UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Faculty of Engineering
Master's Degree in Biomedical Engineering

Multi-structure semantic segmentation of echocardiography images using deep learning

Candidate:
MHD Jafar Mortada

Supervisor:
Prof. Laura Burattini

Co-Supervisors:
Dr. Agnese Sbrollini
Dr. Selene Tomassini

Academic Year 2021-2022

UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Faculty of Engineering
Master's Degree in Biomedical Engineering

Multi-structure semantic segmentation of echocardiography images using deep learning

Candidate:
MHD Jafar Mortada

Supervisor:
Prof. Laura Burattini

Co-Supervisors:
Dr. Agnese Sbrollini
Dr. Selene Tomassini

Academic Year 2021-2022

Università Politecnica delle Marche
Faculty of Engineering
Master's Degree in Biomedical Engineering
Via Brecce Bianche – 60131 Ancona (AN), Italy

Acknowledgments

All thanks and appreciation to professor Laura Burattini and all professors at Università Politecnica Delle Marche for their great support and precious effort and special thanks to Dr. Agnese Sbrollini and Dr. Selene Tomassini who were with me at each step doing this research.
Also, I would like to thank Instituto Superior Técnico in Lisbon where I spent two semesters as a part of my ERASMUS, and I can't talk about learning institutions without thanking Damascus University where I got my bachelor's degree, especially professor Hani Amasheh, and professor Bassam Lala, and Al-Mohsinia School in Damascus, my school where I spent 12 great years.
I would not have been made it this far without you, my family, my father Ehsan, my mother Omayma, my sisters Nour and Nada, and my brother-in-law Mohammad.
To the best thing that ever happened in my life, my nephews Jawad, Kareem, Hadi, and Hasheem, and my niece Princess Alia.
To the guys who made me call Ancona my home, Alessandro, Simone, Claudio, and Eleonora.
To those whom I shared the Portuguese days with, Renata and Pedro and their sweet family, Alice, Leonardo, Pedro, and of course senhorita Ester, thank you for being my Portuguese family.
To my friends in France, Mireille, Jack, Vivian, Claude, Yvan, Michel, and last but most definitely not least Lucile.
To my Tunisian friends who were always there for me whenever I needed them, Bilal and Norhan.
To my family and friends in Germany Yaser, Shahd, Yamen, Siraj, and Zahra.
To all my friends in Syria who were always very close regardless of the long distance, especially Noor Orfahly, Suzan Alnajdi, Mohannad Ali, Ibrahim Saleh, Abdulmalek Hinnawi, AlFata, Hanna's family, and Haidar's family.
To all my friends in Apricot, especially, Wael, Bassam, Abduljwad and, I can't forget my friend and mentor, Mr.Rebhi Alasadi.
Last, but again most definitely not least, to the two brothers with whom I shared everything, who were always here, to YOU and HIM the most amazing friends anyone could have, Hanna and Haidar, may this friendship last ... forever.
Finally, I say Thank you, Obrigado, Grazie, Merci, Danke, and شكرا.

Ancona, Febbraio 2023

MHD Jafar Mortada

Abstract

The heart is a complex organ with multiple structures. It is divided into four chambers, left and right ventricles, and left and right atriums. Although multiple ways can be used to image the heart structures, using UltraSound (US) rays in echocardiography is considered the most economic solution, and yet a great tool that can be used to perform early detection of many hearts malfunctions.
Performing semantic segmentation on echocardiography images is an important step to evaluate the function of the heart. If done manually, it requires both time and experience and is highly prone to error. Thus, an automatic semantic segmentation procedure may be necessary, but it is still challenging due to the low Signal-to-Noise Ratio (SNR) of the US images and the wide range of patient characteristics.
Conventional image segmentation methods, such as edge detection, contour and shape detection, and deformable models, were used to deal with this problem, but with the advancement of deep learning techniques especially convolution neural networks, a lot of researchers tried to create a deep learning module to perform an accurate semantic segmentation of the echocardiography images.
In this thesis, a Convolutional Neural Network (CNN) system based on the YOLOv7 algorithm and U-Net architecture was proposed to automate the segmentation of the Left Ventricular endocardium (LV_{endo}), Left Ventricular epicardium (LV_{epi}) and Left Atrium (LA). The system was trained and tested on the Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) dataset, which consists of clinical exams from 500 patients, acquired at the University Hospital of St Etienne (France), for each patient, exists two views of Apical two champers view (A2C) and Apical four champers view (A4C) during End-Systolic (ES) and End-Diastolic (ED) events, this dataset was divided into training, validation and test sets.
One system was trained for both views, and The system was able to achieve Dice Similarity Coefficient (DSC) of (91.42%, 85.3%, and 88.19%) for LV_{endo}, LV_{epi} and LA segmentation respectively, and Hausdorff distance (HD) of (3.88 pixels, 4.96pixels, and 4.00pixels) for LV_{endo}, LV_{epi} and LA segmentation respectively. To our knowledge, our system is unique in the way that it implements the YOLO algorithm, and in terms of evaluation metrics it achieved good results compared to the literature.

Contents

List of Figures

List of Tables

Introduction

The human circulation system consists of the blood vessels, blood, and the heart, which is a complex organ that plays the role of pumping blood and regulating its pressure.
Like other organs in the human body, there are multiple ways to image the heart, and each has its advantages and disadvantages, while magnetic resonance imaging (MRI) consider the golden standard in terms of image quality, the US is much cheaper and more accessible, and can image the heart from many angels giving many views each has it diagnostical purposes.
In many clinical cases performing segmentation on these images is a crucial step, usually, this is done manually, whereas an expert would perform it. However, this is both time-consuming and subject to errors, in order to solve this, efforts are made to automate this procedure.
These efforts are in two classes, traditional methods, and machine learning methods, traditional methods use techniques like edge detection, *i.e.* robust methods that rely on handcrafted features, which indeed succeed to a limit, the machine learning methods appeared to exceed these limits, and proved itself to be the future of image processing.
Due to limited resources, there was not much research on this particular topic. However, in 2019 a research team introduced a new dataset called CAMUS, consisting of annotated echocardiography images, and launched a challenge to perform semantic segmentation on it.
In this thesis, we introduce a system to perform semantic segmentation on echocardiography images, taken from the CAMUS dataset, to perform semantic segmentation of the LV endocardium, the LV epicardium, and the LA, the system relied on machine learning, and consists of two sequential steps, and evaluates the performance of our system in terms of images segmentation evaluation metrics.

