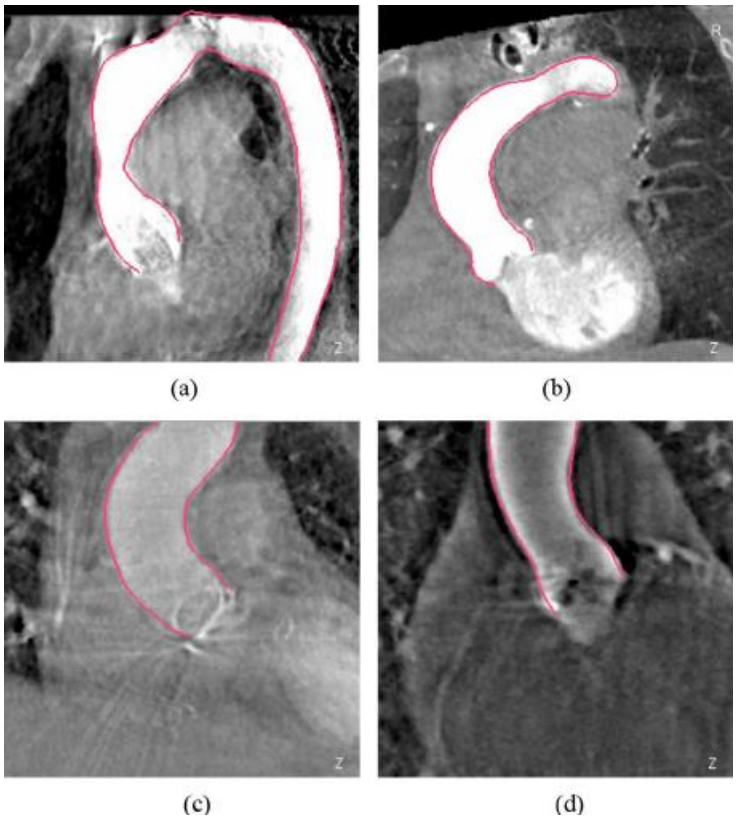


Figure 2.2 The segmentation the aorta [15]

Daoudi et al [16] use a region growth technique similar to Shewaje [9] with additional clarifications of the initial pixel selection method, as well as the use of additional image processing methods. The segmentation process starts by improving the contrast of the image using Adaptive Histogram Equalization (AHE). Using this method of changing contrast, the histogram can be adjusted to widen the edges of areas with poor distribution. A Generalized Hough Transform (GHT) is also used to select the initial pixel. The most important parameter of this method is the selection of the limit value. An iterative algorithm that will find the ideal limit value is implemented, and then the segmentation process is performed. Figure 2.3 shows the described process and the final result of the heart segmentation.

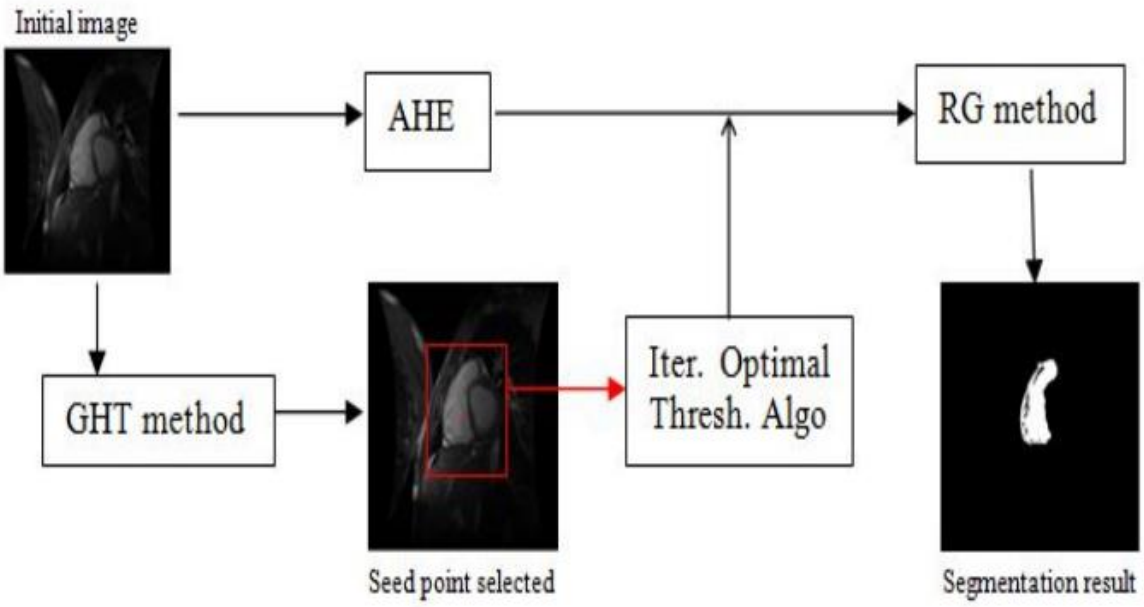


Figure 2.3 Presentation of the method from the paper[16]

Since the research of heart segmentation is an extremely frequently observed area, a number of other methods have been developed, detailed descriptions of which can be found in review papers [13,14].

3. MEDICAL AND TEHNOLOGICAL BACKGROUND

The area of image processing and deep learning is often needed to solve various tasks such as segmentation or classification. Deep learning methods are often applied in medical image processing where it is desirable to know the structure of organs or individual parts of the body. This chapter describes the structure of the human heart and further explains how magnetic resonance imaging works and in what format medical images are saved. It also describes convolutional neural networks and their use in the process of semantic segmentation of medical images.

3.1. The structure of the human heart

The heart is a muscular organ that acts like a pump to continuously send blood throughout the body. It is the center of the circulatory system. This system consists of a network of blood vessels, such as arteries, veins and capillaries. They carry blood to all parts of the human body. The electrical system regulates the heart and uses electrical signals to contract the heart. During a heart contraction, blood is pumped into the circulatory system. The system of inlet and outlet valves in the heart chambers work to ensure blood flow to the right (Figure 3.1). Without pumping the heart, blood cannot circulate within the body [17].

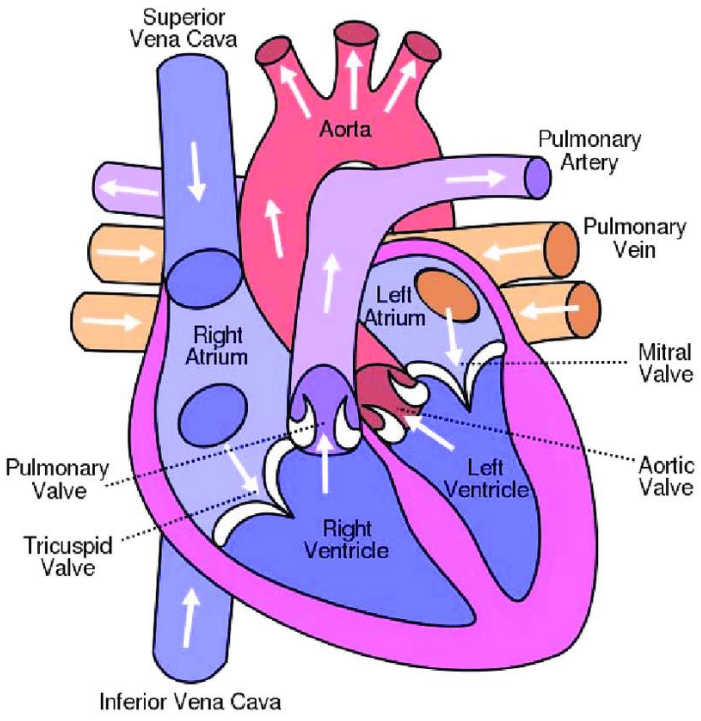


Figure 3.1 Structure of the heart [18]

The heart consists of the left and right halves and is divided into four chambers. The cardiac septum extends in length between the two halves of the heart as shown in Figure 3.2. The septum separates the four ventricles into the left and right atria and into the left and right ventricles. The left ventricle is part of the circulation and is connected to the left atrium. It is responsible for supplying the body through the circulation of the aorta with blood coming from the lungs. The right ventricle is part of the pulmonary circulation and is connected to the right atrium [19].

Figure 3.1 shows the blood flow through the heart. Blood is taken from the heart by the largest artery in the body -the aorta which has branches for the head and neck, heart, liver, intestines, kidneys and lower extremities. The veins collect blood rich in carbon dioxide and decomposed metabolic products from all parts of the body and carry it through the superior and inferior vena cava to the right heart where this blood goes to the lungs through the pulmonary artery. In the lungs, the blood is enriched with oxygen, taking it to the left half of the heart to other parts of the body [20].

Figure 3.2 A more detailed view of heart chambers [21]

3.2. Getting images using Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a medical imaging technique used in radiology to shape images of anatomy and physiological processes in the body. Magnetic resonance imaging creates three-dimensional anatomical images that are used to detect diseases, and to diagnose and monitor treatment. It is based on a technology that excitates and detects a change in the direction of the rotational axis of protons found in water that makes up living tissues [22]. MRI (Magnetic Resonance Imaging) does not involve X-rays or the use of ionizing radiation, which distinguishes it from CT (Computed Tomography) and PET (Positron Emission Tomography) scans. MRI is widely used in hospitals and clinics for medical diagnosis and for setting and monitoring the disease without exposing the body to radiation. MRI scans take a long time and are loud. As shown in Figure 3.3, most MR (Magnetic Resonance) devices have a tunnel-shaped housing in the central part of which a part of the body to be analyzed is placed using an automated examination table [23]. In addition, people with some medical implants or other metal that cannot be removed from the body may not be able to pass an MRI examination completely safely.

Figure 3.3 MRI device [24]

3.2.1. Background physics of Magnetic Resonance

Atoms consist of three basic particles: protons that have a positive charge, neutrons that have no charge and electrons that have a negative charge. Protons and neutrons are located in the nucleus of an atom, while electrons are located in the shells surrounding the nucleus. The characteristic chemical reactions of the elements depend on the specific number of each of these particles. The properties most commonly used to categorize elements are the atomic number and atomic weight. The third property of the nucleus is its spin. It can be considered that the nucleus rotates continuously around its axis at a constant speed, and this rotation around its axis is perpendicular to the direction of rotation (Figure 3.4 shows the rotation) [25]. MRI devices use strong magnets that produce a strong magnetic field that forces protons in the body to align with that field. Figure 3.5 that shows when the radio frequency pulsates through the patient, the protons are stimulated and thrown out of balance, leaning-on the magnetic field pull. When the radio frequency field is off, MRI sensors can detect the energy released when the protons align with the magnetic field. The time required for protons to align with the magnetic field, as well as the amount of energy released, varies depending on the environment and the chemical nature of the molecules. Physicians are able to distinguish between different tissue types based on these magnetic properties [22].

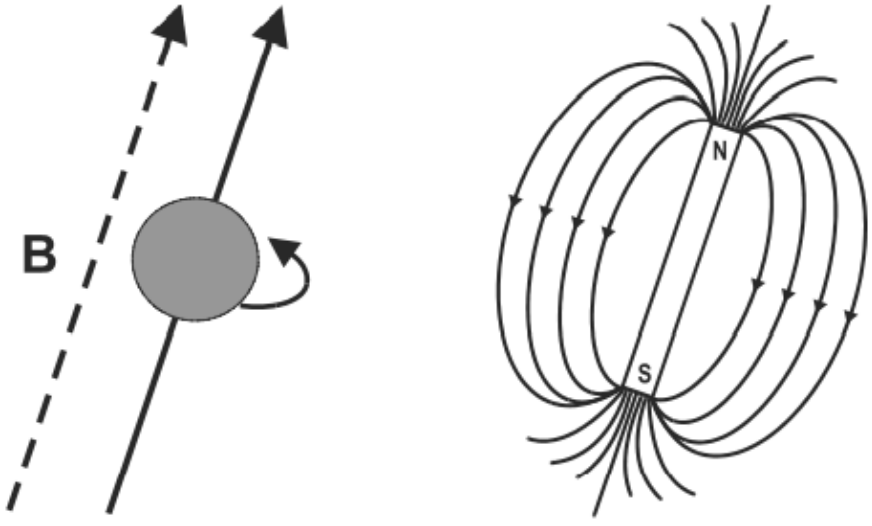


Figure 3.4 Spin of an atomic nucleus [25]

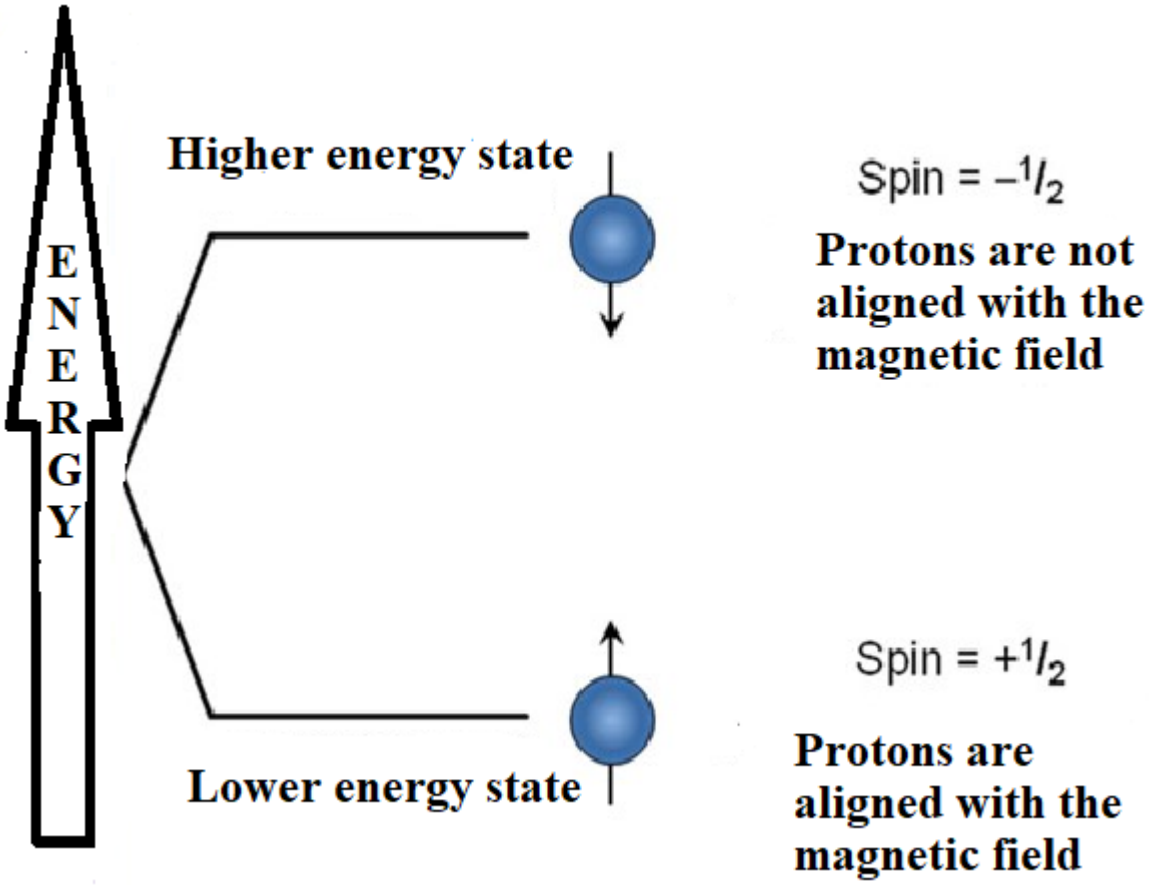


Figure 3.5 Proton alignment [26]

3.3. Medical image recording formats

The main formats used to record medical images are: Analyze, Nifti (Neuroimaging Informatics Technology Initiative), Minc (Medical Image NetCDF) and Dicom (Digital Imaging and Communications in Medicine). A medical image is a representation of the internal structure or function of an anatomical region in the form of a series of image elements called pixels or voxels. What the numerical value of a pixel expresses depends on the image modality, acquisition protocol, reconstruction, and finally, post-processing [27].

Pixel depth is the number of bits used to encode the data of each pixel. Each image is stored in a file and stored in computer memory as a group of bytes. Bytes are a group of 8 bits and represent the smallest amount that can be stored in computer memory [27].

Photometric interpretation determines how pixel data should be interpreted to correctly display an image as a monochrome or color image. Monochrome images have one pattern per pixel and no color data is stored on it. Grayscale is used to display images. The number of shades of gray depends on the number of bits used to store the pattern that matches the depth of the pixels. Clinical radiological images, such as magnetic resonance imaging and computed tomography, have a photometric interpretation of the gray scale. Colour is used to encode the direction of blood flow in Doppler ultrasound, to show additional information in the anatomical image [27].

Metadata is information that describes an image. Metadata is saved at the beginning of the file as a header and contains at least the image's matrix dimensions, spatial resolution, pixel depth, and photometric interpretation. Images that come from diagnostic modalities have information on how to create an image. For example, an MRI image will have parameters related to the pulse sequence used, time information, number of acquisitions, etc. Table 3.1 shows a description of each format used in recording medical images [27].

Table 3.1 Medical image formats[27]

Format	Heading	Extension	Data type
Analyze	Fixed length: 348 bytes of binary format	.img, .hdr	Integer, float, complex number
Nifti	Fixed lengths: 352 bytes of binary format (348 bytes in case of saving data as .img or .hdr)	.nii	Integer, float, complex number
Mine	Expandable binary format	.mne	Integer, float, complex number
Dicom	Binary format of variable length	.dem	Integer

Depending on the data type, numeric pixel values are stored as integers or floating point numbers using the minimum number of bytes required to represent the value.

Medical image formats can be divided into two categories. The first formats were intended to standardize images generated by diagnostic modalities such as Dicom. The second category is designed to facilitate post-processing analysis, such as Nifti, Minc, and Analyze. Medical image files are typically stored using one of the following two possible configurations. One in which one file contains both metadata and image data, with metadata stored at the beginning of the file. This paradigm is used by the file formats Dicom, Minc and Nifti. The second configuration stores metadata in one file and image data in another. The Analyze file format uses a two-file paradigm [27].

Analyze format used to be the standard for post-processing medical images. It is designed for multidimensional data. It is possible to store three-dimensional or four-dimensional data in one file (the fourth dimension represents time data). It consists of two binary files: image files with an ".img" extension, which contain voxel data, and a header with an ".hdr" extension, which contains metadata such as the number of pixels in the x, y, and z directions, the voxel size, and the data type. The header is a fixed length of 348 bytes and is described as a structure in the C programming language. This format is considered obsolete today, but is still used [28].

The Nifti format was created with the intention of taking advantage of the Analyze format, but also to solve its disadvantages. Although the format also allows header and pixel data to be stored in separate files, images are usually saved as a single ".nii" file, combining header and pixel data. The header is 348 bytes long if ".img" and ".hdr" files are saved, and 352 bytes if it is stored as a single ".nii" file. The Nifti format allows you to record in two ways. The first method involves rotation and translation, which is used to map voxel coordinates. The second method is used to store 12 linear transformation parameters that define the alignment of the image volume according to the coordinate system. The nifti format quickly replaced the Analyze format in medical image processing and is today the most commonly used format in the field [27].

Dicom is not only a file format, but also a network communication protocol. The Dicom standard describes how to format and share medical images and related data inside and outside the hospital. Dicom as a file format made it possible for pixel data not to be separated from the description of the medical procedure that led to the creation of the image itself, and this means that metadata is not separated from the image. Metadata and pixels are merged into a single file, and

the Dicom header, in addition to the image matrix data, contains the most complete description of the entire process used to generate the image. The header also contains patient information such as name, gender, age, weight, and height [27, 29].

The Minc system includes three components: a specification and file format, a program library, and a large number of tools. Minc has an expandable design and is ideal for use in larger institutions and databases. The header can have a lot of information, unique identifiers, and an explicit processing history. Newer versions of this format support 64-bit files [30].