Chapter 1

Anatomy and physiology of the human heart

1.1 Introduction

The human circulation system consists of the heart, blood vessels, and blood, the blood vessels circulate the entire human body, providing blood to its organs and tissues. Blood carries oxygen and nutrients to all the parts of the body and carries carbon dioxide and other waste materials away from tissues, while the heart plays the role of pumping the blood and regulating its pressure. Figure 1.1 shows an illustration of the human circulation system with its parts and connections to the body organs.

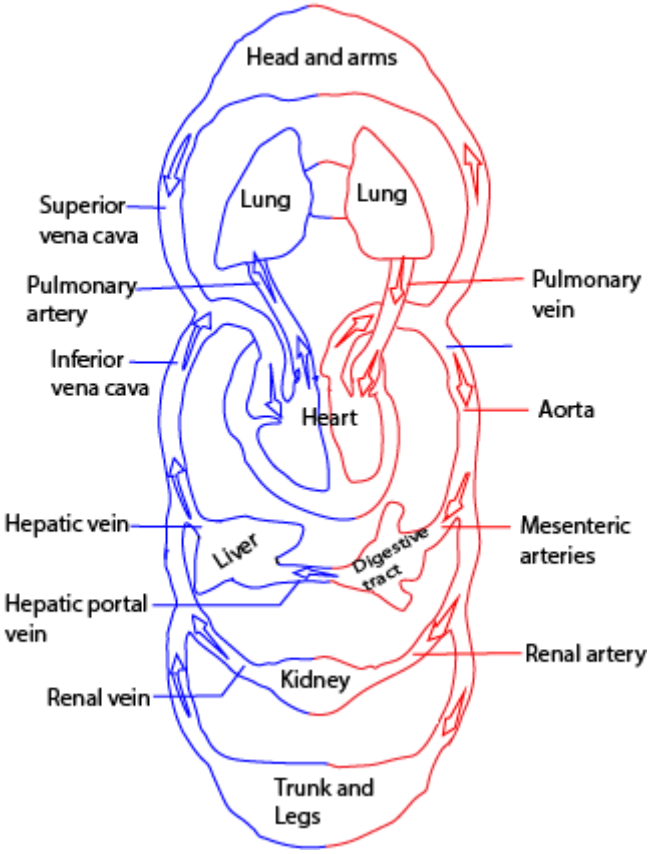

Figure 1.1: Human circulation system.

1.2 Anatomy of the human heart

1.2.1 Positioning

The human heart lies in the protective thorax, posterior to the sternum and costal cartilage, it rests on the superior surface of the diaphragm, occupying a space between the plural cavities called the (middle mediastinum), which can be defined as the space inside the pericardium. It is located between the two lungs, which occupy the lateral spaces, called the pleural cavities. The space between these two cavities is referred to as the mediastinum, and it assumes an oblique position in the thorax, with two-thirds to the left of the midline[1]. Figure 1.2 shows the position of the heart inside the protective thorax and the surrounding tissues.

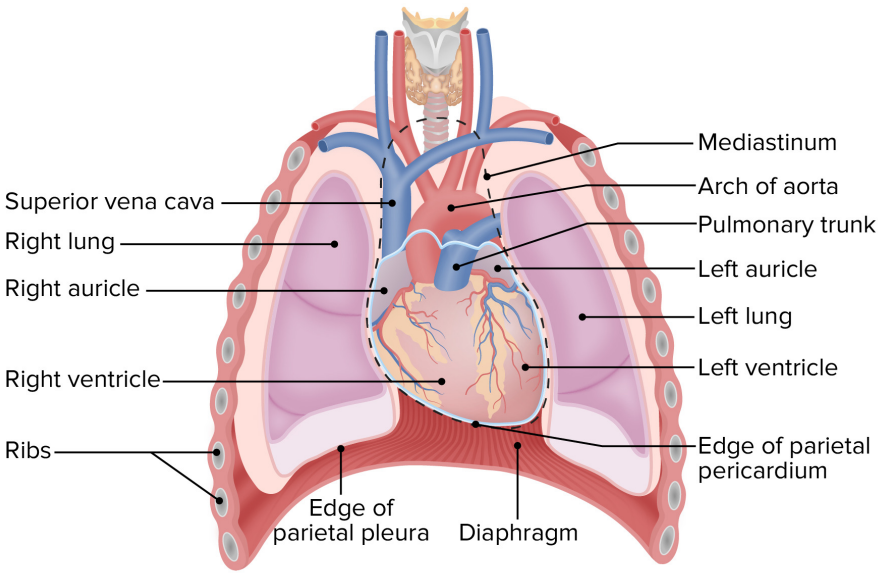

Figure 1.2: Position of the heart inside the protective thorax.

In a rare condition called "Dextrocardia", the Heart is located on the right side of the thorax (Medial), which in itself does not provide any life-threatening and accords only once in 12,000 pregnancies [2].

1.2.2 Shape and dimensions

The heart is a conical hollow muscular organ, measures (12cm x 8.5cm x 6cm), it weighs around (280g - 340g) in males and (230g -280g) in females [3].

1.2.3 Structure

The heart is divided into right and left sections, each one has two sections called chambers, in total the heart has four chambers:

- Right Ventricle:
  The right ventricle (RV), is the most anteriorly positioned chamber, sitting directly posterior to the sternum. Anteriorly, the RV is convex, with the pericardium separating it from the thoracic wall[4]. RV is connected to the pulmonary artery.

- Left Ventricle:
  The left ventricle (LV), is situated posterior to the right ventricle, it has a cone shape similar to the RV but more extensive and narrower, and its wall is three times thicker than the walls of the RV with a typical thickness of (12–15)mm, and this gets thinner as we approach the apex with the wall of the apex measures only (1–2)mm thickness.[5], LV is connected to the aorta which is the main artery that carries blood away from your heart to the rest

of your body, and separating it from the RV is The "interventricular septum", also known as the "ventricular septum", which is a triangular wall of cardiac tissue that separates the left and right ventricles (**i.e.**, the lower chambers) of the heart. The entire interventricular septum can be further divided into two parts: a muscular portion and a membranous portion.

- Right Atrium:
  The right atrium (RA), is positioned anteriorly, located above the RV. and separated from the LA by the interatrial septum. It can be described as anterolateral to the right side of the left atrium. [6]. The main vessels entering RA are the superior vena cava and the inferior vena cava

- Left Atrium:
  The left atrium (LA), It has cuboidal-shaped and is housed at the base of the heart, and is the most posterior of all the cardiac chambers. [6]. The pulmonary veins enters the LA.

Figure 1.3 shows the chambers of the heart and the anatomical structures that separate/connect them together.

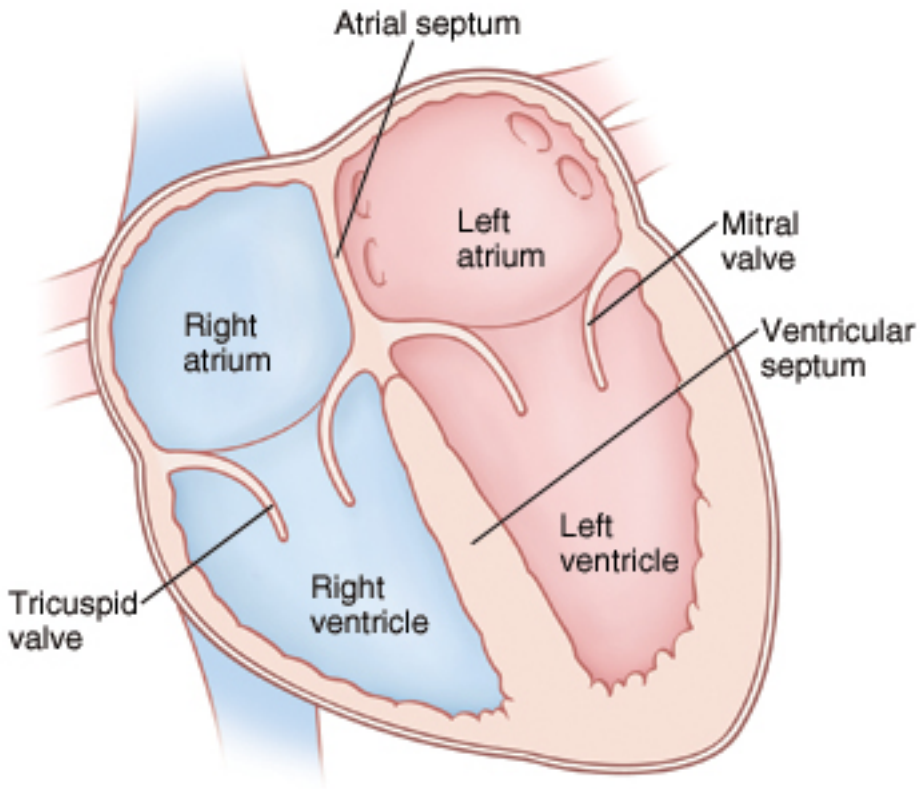

Figure 1.3: Heart chambers.

1.2.4 Valves

Valves are anatomical structures that maintain unidirectional blood flow, in another word, it controls the flow of blood and make sure that it travels in one direction, also it can open and close, which would stop the flow. There are four valves in the human heart, two are between heart chambers called "atrioventricular valves", and two between chambers and vessels called "semilunar valves", they are as follows:

- Tricuspid valve:
  The tricuspid valve (TV), is located between the RV and the RA, it prevents the blood from flowing back to the RA, typically composed of 3 leaflets of unequal size. However, in some variants, it was found that two leaves or more than three, these leaflets are referred to as the septal, anterior, and posterior leaflets. [7]

- Mitral valve:
  The mitral valve (MV) is also known as the bicuspid valve. It is located between the LV and LA [8], preventing the blood from flowing back to the RA, It has two leaflets, and the opening of the mitral valve is surrounded by a fibrous ring known as the mitral annulus.

- Pulmonary Semilunar valve:
  The pulmonary semilunar valve (PV), located between the RV and the pulmonary artery, is composed of three valve leaflets, each attached to its respective sinus, which prevents the blood from flowing back to the RV[8].

- Aortic Semilunar valve:
  The aortic semilunar valve (AV), located between the LV and the aorta, is similar to the PV as it is also composed of three valve leaflets[8].

These valves are shown in Figure 1.4.

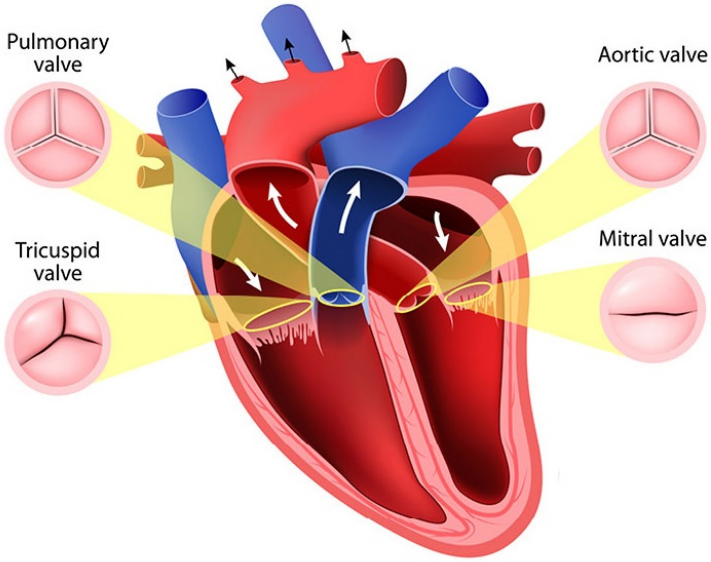

Figure 1.4: Heart valves.

1.2.5 Layers of the walls

The heart wall consists of three layers: the endocardium, myocardium, and epicardium. as shown in Figure 1.5.
The endocardium: is the thin membrane that lines the interior of the heart, while, the myocardium: is the middle layer of the heart. It is the heart muscle and is the thickest layer of the heart, and the epicardium: is a thin layer on the surface of the heart in which the coronary arteries lie. The covering that directly surrounds the heart and defines the pericardial cavity is called the pericardium or pericardial sac. It also surrounds the "roots" of the major vessels or the areas of closest proximity to the heart. The pericardium, which literally translates as "around the heart," consists of two distinct sublayers: the sturdy outer fibrous pericardium and the inner serous pericardium. The fibrous pericardium is made of tough, dense connective tissue that protects the heart and maintains its position in the thorax. The more delicate serous pericardium consists of two layers: the parietal pericardium, which is fused to the fibrous pericardium, and an inner visceral pericardium, or epicardium, which is fused to the heart and is part of the heart wall. The pericardial cavity, filled with lubricating serous fluid, lies between the epicardium and the pericardium. The epicardium consists of a simple squamous epithelium called a mesothelium, reinforced with loose, irregular, or areolar connective tissue that attaches to the pericardium. This mesothelium secretes the lubricating serous fluid that fills the pericardial cavity and reduces friction as the heart contracts. Figure 1.5 illustrates the pericardial membrane and the layers of the heart[9].