3.4. Convolutional neural networks

Convolutional neural networks are similar to conventional neural networks. They consist of neurons that have weight and bias. Each neuron receives some inputs, derives a point product, and optionally follows it with nonlinearity. The whole network expresses one differential grading function and has a loss function. Convolutional neural networks take advantage of the fact that the input consists of an image, and the layers of the convolutional neural network are arranged in three dimensions: width, height, and depth [31]. Each layer in CNN (Convolutinal Neural Network) applies a different set of filters, usually hundreds or thousands, and combines the results. The last layer of CNN uses these higher-level features to predict image content. CNN gives us two key benefits: local invariance and composition. The concept of local invariance allows us to classify an image that contains a particular object regardless of where that object is located in the image. Each filter composes a local patch of lower-level features into a higher-level display, similar to how we can compose a set of mathematical functions that build on the result of previous functions: f (g (x (h (x ())). Composition allows the network to learn as much as possible from the given features [6].

There are many types of layers used to build convolutional neural networks, but the most commonly used are: the Convolutional layer (CONV), activation layer (ACT or RELU), pooling layer (POOL), fully connected layer (FC), batch normalization (BN), dropout layer (DO). Stacking a series of these layers in a specific way is obtained by CNN. CONV, RELU, POOL and FC are the most important in defining the network architecture [6].

3.4.1. Layers of convolutional neural networks

The CONV layer is the most important layer of the convolutional neural network. The CONV layer consists of a set of K filters, where each filter has a height and a width and are almost always square. These filters are small but extend through the entire depth of the volume. For volumes deeper in the network, the depth will equate number of filters applied in the previous layer. Each K filter can be imagined as “sliding” across the input area, calculating multiplication by elements, summation, and then storing the output values in a two-dimensional output called an activation map (Figure 3.6) [6]. After applying the K filter to the input volume, the K two-dimensional activation maps were created. Activation maps are arranged along the depth dimension of this sequence to obtain the final output volume (Figure 3.7) [6,31].

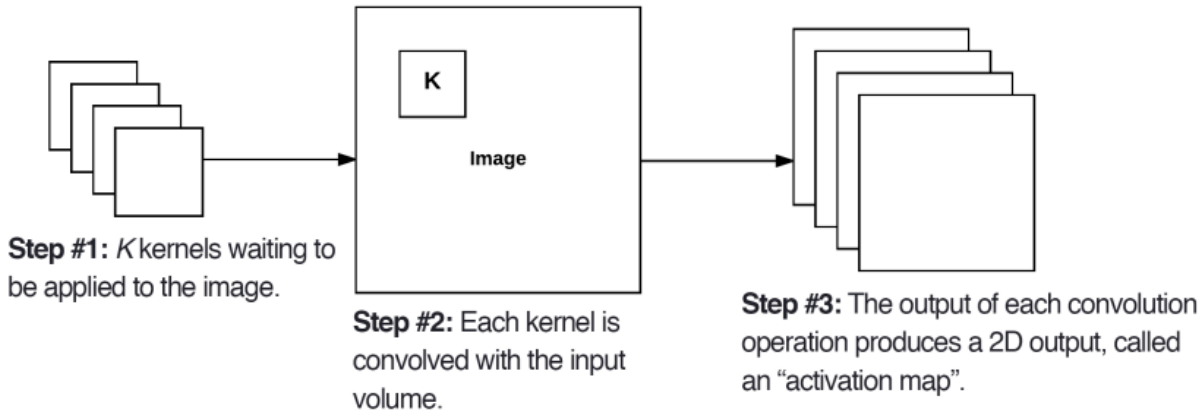


Figure 3.6 K filter implementation [6]

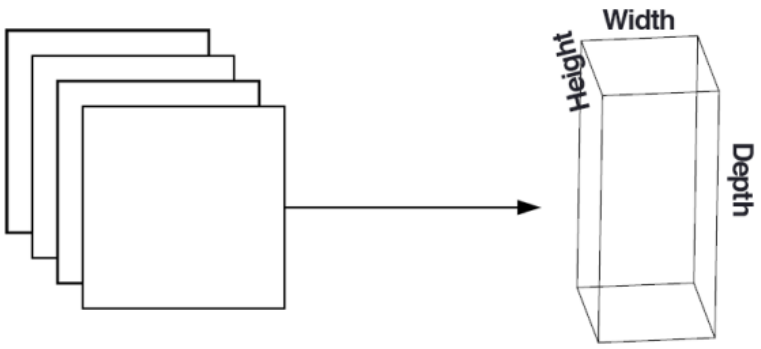


Figure 3.7 Stacking K maps into final output [6]

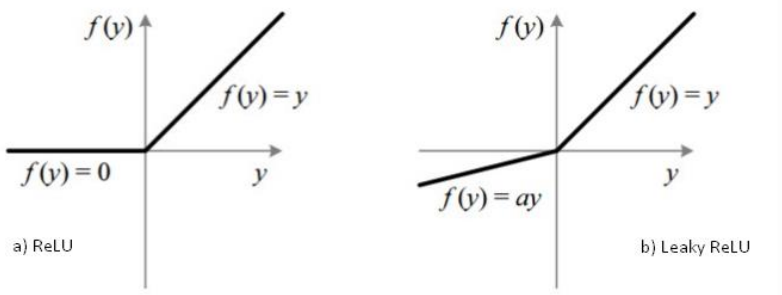
When working with images, it is often impractical to link images of neurons to the current volume with other previous volumes. There are simply too many connections and too much weight, which makes it impossible to train deep networks on large images. Instead, when CNN is used, each neuron connects only to the local volume input region. The size of such a region is called the receptive field of neurons. There are three parameters that control the size of the output volume: depth, step , and zero filling [6].

The depth of the output volume controls the number of neurons (filters) in the CONV layer that connect to the local area of the input volume. Each filter creates an activation map [6].

Figure 3.6 shows a convolutional matrix with a smaller matrix "sliding" through a large matrix. It stops at each coordinate, executes a calculation and saves the output. Thus, an arbitrary number of steps can be selected when convolution is applied to the input volume. The most common number of steps is one or two [6].

The method of filling the input with zeros along the border of the image gives the possibility of equalizing the dimensions of the input and output. This method becomes important when several CONV layers are stacked on top of each other [6].

After each CONV layer, a nonlinear activation function is applied: a Rectified Linear Activation Function (ReLU) and an Exponential Linear Unit (ELU) or some other variant of the two. Activation functions are used to determine the output from the previous layer. The ReLU activation function is most commonly used [32]. As shown in Figure 3.8 a), when y is less than zero, then f (y) is equal to zero, and when y is equal to zero or greater than zero, f (y) is equal to y. The problem with this ReLU function is that all negative values immediately become zero, which reduces the model's ability to adjust properly to the data set. For this reason, the so-called Leaky ReLU (Figure 3.8 b)) is increasingly used, which allows the use of negative values. The activation layer receives an input volume of dimensions W (width) x H (height) x D (depth) and applies the activation function. The output from the activation layer has the same dimensions as the input [6, 31].



Slika 3.8 ReLU and Leaky ReLU [31]

The Pooling layer (POOL) is often introduced between convolutional layers. The primary function of the POOL layer is to gradually reduce the dimensions of the input volume. This reduces the number of parameters and computations in the network, it also helps to control overfitting of the network. POOL layers use the maximum or average function. Maximum compression is typically performed in the middle of a CNN architecture to reduce space size, while average compression is used as the final layer of the network. Maximum compression is most commonly used, and the compression size is 2 x 2 in most cases, but for larger images, 3 x 3 is also used [6]. Figure 3.9 shows the use of maximum compression. The upper matrix has a step of size 1, and a step lower of size 2.

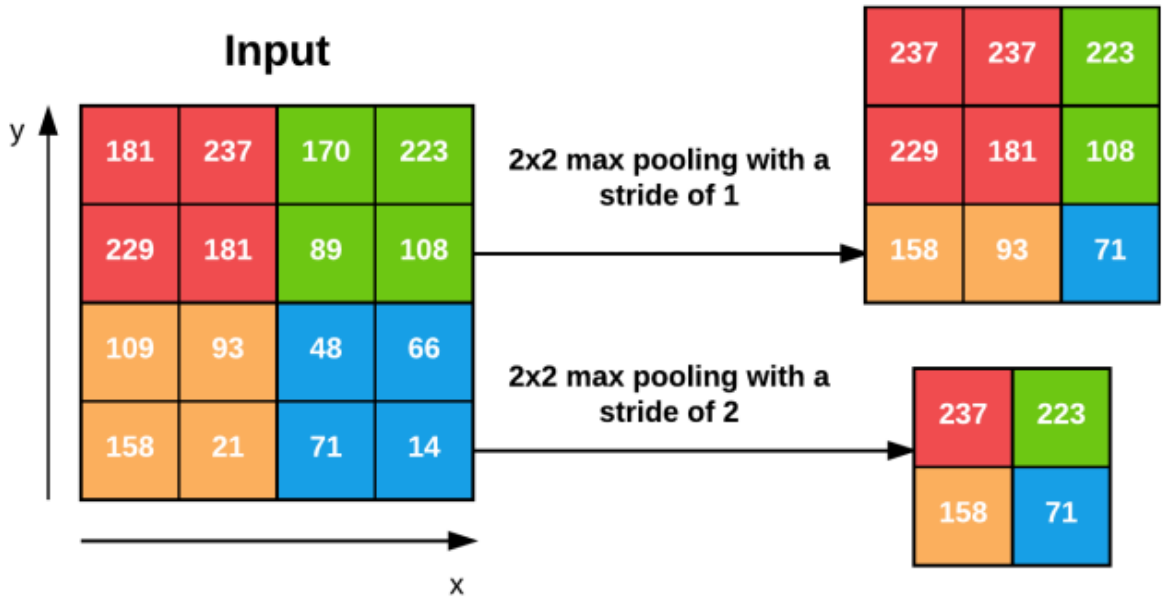


Figure 3.9 Output result after the POOL layer [6]

Fully Connected (FC) neurons are associated with all activations in the previous layer. FC layers are always placed at the end of the network. It is common to use one or two FC layers before applying the classifier [6].