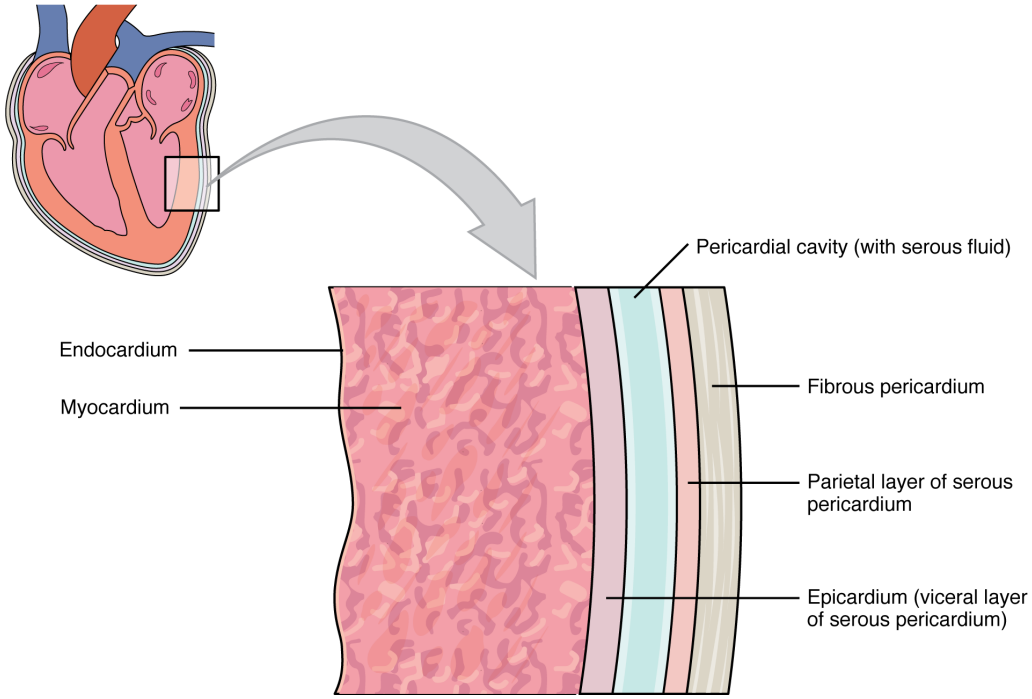

Figure 1.5: Layers of the heart walls.

1.2.6 Conduction system

The heart is able to create its own electrical impulses and control the route the impulses take via a specialized conduction pathway. This pathway consisted of five elements:

- Sinoatrial node:
  The sinoatrial node (SA): is a small, flattened, ellipsoid strip of specialized cardiac muscle, it is 10 to 20mm long and 2 to 3mm wide and tends to narrow caudally toward the inferior vena cava. It is located in the superior posterolateral wall of the RA immediately below and slightly lateral to the opening of the superior vena cava[10].

- Atrioventricular node:
  The atrioventricular node (A-V) node is located in the posterior wall of the RA immediately behind the TV, Morphologically, the A-V can be subdivided into the lower nodal bundle and compact node (CN). From the lower nodal bundle, the rightward inferior nodal extension spreads along the TV toward the coronary sinus and the leftward nodal extension spreads from the CN along the tendon of Todaro[11].

- Bundle of His:
  The bundle of His is a bundle of specialized muscles for electrical conduction.

- The left and right bundle branches:
  The Bundle of His is divided into two branches, one to conduct the electrical pulses to the LV, and the other for the RV.
- The Purkinje fibers:
  Purkinje fibers (or Purkyne tissue) are located in the inner ventricular walls of the heart, just beneath the endocardium.

Figure 1.6 shows the positions of these structures inside the heart.

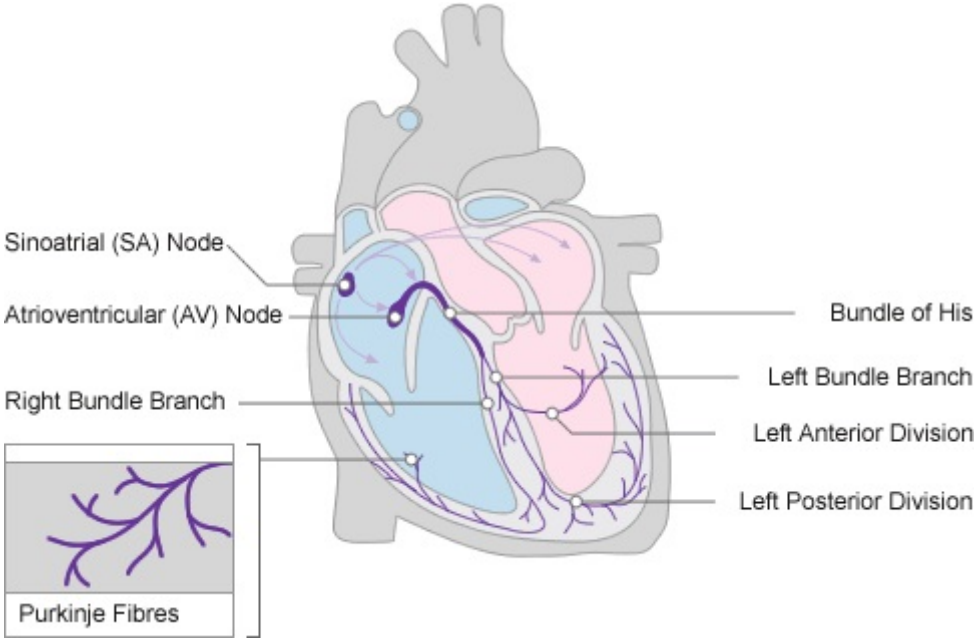

Figure 1.6: Conduction system of the heart.

1.3 Heart physiology

1.3.1 Function

The heart has three main functions: moving blood throughout the body, controlling the rhythm and speed of the Heart rate, and maintaining blood pressure. The three functions are related and integrated together. The heart acts as the pump of the blood circulation system, this function can be described using two separate parts: systemic circulation which can be summarized as moving the oxygenated blood from the lungs to the RV, then into LV, and from that to the body organs, and pulmonary circulation which can be summarized as moving deoxygenated blood from body organs to the RA, then into LA and from that to the lungs where it can be oxygenated again, this is done during what is known as the heart cycle, which can be divided into two periods: one during which the heart muscle relaxes and refills with blood, called diastole, following a period of robust contraction and pumping of blood, called systole.