Batch Normalization (BN) is used to normalize the activation of the input volume before it is passed to the next layer in the network. BN has been shown to be extremely effective in reducing the number of epochs required to train the neural network. Batch normalization also has the added benefit of helping to “stabilize” training. It allows the use of a larger range of learning rates [6].

The Dropout (DO) layer will randomly disconnect the input connections from the previous to the next layer in the network for each mini batch in the training set. Figure 3.10 shows what the network looks like after using the DO layer when a random break is associated with a 50 percent probability. The DO layer is used to prevent network overfitting [6].

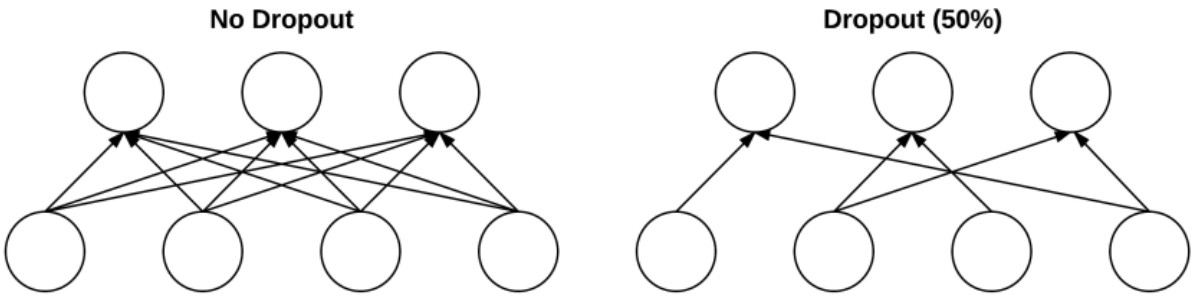


Figure 3.10 DO layer example [6]

Figure 3.11 shows what the architecture of a convolutional neural network would look like after connecting the most important layers.

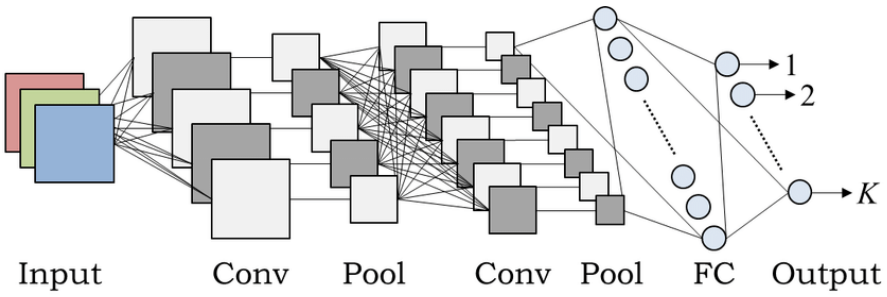


Figure 3.11 CNN architecture [33]

3.5. Semantic segmentation of medical images using convolutional neural networks

Semantic segmentation is a fairly new biomedical image processing technology, but it has already made a major contribution to sustainable medical care. With the rapid development of deep learning, medical image processing based on convolutional neural networks has become the focus of today's research. Despite the great achievements of segmentation in the field of medical image processing based on deep learning, there are still difficulties in research. For example, the segmentation accuracy is not too high, the number of medical images in the data set is small, and the resolution is low. Inaccurate segmentation results are unable to meet actual clinical requirements. Therefore, attempts are being made to improve current segmentation solutions using convolutional neural networks [34].

When performing image segmentation operations, convolutional neural networks have excellent feature extraction and decent feature expression capabilities. They do not require manual extraction of image features or excessive image processing. Therefore, CNN is used in medical image segmentation. Currently popular algorithms for semantic segmentation of medical images other than U-net are: SegNet, PSPNet, DeepLab [34,35].

For the general classification of CNN networks, such as VGG and ResNet, fully connected layers are added to the end of the network. Category probability information can be obtained after the softmax layer, but this probability information is one-dimensional. That is, only the category of the entire image can be identified, not the category of each pixel. Thus, this fully related method is not suitable for image segmentation. FCN is used to solve this problem. In the usual CNN structure, the first five layers are convolutional layers. The sixth and seventh layers are completely connected layers of length 4096 (one-dimensional vector). The eighth layer is a fully connected layer of length 1000, which corresponds to a probability of 1000 categories. The FCN changes the last three layers to convolution layers whose convolution filters are 7 x 7, 1 x 1, and 1 x 1. This will give a two-dimensional feature map of each pixel. This is followed by a softmax layer to obtain the classification information of each pixel. In this way, the problem of segmentation is solved. A fully convolutional network can accept input images of any size. The FCN uses a deconvolution layer to return the image to the same size as it was at the input. Thus, a prediction can be generated for each pixel, retaining the original dimension of the image. Finally, a pixel classification is performed on the sampled feature map to complete the final image segmentation [34]. Figure 3.12 shows the structure of FCN.

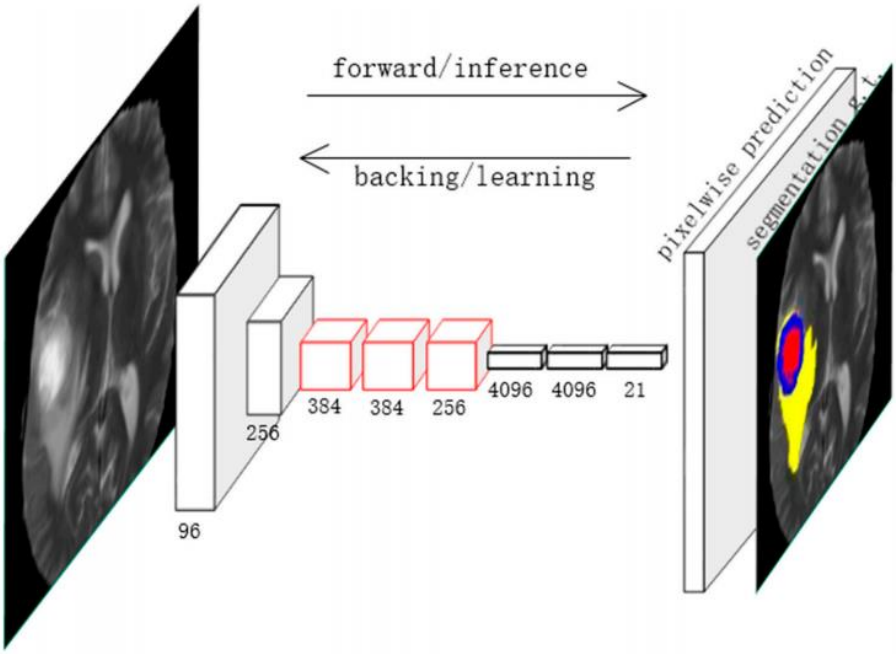


Figure 3.12 FCN [32]

DeepLab v1 is based on the VGG-16 network, removing the last fully connected layer of the VGG network and using full convolution instead, as using too many compression layers will result in an insufficient feature layer size. The features contained are too rare, which is not suitable for semantic segmentation. DeepLab v1 uses atrous convolution. Compared to traditional convolution, the receptive field can expand without increasing the amount of calculation and the density of features can increase. Finally, DeepLab v1 uses a conditional random field to improve the accuracy of segmentation boundaries. DeepLab v2 is an improvement based on DeepLab v1. DeepLab v2 solved the segmentation problem caused by the difference of the same scale of the object in the same image. DeepLab v3 uses the ResNet-101 architecture. Its goal is to solve the problem of multiscale segmentation and therefore a cascade or parallel atrous convolutional module has been designed. DeepLab v3 + is an extension of DeepLab v3. A simple decoder module has been added to refine the segmentation results, especially for segmentations at the edges of the object [34].

SegNet has a symmetric encoder-decoder structure based on the semantic segmentation of the FCN to achieve image segmentation at the pixel level. The network mainly consists of an encoder and a decoder. The encoder is a network model that uses the VGG-16, mostly to analyze object data. The decoder converts the parsed information into the final image format, that is each pixel is represented by a color or label that corresponds to the object information. The decoder uses a large compression index that is transmitted to the encoder to take nonlinear samples of its inputs, so no

learning is required to take samples. Then a training convolutional filter is used and it creates a dense feature folder. When the feature map is returned to its original form, they are sent to the softmax classifier for pixel classification [34].

PSPNet is a semantic segmentation model that uses ResNet-101 as a feature extraction layer and introduces a pyramidal compression module to identify previous image information [35]. Figure 3.13 shows the architecture of PSPNet. The architecture uses a pyramid of 4 levels, each level uses convolution and they cover the whole, half and smaller parts of the image. These convolutions are linked and will eventually merge with the initial feature map.

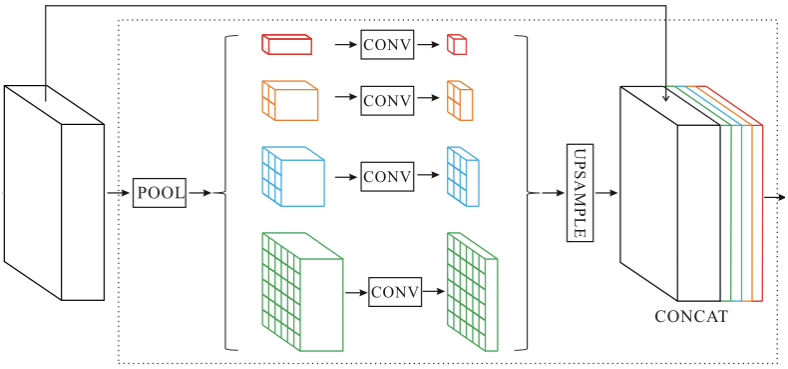


Figure 3.13 PSPNet architecture [36]

4. DEVELOPED SEGMENTATION SYSTEM

As described in the previous chapter, convolutional neural networks are used to segment medical images. The U-Net convolutional neural network has shown remarkable results in this area and is the most commonly used network for this task. Pre-processing the data is a common step before training the network. This can include a variety of image processing methods, data filtering, data categorization, and more. This chapter describes what pre-processing method was performed on the data and describes the U-net convolutional neural network.