1.3.2 Mechanical events of cardiac cycle

The mechanical events during this cycle can be described as the following: The Heart cycle starts with the contraction of the two ventricles, which happens almost synchronously. As contraction starts in the ventricle, the blood pressure there grows rapidly. At this stage, the AV is still closed because the pressure in the aorta exceeds that in the ventricle. As the pressure in the ventricle grows larger than in the atrium and, after a very short period of backward flow into the atrium, the mitral valve closes. The valve closure is accompanied by a sound that is audible at the chest. It is known clinically as the first heart sound. This sound marks the start of the systole, which is the period of ventricular contraction. The pressure in the ventricle keeps rising until it exceeds that in the aorta. During this phase, there is no change in ventricular volume as there is no flux through the valves and the blood is effectively incompressible. This phase is known as the isovolumetric period. When the pressure in the ventricle exceeds that in the aorta the aortic valve opens. At this moment the blood ejection into the systemic circulation starts. As the tension in the ventricle wall falls, the ventricular pressure starts to decrease. The pressure gradient between the ventricle and the aorta is reversed and flow starts to decelerate. After a short period of backflow into the ventricle the aortic valve closes. This generates the second heart sound, which marks the onset of the diastole. At this stage, all valves are closed and a second isovolumetric period occurs during which the ventricular muscle relaxes and the pressure in the ventricle decrease. At the same time, the pressure in the atrium rises again as the left atrium is filled by the pulmonary venous system. When the pressure in the atrium exceeds that of the ventricle the mitral valve reopens. At this stage, the blood flow refills the ventricle. This process is initially passive, driven by a pressure difference between the atrium and the ventricle (80% in volume). Then, it becomes active as the atrium contracts and the atrial systole pushes the remaining 20% of blood volume. Shortly after that the ventricle contracts again, starting the same cycle again. [8]

1.3.3 Electrical events of cardiac cycle

The SA node generates the rhythmic pulse through an action potential, *i.e.* an electrochemical signal that propagates as a traveling wave along the neurons. The pulse goes from the SA node to the atria and, through the internodal pathways, to the AV node. There it is delayed to let the atria empty into the ventricles before starting the ventricular contraction. From the AV node, the pulse moves through the atrioventricular bundle, which splits into the right and left branches, reaching the ventricles about 0.16s after the initial SA node impulse. The term used for the release (discharge) of an electrical stimulus is "depolarisation", and the term for recharging is "repolarization", which means the electrical events can be described as three stages atrial depolarisation, ventricular depolarisation atrial, and ventricular repolarization. The electrical events produced by the heart can be captured and

recorded by using electrodes placed on the skin in specific places, this recording is known as the "electrocardiogram" or "ECG", Figure 1.7 shows an ECG for a healthy subject during one heart cycle

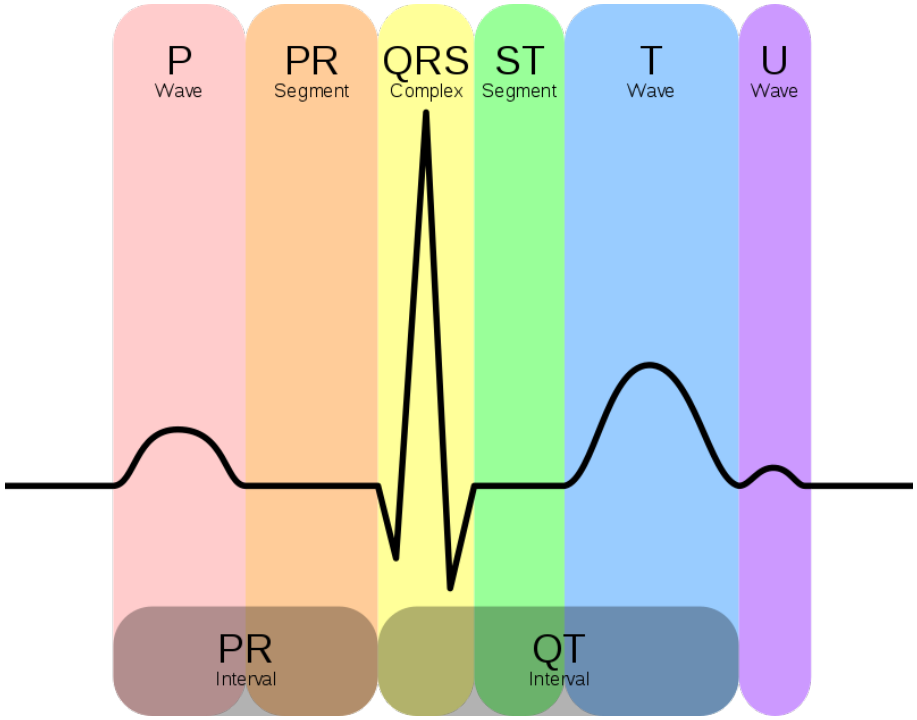

Figure 1.7: Electrocardiogram for a healthy subject during one cardiac cycle.

The first wave (P-wave) represents atrial depolarisation, in fact, the contraction of the atriums is required to empty only around 30% of the blood which means only a small muscle mass is required thus explaining the relatively lower voltage, the flat line after the P wave (P-R segment) is because the stimulus is delayed in the bundle of His, as the electrical stimulus passes from the bundle of His into the bundle branches and Purkinje fibers, it causes the depolarisation of the ventricles, and appeares on the ECG as the (QRS-complex), which can be further explained as during Q wave the depolarisation is in the septum, while the R wave represents the electrical stimulus as it passes through the main portion of the ventricular walls, and S wave represents depolarisation in the Purkinje fibre. Both ventricles repolarise before the cycle repeats itself and therefore the (T wave) is visible representing ventricular repolarization. The U wave occurs when the ECG machine picks up the repolarisation of the Purkinje fibers. However, it is very common not to see the U-wave ECG.[12] The connection between the electrical and mechanical events can be shown by the "Wiggers' diagram" illustrated in Figure 1.8, the Figure shows the cardiac cycle events occurring in the left ventricle. In the atrial pressure plot: wave "a" corresponds to atrial contraction, wave "c" corresponds to an increase in pressure from the mitral valve bulging into the atrium after closure, and wave "v" corresponds to passive atrial filling [13].

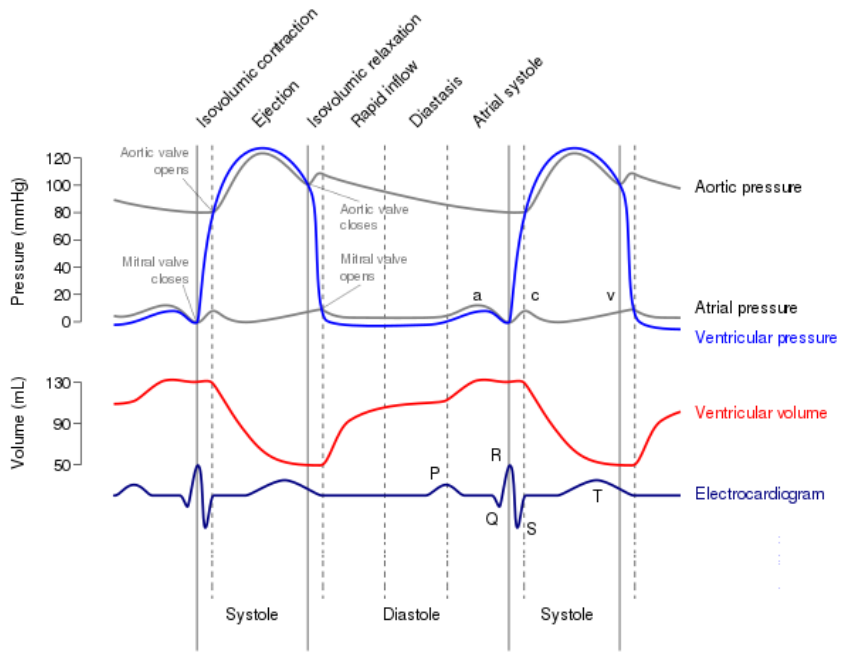

Figure 1.8: Wiggers diagram

1.3.4 Ejection fraction

Left ventricle ejection fraction or is the fraction of chamber volume ejected in systole (stroke volume) in relation to the volume of the blood in the ventricle at the end of diastole, where stroke volume is calculated as the difference between end-diastolic volume and end-systolic volume, equation 1.1 describes how we can get this parameter mathematically[14].

$$EF\% = \left(\frac{EDV - ESV}{EDV}\right) * 100 \tag{1.1}$$

EF corresponds to the ejection fraction represented as a percentage, EDV is the end diastole volume, and ESV is the end-systole volume.
EF plays an important role in diagnosing early heart failure and in the assessment of heart function overall. It can also be used as an assessment of global and segmental left ventricular function: qualitative and quantitative [14].

Chapter 2

Echocardiography

2.1 Introduction

The heart can be imaged by multiple methods, including magnetic resonance imaging (MRI), computed axial tomography (CAT), catheterization laboratory (cath-lab), and ultrasound. Although the MRI is considered the clinical gold standard for noninvasive imaging in coronary heart disease [14], the ultrasound option or Echocardiography is considerably faster and cheaper than other options as it uses non-ionizing rays (compared to Xray used by CT and cath-lab), and non-invasive (compared to cath-lab).

2.2 Physics of ultrasound imaging

US machines use mechanical natural aquatic waves, with a frequency of more than $20Khz$ which makes it not audible by humans hence the name (ultrasound). The basic principle is the same as that used in radar and sonar and is similar to the echo-location method of bats. An ultrasound machine will produce a mechanical wave, this wave will pass through the tissue, suffering from refraction, transmission, scattering from irregular boundaries, absorption, and diffraction. the part that will reflect is detected by the machine and is used to generate the image.

2.2.1 Emission and receiving of ultrasound rays

Ultrasound waves are generated using piezoelectric crystals that, when electrical impulses are applied, produce waves at frequencies determined by equation 2.1 [15].

$$f = \frac{t}{2} \tag{2.1}$$

Where f is the frequency generated by the crystal measured in hertz (Hz), t is the thickness of the crystal measured in meters (m). The lower the ultrasound frequency, the larger the penetration depth is reached but the less resolution it gives, while higher frequency has less penetration depth and more resolution. The same crystals are used to receive the reflected wave, *i.e.* it will transfer the mechanical waves into an electrical signal, this is possible because the machines apply energy and,

ultimately sound waves, in pulses [15]. The pulsatile nature of ultrasound waves produced facilitates the emission and reception of sound waves, the time between the beginning of one pulse and the beginning of the next pulse is called the pulse repetition period (PRP) or it is represented as frequency we call pulse repetition frequency (PRF) as shown in Figure 2.1.

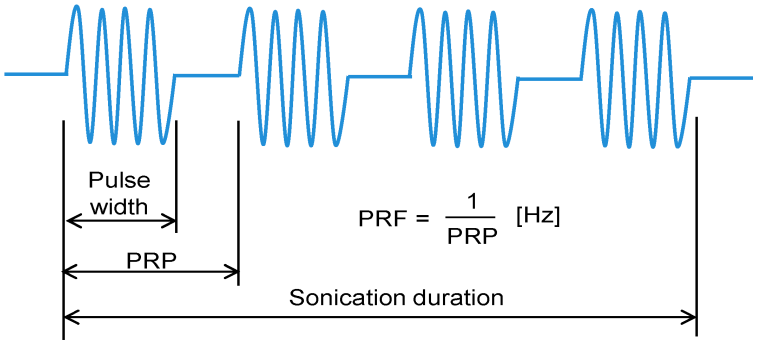

Figure 2.1: Typical ultrasound wave showing the pulse, PRP, and PRF.

Crystals are packed inside what is called a "Transducer" which is the instrument that controls emitting and receiving ultrasonic rays, different types of transducers are used for different medical applications, this will be discussed later.

2.2.2 Acoustic imaging

The depth information is encoded in the time that an aquatic wave takes to travel to a tissue an back to the transducer, hence the creation of an image relays on the knowledge of the ultrasonic waves propogation speed inside a medium, *i.e.* the rate at which waves pass through a medium. Table 2.1 shows the speed of sound in different mediums [15].

Table 2.1: Sound speed in different meduims.

Medium	Speed [m/s]
Air	330
Fat	1450
Water	1480
Soft tissue	1540
Liver	1560
Blood	1500
Muscle	1600
Tendon	1700
Bones	3500

Propagation speed depends on the characteristics of the medium that waves are traveling through and is independent of the frequency, as tissue density increases, the propagation speed decreases, by contrast, the stiffer the tissue, the higher the propagation speed.[15] As seen in Table 2.1 various tissues in the human body differ from each other in terms of their specific speed of sound, when performing abdominal scans, aberration distortions can become significant due to the change in the speed of sound between connective tissues, fat layers, muscles, and abdominal organs, despite this variation, clinical ultrasound scanners typically use an assumed speed of sound (1540 m/s) for image reconstruction.