4.1. Data pre-processing

The dataset used in this thesis presents magnetic resonance images of the heart chambers. The dataset contains 20 images of heart chambers in Nifti format and the same number of labels that are also in Nifti format. The labels are masks for each image, and they are in fact annotated images that indicate the exact position of the heart chambers from the MRI images. Figure 4.1 shows the images using ITK-SNAP tools [37, 38]. ITK-SNAP allows work with magnetic resonance images. Figure 4.1 shows the exact position of the heart using a label.

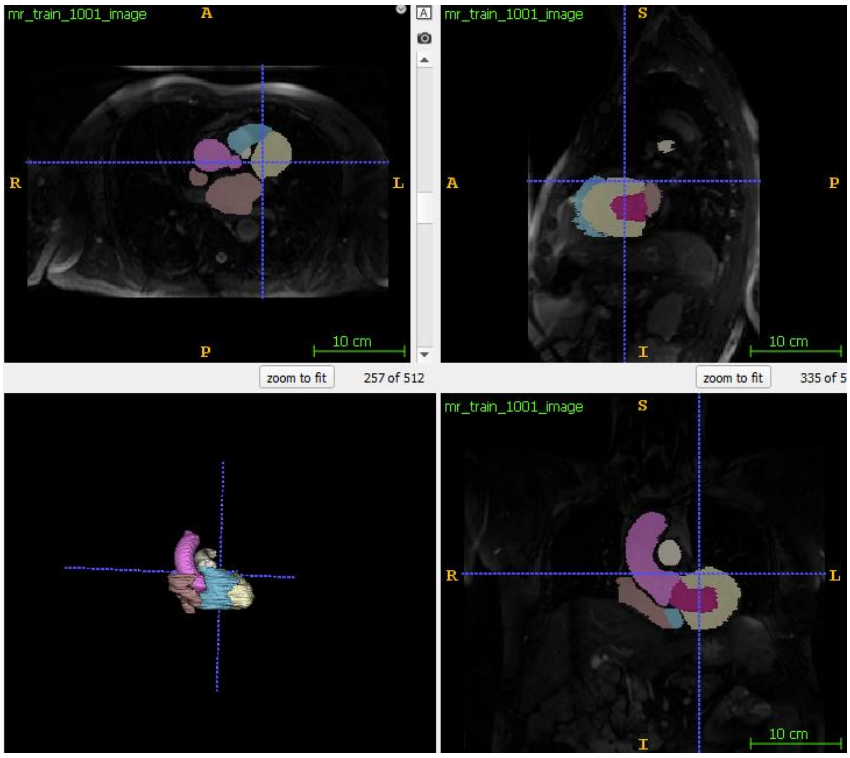


Figure 4.1 Magnetic resonance imaging using ITK-SNAP [37, 38]

To enable 2D CNN training, images from the Nifti format are converted to more familiar formats with the ".png" or ".jpeg" extension(Figure 4.1). When converting an image to another format, the images will go through each of these spatial dimensions and save all the images that appear to be the x, y, and z dimensions, with the fourth dimension representing time. The same procedure was done on the labels. Now the images are saved in a format that is more suitable for training.

Of course, a lot of images are saved and there are images that have no annotation, that is there are magnetic resonance images from the x, y and z dimensions where the heart is not visible at all. Such images are then ejected and are not required when training the network.

The images are then divided into a validation set and a training set. 75 percent of the images belong to the training set while 25 percent of the images are set aside for network validation. Images in the x and y dimension are trained together, separately from z dimension. The reason it is trained separately is because the difference in image dimension between x, y, and z dimensions is too significant and it is not possible to determine a satisfactory resolution corresponding to each spatial dimension, so when comparing the results, there will be a network which is trained with images from the x and y spatial dimensions and a network trained with images only from the z spatial dimension. Figure 4.2 shows what MRI images from each spatial dimension look like along with their label.

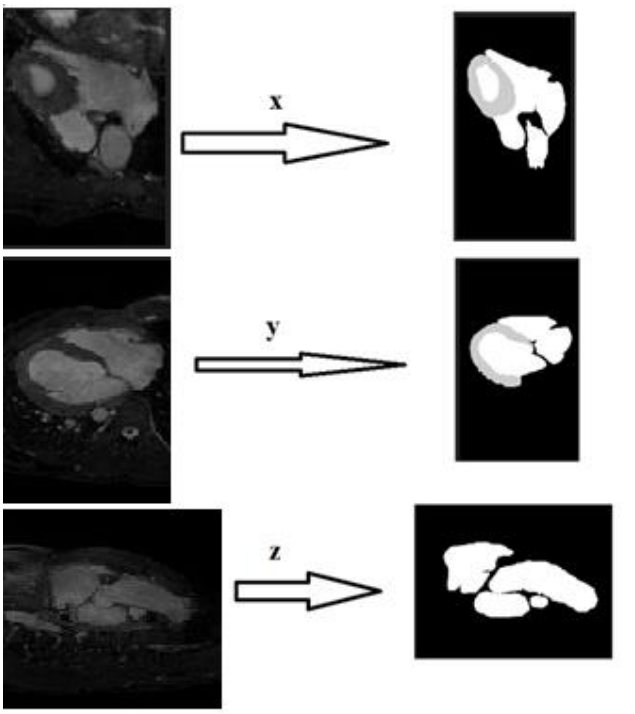


Figure 4.2 MRI images with corresponding labels

Before training, the images are set to a certain resolution because all of them must be the same size during training, and they are converted into monochrome images. The pre-processing method used in this thesis before image training is the Bayesian method for noise reduction.

4.1.1. Bayesian method for noise reduction

Bayesian method for noise reduction is based on the Bayesian classifier. The naive Bayesian classifier is a probabilistic model used to classify data. The Bayesian classifier is based on the Bayesian theorem given by formula 4-1:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{4-1}$$

Formula 4-1 represents the Bayes' theorem. The theorem shows how the probability of event A is found with respect to event B. Here B is the proof and A is the hypothesis [39].

Noise is a common problem when magnetic resonance imaging is being recorded. Fahmy [40] described a probability model for blood and tissue signals using the Bayesian classifier. The classifier is used to recognize and filter the background signal and thus the background noise can be removed from the image.

$$d(\bar{v}) = \log\left(f_{s_1 s_2}(S_1, S_2 |\, tkivo)\right) - \log\left(f_{s_1 s_2}(S_1, S_2 |\, pozadina)\right) \quad (4\text{-}2)$$

Formula 4-2 [40] shows the probability model used based on the Bayesian classifier. S_1 and S_2 represent the obtained images in the form of vectors, and they can be classified into two classes: background or tissue. The probabilistic model works in such a way that for each pixel it will calculate the probability of the event whether that pixel is the background or tissue and thus the noise in the image will be reduced because the noise will be removed from the image by classifying it as background [40]. By further announcing formula 4-2, formula 4-3 is obtained [40]:

$$d(\bar{v}) = \log\left(I_o\left(\frac{\hat{p}(x,y)sinc(\partial\hat{\omega})S_1}{\sigma^2}\right)\right) + \log\left(I_o\left(\frac{\hat{p}(x,y)sinc(1-\partial\hat{\omega})S_2}{\sigma^2}\right)\right) - \hat{p}(x,y)\frac{sinc(\partial\hat{\omega})^2 + sinc(1-\partial\hat{\omega})^2}{2\sigma^2} \quad (4\text{-}3)$$

Equation 4-3 will give the final probability amount for each pixel. If the value of $d(\bar{v})$ is greater than zero the pixel is classified as tissue, if it is less than zero it is classified as background. I_o represents the zero value of the modified first-order Bessel function, $\hat{p}(x,y)$ is the currently observed pixel, $\partial\hat{\omega}$ represents the deformation on the tissue, σ represents the standard noise deviation [40].

4.2. U-Net neural network architecture

U-net was originally invented and first used for biomedical image segmentation. The architecture of the U-net network is shown in Figure 4.3. At first glance, it has the shape of the letter "U". The architecture is symmetrical and consists of two main parts: the left part, called the contraction path, which conducts the convolution process, and the right part (expansion path), which consists of transposed two-dimensional convolution layers [41].

The contraction path constantly repeats the double convolution of dimension 3 x 3 is constantly increasing the depth of the image. A maximum compression layer of 2 x 2 is then applied and it will halve the image dimensions. This process is repeated 3 times until it reaches the bottom where 2 convolutional layers are carried out, but without the compression layer [42].

The expansive path will return the image back to its original dimension. This is achieved by transposed convolution. Transposed convolution is a pattern enlargement technique that expands the size of images. The transposed convolutional layers have dimensions of 2 x 2. After the transposed convolution, the image is merged with the corresponding image from the contraction

path and this is done to achieve more precise segmentation. This process is also repeated 3 times [42].

At the very top, a convolution layer with a filter size of 1 x 1 is applied and the image dimension is changed to meet the prediction requirements.

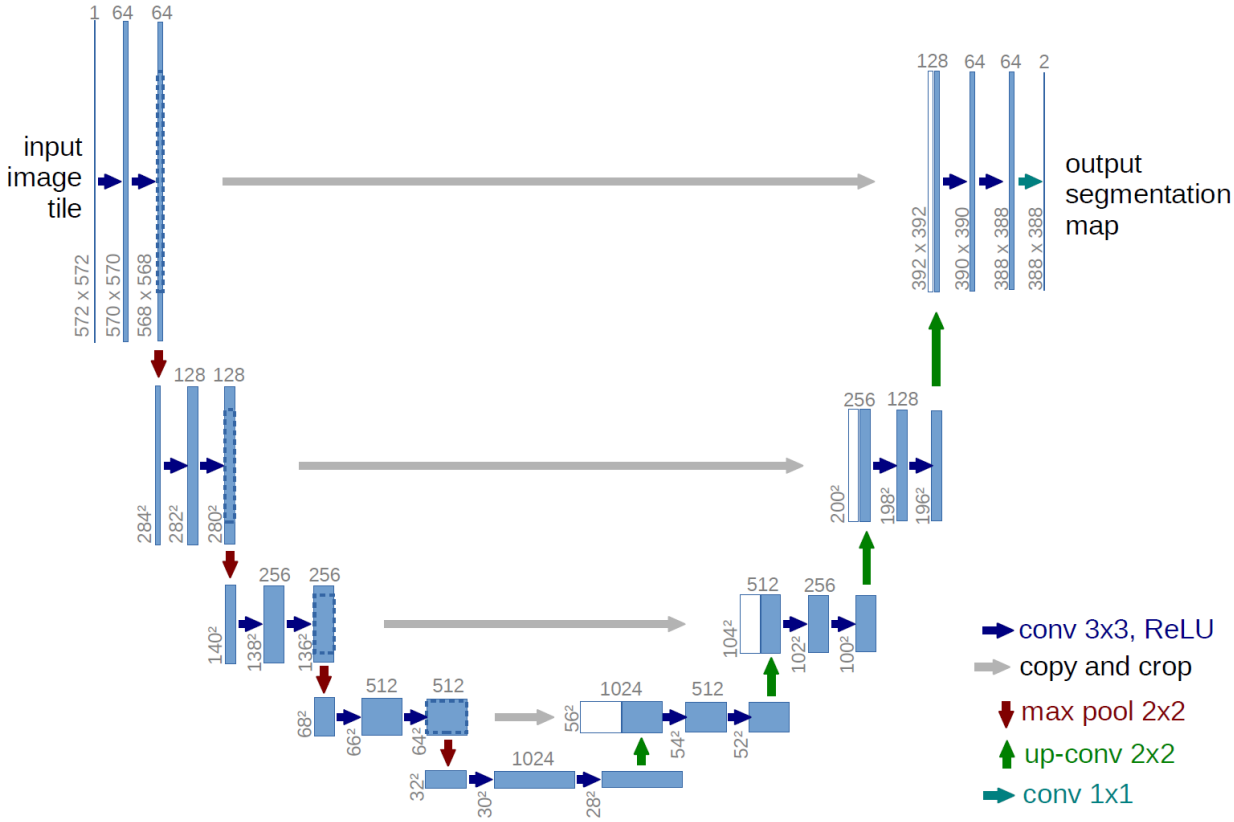


Figure 4.3 U-net architecture [41]

5. RESULTS

During network training, certain measures are used to conclude whether a model has good or bad performance. The aim of this chapter is to present the results of data pre-processing, to compare the results of segmentation with and without data pre-processing, and to evaluate the results of segmentation.

5.1. Data pre-processing results

The previous chapter described Bayes' method for noise reduction. This method is used to pre-process the data before training the convolutional neural network. Figure 5.1 and Figure 5.2 show the results after pre-processing from each spatial dimension. The heart is more pronounced in Figures 5.1 and 5.2 after pre-processing and the noise decreased. The noise is not completely removed because with too much pre-processing the heart shape would also be removed.

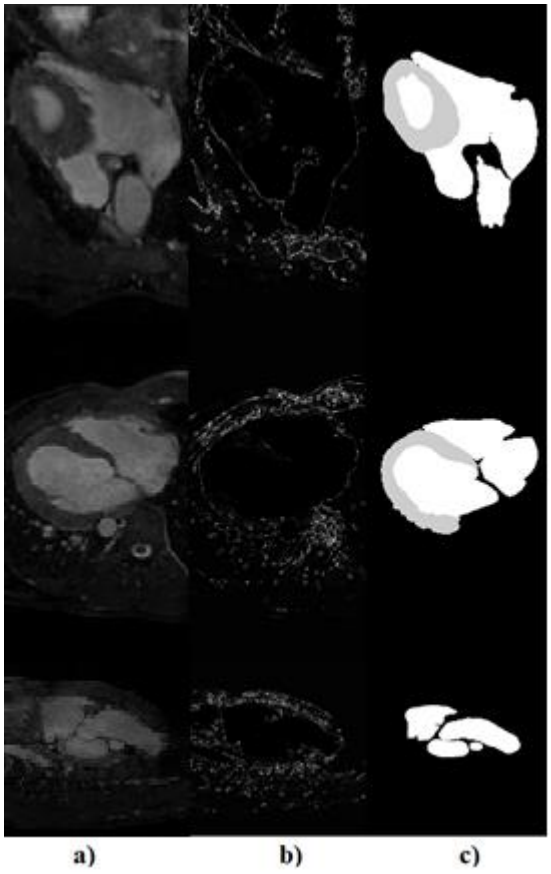


Figure 5.1 a) MRI image b) Pre-processing result c) Label

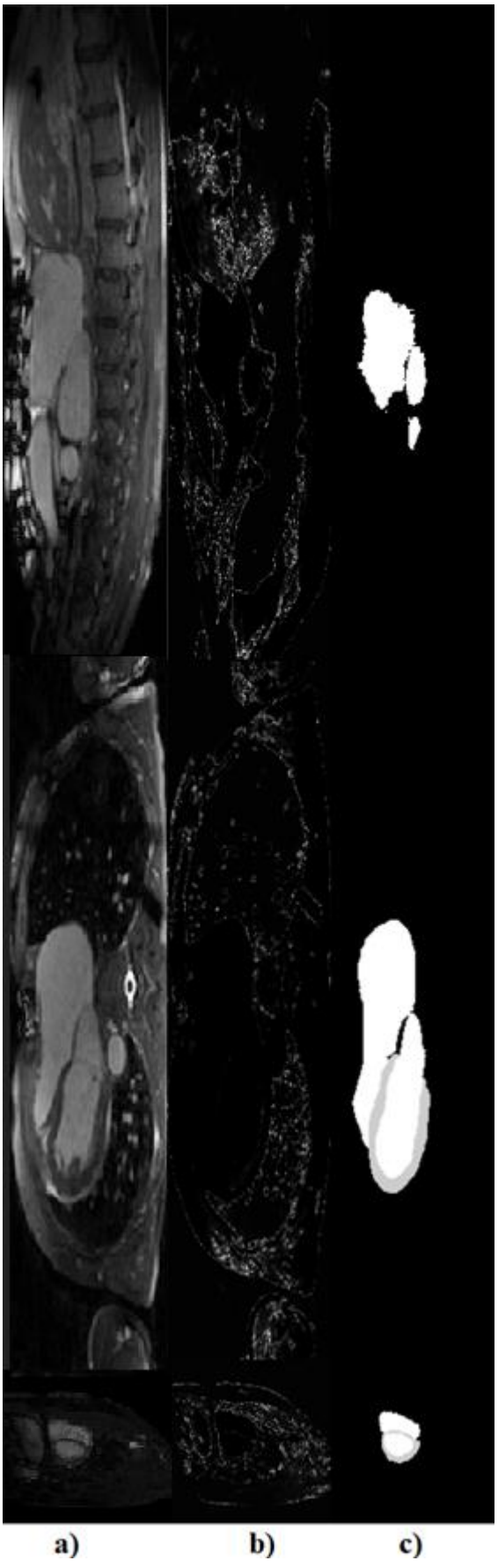


Figure 5.2 a) MRI image b) Pre-processing result c) Label

5.2. Comparison of segmentation results with and without data pre-processing

U-net convolutional neural networks were also trained on data without using the Bayesian noise removal method. Figure 5.3 and Figure 5.4 show the segmentation results with and without pre-processing.

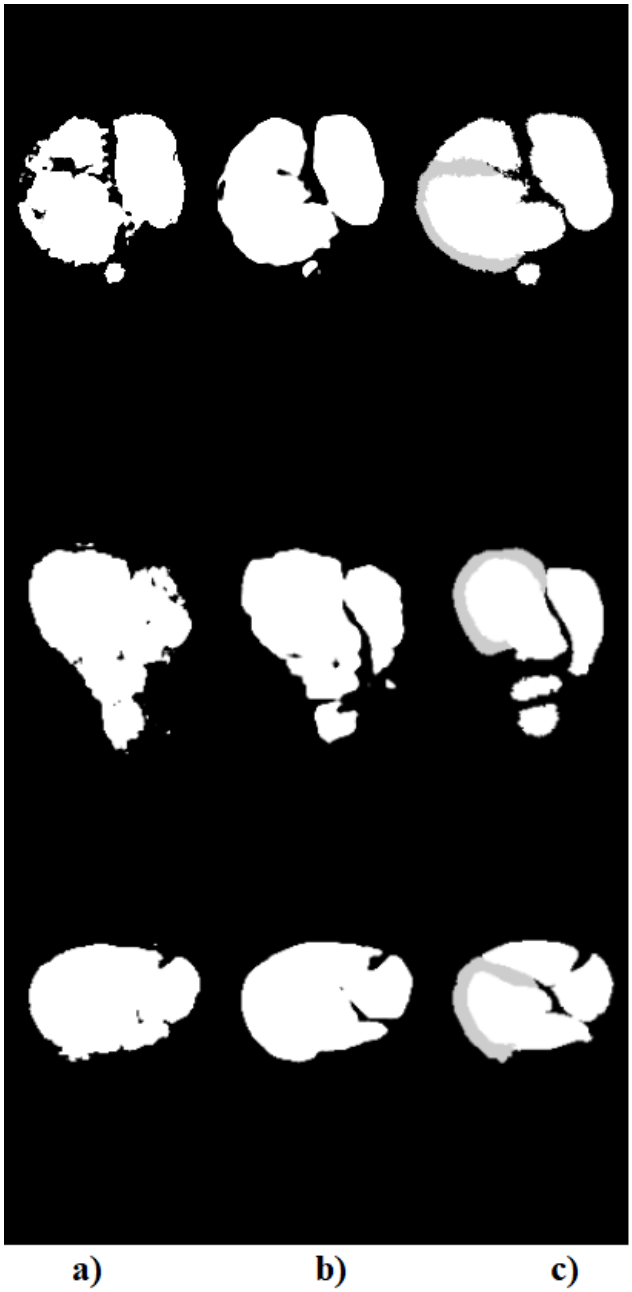


Figure 5.2 a) Segmentation without pre-processing b) Segmentation with pre-processing c) Labels