2.3 Imaging mode

Different types of ultrasound imaging are used in medicine.

2.3.1 Amplitude mode

Amplitude mode (A-mode) is the simplest type of ultrasound. A single transducer scans a line through the body with the echoes plotted on screen as a function of depth.

2.3.2 Brightness Mode

Brightness Mode (B-mode) is the most commonly used in medicine, an array of transducers simultaneously scans a plane through the body that can be viewed as a

two-dimensional image on the screen. Figure 2.2 shows a B-Mode obstetric b-mode ultrasound image.

Figure 2.2: B-MODE ultrasound image.

2.3.3 Motion mode

Motion mode (M-mode), in which a rapid sequence of B-mode scans whose images follow each other in sequence on screen enables doctors to see and measure a range of motion, as the organ boundaries that produce reflections move relative to the probe. it is very common to use it in cardiac applications. Figure 2.3 shows an M-Mode cardiology ultrasound image.

Figure 2.3: M-Mode cardiology ultrasound image

2.3.4 Doppler mode

This mode is used to measure the speed and direction of moving objects and plot it, although, it is most commonly used to measure the speed of blood flow, it is can be also used to measure the speed of tissue such as the tissue of the heart. This mode depends on the Doppler effect: which is the change in frequency of a wave in relation to an observer who is moving relative to the wave source. equation 2.2 describe this shift [16]:

$$f_0 = \frac{\nu + \nu_o}{\nu + \nu_s} \tag{2.2}$$

Where f_o represents observer frequency of sound in (Hz) ,ν represents speed of sound waves in (m/s),ν_o is observer velocity (m/s),ν_s is source velocity (m/s),and f_s is actual frequency of sound waves (Hz).

2.4 Types of transducers

Transducers can be categorized by the type of image they produce (1D, 2D, and 3D) or by the invasion criteria (invasive like the vaginal probe or the transesophageal probe (TEE) showed in Figure 2.4 and noninvasive).

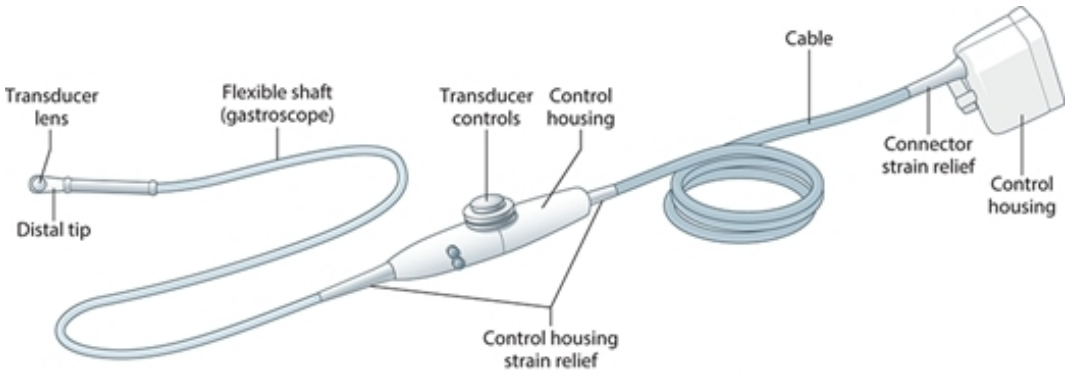

Figure 2.4: TEE Transducer and its parts.

They can also be categorized by the shape and alignment of the piezoelectric crystals array inside the transducer into:

2.4.1 Linear probes

Linear probe usually has a range of frequencies of ($3\breve{}12MHz$), as its name suggests, this type of transducer has a linear arrangement of individual transducer elements, images from a linear array are generally rectangular, and the image width corresponds to the width of the array, a set of adjacent elements is used to fire a single image line or a portion thereof, its applications are small parts, superficial vascular, obstetrics [17]

2.4.2 Curved arrays probes

Curved arrays probes form sector images, because of the shape of the aperture, a relatively wide image can be achieved using a smaller footprint aperture, scan lines are no longer parallel to each other but form a fan beam arrangement with field of view angles of up to 85°, (150° for some endocavitary arrays). Typical bandwidths range from 2–8 MHz, which is a lower range than for linear arrays since this type of probe is intended for large penetration depths where frequency-dependent attenuation prohibits very high frequencies, its applications are abdominal, obstetric, transabdominal, transvaginal, transrectal, and pediatric imaging [17].

2.4.3 Phased array probes

Phased arrays are also designed to form sector images, but contrary to curved arrays, where the natural shape of the physical aperture provides the basis for the sector shape, phased arrays steer the beam to form the image, specific timing delays for the sub-aperture can not only focus on a specified depth but also steer the beam in the lateral direction, large fields of view can be achieved this way, but the development of increased side and grating lobes is a trade-off. Anatomical locations with small diameter access to larger distal regions can be imaged with this type of ultrasound array, cardiac imaging typically relies on phased arrays due to acoustic shadowing from the rib cage, where one needs to image between narrowly spaced ribs in order to interrogate the much larger-sized heart chambers. Its applications are Cardiology, liver, spleen, fontanelle, and temple [17]. Figure 2.5 shows the three different types of probes.

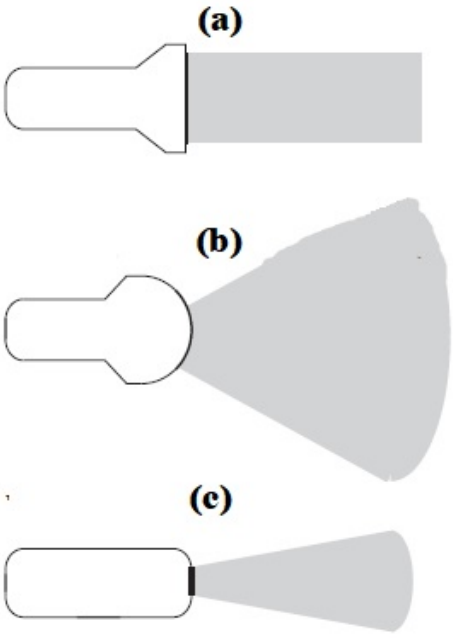

Figure 2.5: (a): Linear, (b): Curved (c): Phased array probes.