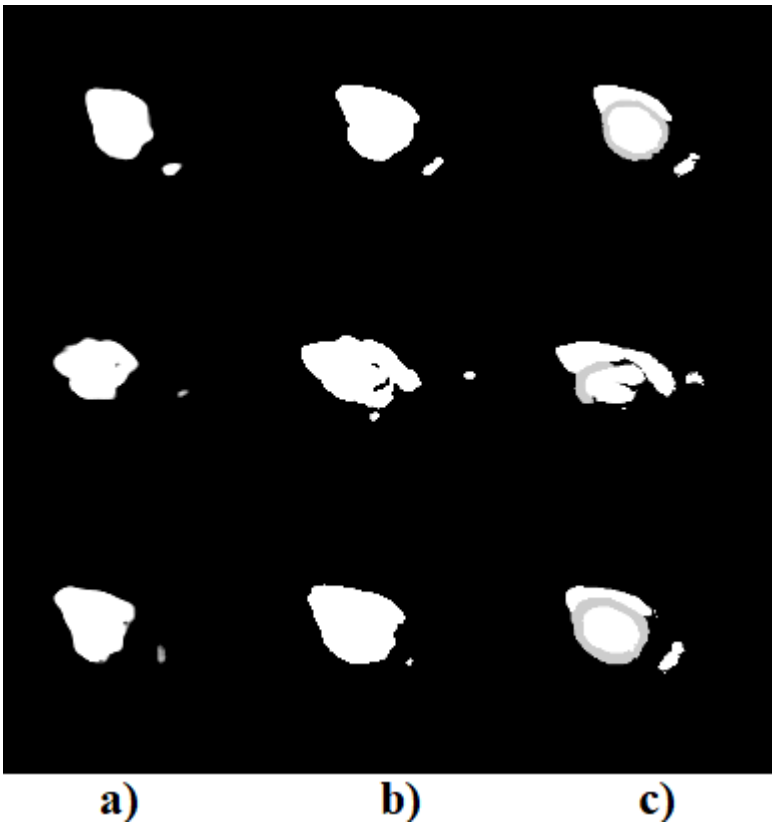


Figure 5.3 a) Segmentation without pre-processing b) Segmentation with pre-processing c) Labels

It can be seen from the figures that the data pre-processing had an impact on the final segmentation result. Images obtained by pre-processing have a shape similar to the original mask. Although the results obtained without pre-processing do not give bad results, it is evident that the pixel locating is less accurate.

5.3. Evaluation of segmentation results

During and after network training, various measures are used to evaluate the results. For models that use semantic segmentation, the results are evaluated by measuring: IoU (Intersection over Union), pixel precision or dice coefficient [43].

Pixel accuracy is the percentage of accurately classified pixels. This measure can be misleading when the representation of classes in the picture is small, because the measure will be biased mainly on how well it recognizes cases where the class is not present [43].

IoU, also known as the Jaccard index, is one of the most commonly used metrics in semantic segmentation. IoU metrics are very direct and extremely effective. IoU represents the ratio of the

area of overlap between the predicted segmentation and the reference data with the area of the union between the predicted segmentation and the reference data (Figure 5.5). Values range from zero to one (0% - 100%), with zero meaning no overlap and 1 indicating perfectly overlapping segmentation [43].

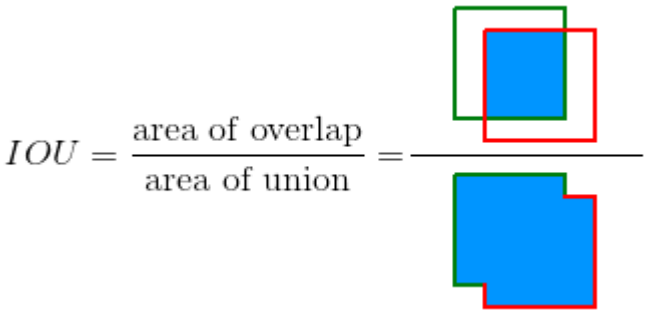


Figure 5.4 IoU calculation [44]

The dice coefficient is 2 x the overlap area divided by the total number of pixels in both images (Figure 5.6) [43].

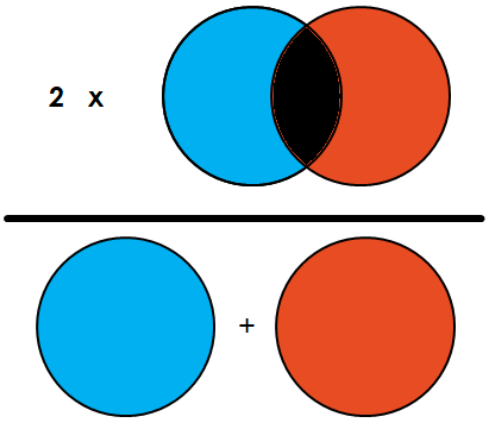


Figure 5.5 Dice coefficient calculation [43]

Another important measure in network training is the loss function. The loss function quantifies how well the reference values match the predicted class value. The higher the level of agreement, the lower the loss, and this is an indication that the accuracy of the classification is higher [6].

The dice coefficient and the loss function were used to evaluate the results. The evaluation was done on 4 models. The networks were trained on images from the x, y spatial dimension with and without pre-processing and on images from the z spatial dimension with and without data pre-processing. Figure 5.7 shows graphs of the loss function. Figure 5.7 a) shows the loss functions before data pre-processing, the upper graph is from x, y spatial dimensions, the lower graph is from the z spatial dimension. The graphs in Figure 5.7 b) used data pre-processing. It can be seen from the figure that there is not too much difference in the loss function between the models obtained from images from x, y spatial dimensions, but it can be seen that the loss gradually decreases, which is an indication that the network should have high classification accuracy. As for the z spatial dimension, images with pre-processing have loss lower than images without pre-processing. The learning rate during training is 0.0001. It controls how quickly the model adapts to the data set.

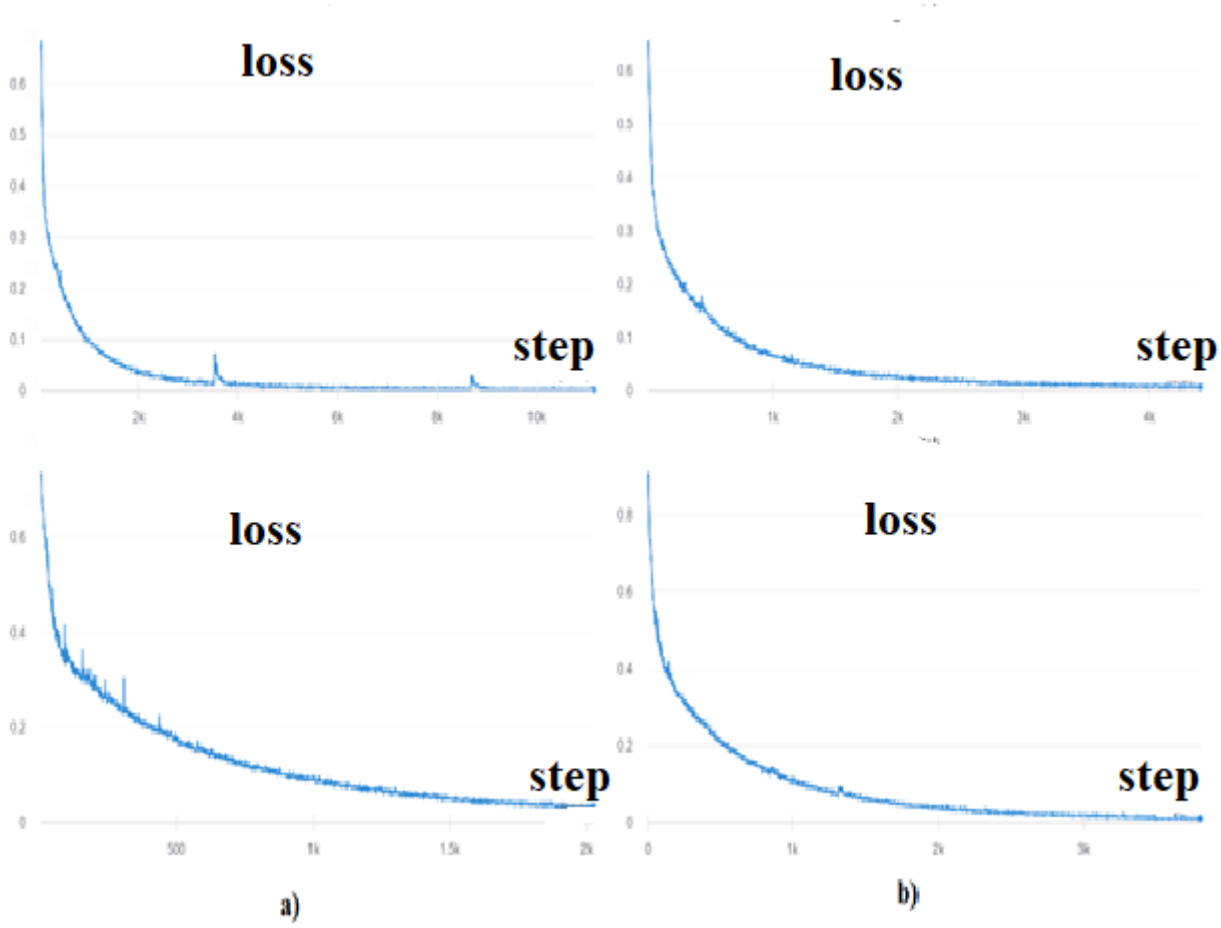


Figure 5.6 a) Loss function without pre-processing b) Loss function with pre-processing

Loss functions show satisfactory results for cases with and without data pre-processing. The biggest difference is visible in the value of the dice coefficient. The dice coefficient for images from x, y spatial dimensions without pre-processing is 87.14 percent, while for data with pre-processing, the dice coefficient is 88.89 percent. Images from the spatial dimension without pre-processing have a dice coefficient of 86.59 percent, and with pre-processing it is 88.74 percent.

Although seemingly this difference in percentages does not seem too great, a higher dice coefficient in such differences suggests that segmentation is more accurate. These results are also visible in the figures of the previous chapter (Figure 5.3 and Figure 5.4). By removing unnecessary noise and background, images are created where the heart shape is more visible and this affected the final results of the dice coefficient.

Globally, data pre-processing helped and segmentation is mostly more accurate, but there are also cases where pre-processing did not yield effective results. An example is given in Figure 5.8, where it is visible that due to pre-processing, the model did not successfully segment the image. The standard noise deviation σ affects the pre-processing intensity of the data. Since the same standard deviation value is applied to all images in the data set, there will be cases where this value is too large and therefore the heart shape in the image would be lost.

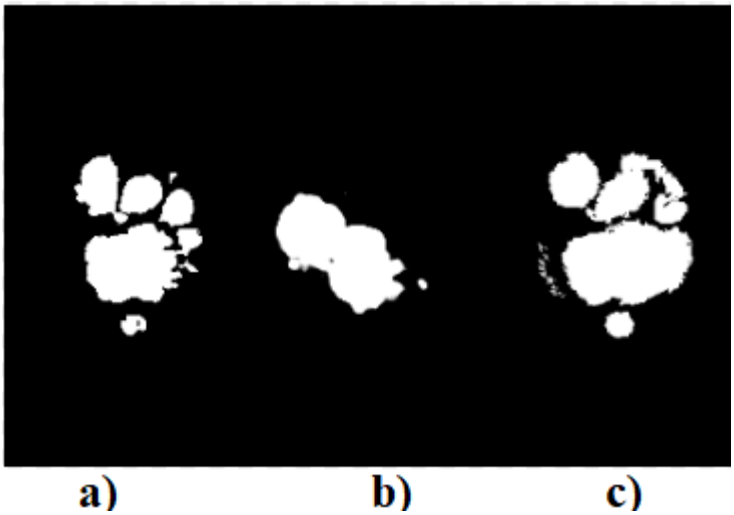


Figure 5.7 a) Segmentation without pre-processing b) Segmentation with pre-processing c) Labels

Bayes' method for noise reduction has proven to be a successful image processing method. There is a possibility that due to too strong or too small standard deviation, the noise reduction will not always successfully give satisfactory results. The results showed that with the Bayesian

method for noise reduction, a higher dice coefficient is obtained and the loss function is smaller. It is possible that the method could be improved by finding the ideal value of the standard deviation, and then the network model would give more precise segmentations. The results would also be affected by a larger set of data from more annotated images. Further improvement of the results would also be obtained by using variations of the U-net convolutional neural network.

6. CONCLUSION

Deep learning methods are increasingly present in various fields of science. Thus, segmentation is one of these methods that has begun to develop, including the field of medical image processing. Medical image processing is advancing rapidly and applies specific convolutional neural networks over magnetic resonance images and computed tomography images. PSPNet, SegNet, and DeppLab are used to segment MRI images, but remarkable results have been achieved by the U-net convolutional neural network. In addition to the use of convolutional neural networks, various image processing methods are also used. Through this thesis, the Bayesian method for noise removal was used and it is evident from the obtained results that it separates the most important parts of the image from the background. The Probabilistic Model is designed to distinguish tissue from a background signal. This pre-processing of the data yielded satisfactory results, but it should be taken into account that due to the uniform standard deviation of the noise the results are not always satisfactory, and the segmentation sometimes does exactly not find every pixel. Finally, with Bayes' method for noise reduction, images in which the U-net convolutional neural network was trained were obtained, thus achieving low values of the loss function and a high dice coefficient, which shows that the segmentation accuracy is high.

ACKNOWLEDGMENTS

This work was co-financed by the Croatian Science Foundation project UIP-2017-05-4968.

REFERENCES

[1] How to do Semantic Segmantaion using Deep learning, How to do Semantic Segmentation using Deep learning (nanonets.com), 2021

[2] Image Segmentation, segmentation.dvi (toronto.edu), 2021

[3] Research in Medical Imaging Using Image Processing Techniques | IntechOpen, 2021.

[4] Medical Image Processing, What is Medical Image Processing | Synopsys, 2021

[5] MRI scan, MRI scan - NHS (www.nhs.uk), 2021

[6] Deep Learning for Computer Vision with Python, Dr. Adrian Rosebrock, 2017, PyImageSearch.com

[7] U-net, U-Net: Image Segmentation Network (neurohive.io), 2021

[8] A fully automatic cardiac segmentation method using region growing technique, Jörg Sander, Bob D. deVos, Ivana Išgum

[9] Cardiac MR Image Segmentation Techniques: an overview, Tizita Nesibu Shewaye

[10] Payer, C., D. Štern, H. Bischof, and M. Urschler. Multilabel whole heart segmentation using cnns and anatomical label configurations. In: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges, edited by M. Pop, M. Sermesant, P. M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard. Cham: Springer, 2018, pp. 190–198

[11] Wang, C., and O. Smedby. Automatic whole heart segmentation using deep learning and shape context. In: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges, edited by M. Pop, M. Sermesant, P. M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard. Cham: Springer, 2018, pp. 242–249

[12] Bai W, Suzuki H, Qin C, Tarroni G, Oktay O, Matthews PM, et al. Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. 21st International Conference on Medical Image Computing and Computer Assisted Intervention—MICCAI, 2018. Granada: Springer International Publishing (2018). p. 586–94.

[13] Deep Learning for Cardiac Image Segmentation: A Review, Chen Chen, Chen Qin, Huaqi Qui, Giacomo Tarroni, Jinming Duan, Wenjia Bai, Daniel Rueckert

[14] Overview of the Whole Heart and Heart Chamber Segmentation Methods, Marija Habijan, Danilo Babin, Irena Galić, Hrvoje Leventić, Krešimir Romić, Lazar Velicki, Aleksandra Pižurica

[15] Segmentation of the aorta, Automatic aorta segmentation on a few example volumes. (a) Good... | Download Scientific Diagram (researchgate.net), 2021

[16] A fully automatic cardiac segmentation method using region growing technique, Abdelaziz Daoudi, Said Mahmoudi

[17] Canterbury District Health Board, Cardiology, Self learning package. Module 1: Anatomy and Physiology of the Heart

[18] Blood flow through the heart, Diagram of the human heart (cropped) - Heart - Wikipedia, 2021.

[19] Heart chambers , KOMORA SRCA - STRUKTURA, FUNKCIJA I BOLESTI - ANATOMIJA (healthandmedicineinfo.com), 2021

[20] Circulatory system of the heart, Anatomija i fiziologija srca – Kardiologija , 2021.

[21] Heart chambers, CG Heart - Srčana komora - Wikipedia, 2021

[22] MRI, Magnetic Resonance Imaging (MRI) (nih.gov), 2021

[23] MRI, Što je magnetska rezonancija? - Specijalna bolnica Akromion, 2021

[24] MRI device, Cardiac MRI Becoming More Widely Available Thanks to AI and Reduced Exam Times | DAIC (dicardiology.com), 2021

[25] MRI BASIC PRINCIPLES AND APPLICATIONS THIRD EDITION, Mark A.Brown, Richard C. Semelka,

[26] Physics background of MRI Kako funkcionira MRI (uređaji za magnetsku rezonanciju)? - Znanost Blog, 2021

[27] Medical Image File Formats, Michele Larobina, Loredana Murino

[28] Robb RA, Hanson DP, Karwoski RA, Larson AG, Workman EL, Stacy MC. Analyze: a comprehensive, operator-interactive software package for multidimensional medical image display and analysis. *Comput Med Imaging Graph.* 1989;13(6):433–54. doi: 10.1016/0895-6111(89)90285-1.

[29] Bidgood WD, Jr, Horii SC, Prior FW, Van Syckle DE. Understanding and using DICOM, the data interchange standard for biomedical imaging. *J Am Med Inform Assoc.* 1997;4(3):199–212. doi: 10.1136/jamia.1997.0040199

[30] MINC 2.0: A Flexible Format for Multi-Modal Images, Robert D. Vincent, Peter Neelin, Najmeh Khalili-Mahani, Andrew L. Janke, Vladimir S. Fonov, Steven M. Robbins, Leila Baghdadi, Jason Lerch, John G. Sled,4,5 Reza Adalat, David MacDonald, Alex P. Zijdenbos, D. Louis Collins. Alan C. Evans

[31] Convolutional Neural Networks, CS231n Convolutional Neural Networks for Visual Recognition, 2021

[32] ReLU, Activation Functions in Neural Networks | by SAGAR SHARMA | Towards Data Science, 2021

[33] Arhitektura CNN-a An example of CNN architecture. | Download Scientific Diagram (researchgate.net), 2021

[34] A Review of Deep-Learning-Based Medical Image Segmentation Methods, Xiangbin Liu, Liping Song, Shuai Liu and Yudong Zhang

[35] Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medical Imaging Analysis, Ruixin Yang, Yingyan Yu

[36] How PSPNet works? | ArcGIS for Developers, 2021.

[37] Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* 2006 Jul 1;31(3):1116-28

[38] http://www.itksnap.org/, 2021

[39] Bayes Classifier, Naive Bayes Classifier. What is a classifier? | by Rohith Gandhi | Towards Data Science, 2021

[40] Background Noise Removal in Cardiac Magnetic Resonance Images Using Bayes Classifier, Ahmed S. Fahmy

[41] U-Net: Convolutional Networks for Biomedical Image Segmentation, Olaf Ronneberger, Philipp Fischer, and Thomas Brox

[42] U-net architecture, UNet — Line by Line Explanation. Example UNet Implementation | by Jeremy Zhang | Towards Data Science, 2021

[43] Metrics for evaluating semantic models, Metrics to Evaluate your Semantic Segmentation Model | by Ekin Tiu | Towards Data Science, 2021

[44] IoU picture, Intersection over Union - visual equation - Jaccard index - Wikipedia, 2021

ABSTRACT

Title: Segmentation of heart chambers from 2D MRI images using U-net convolutional neural network

In this thesis, a system for segmentation of 2D MRI images of heart chambers is presented. The U-net convolutional neural network was used for the segmentation process. The clinical background on the structure of the heart and the physical background of the operation of MRI device were given throughout the thesis. Magnetic resonance image recording formats and convolutional neural networks are described as well. The system applies a special image processing method based on Bayes' theorem. The Results evaluation compares the segmentation results with and without data pre-processing. The conducted evaluation achieves better results with data pre-processing, but there are cases in which due to too strong data pre-processing, the results are not always satisfactory.

Keywords: Segmentation, MRI, U-net, Bayes' theorem, Heart