2.5 Echocardiography

Echocardiography is a type of medical imaging of the heart, using standard ultrasound or doppler ultrasound, it can be both invasive: known as a transesophageal echocardiogram (TEE), and noninvasive which is known as transthoracic echocardiogram (TTE). The TTE usually uses a phased array probe, which produces different 2D images, these images depend on the location the probe is placed on which is known as the window, each window has multiple views depending on the rotation of the probe. These windows are in five categories:

2.5.1 Parasternal long axis window

The parasternal long axis (PSLA) window can show the structures of RV, LV, LA, AV, MV, aorta (AO), and descending aorta pericardium (DA), Figure 2.6 shows an illustration of this view.

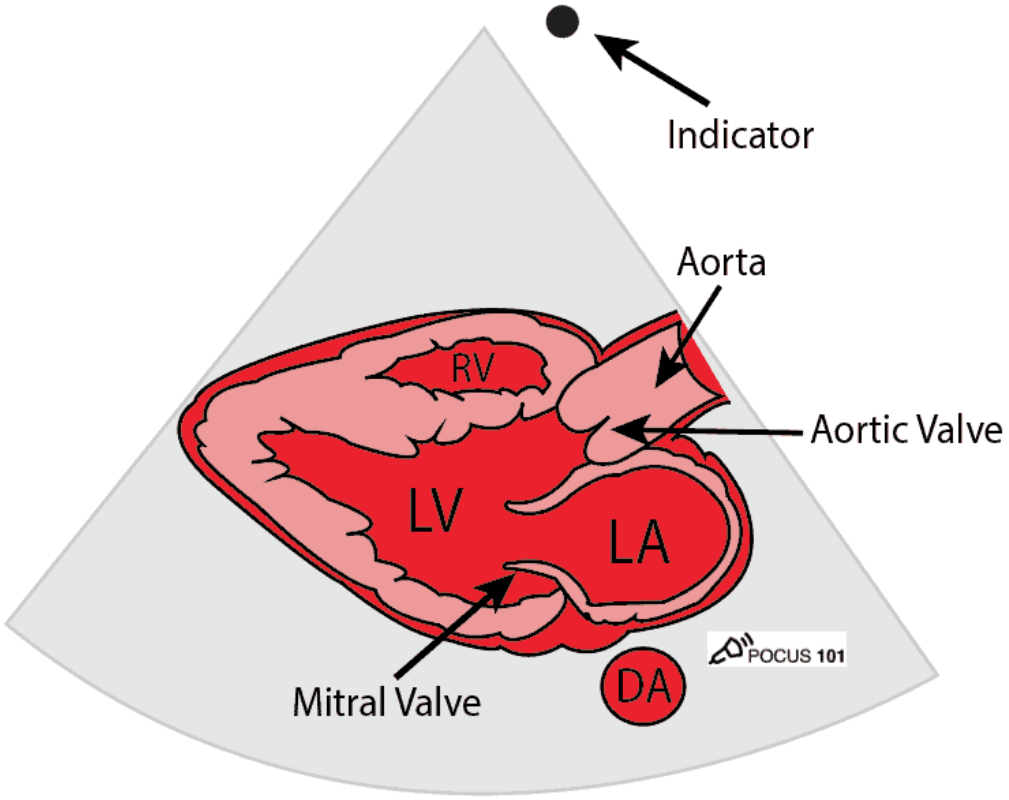

Figure 2.6: Parasternal long axis.

2.5.2 Parasternal short axis window:

The parasternal short axis (PSSA) window has three main views:

- Mid-papillary view: which shows RV and LV structures.

- Mitral valve view: which shows RV, LV, and MV: both anterior and posterior leaflets structures.
- Aortic valve view: which shows RV, TV, AV, PV, right ventricular outflow tract (RVOT), RA, and LA structures.

Figure 2.7 shows this window and the three different views.

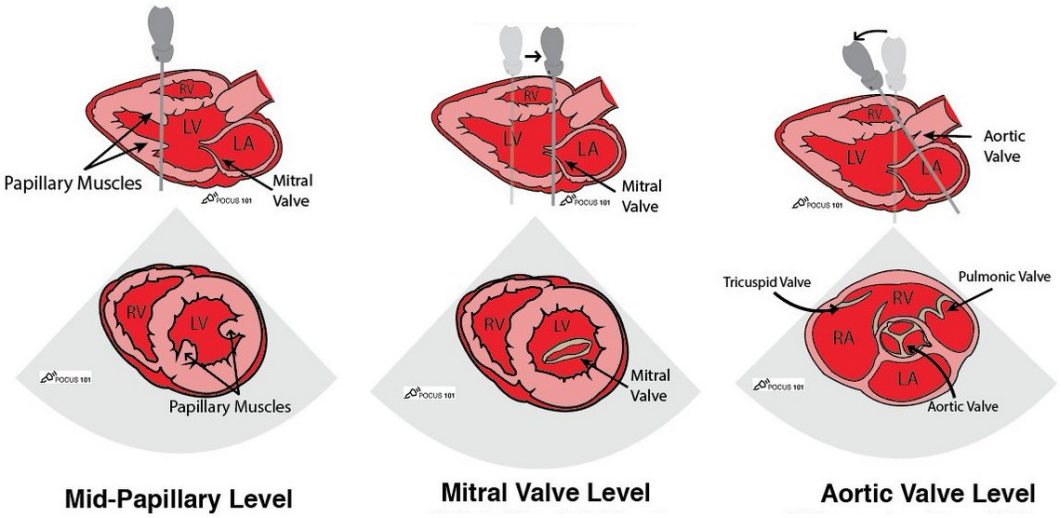

Figure 2.7: Parasternal short axis window.

2.5.3 Apical window

The apical windows are some of the most important views to be able to obtain when doing a hemodynamic assessment of the heart. This includes looking at diastolic dysfunction, valvular regurgitation, cardiac output, etc. This window has multiple views:

- Two chambers view (A2C): in which we can identify LV, anterior leaflet (AL), posterior leaflet (PL), LA, left atrial appendage (LAA), abdominal aorta(Abd Ao), and descending thoracic aorta (DTA) structures.
- Four chambers view (A4C): in which we can identify LV, interventricular septum (IVS), RA, RV, AL, PL, septal leaflet (SL), TV, AV, interatrial septum (IAS), LA, RA, right upper pulmonary vein (RUPV), and DTA structures.
- Five chambers view (A5C): in which we can identify LV, IVS, RA, RV, TV, AV, IAS, LA, RA, Ao, and aortic root (AoR) structure.
- Three-Chamber View (A3C): in which we can identify LV, IVS, RA, RV, AL, PL, SL, TV, AV, IAS, LA, RA, RUPV, and ascending aorta (AscAo) structure.

Figure 2.8 shows this window with its view.

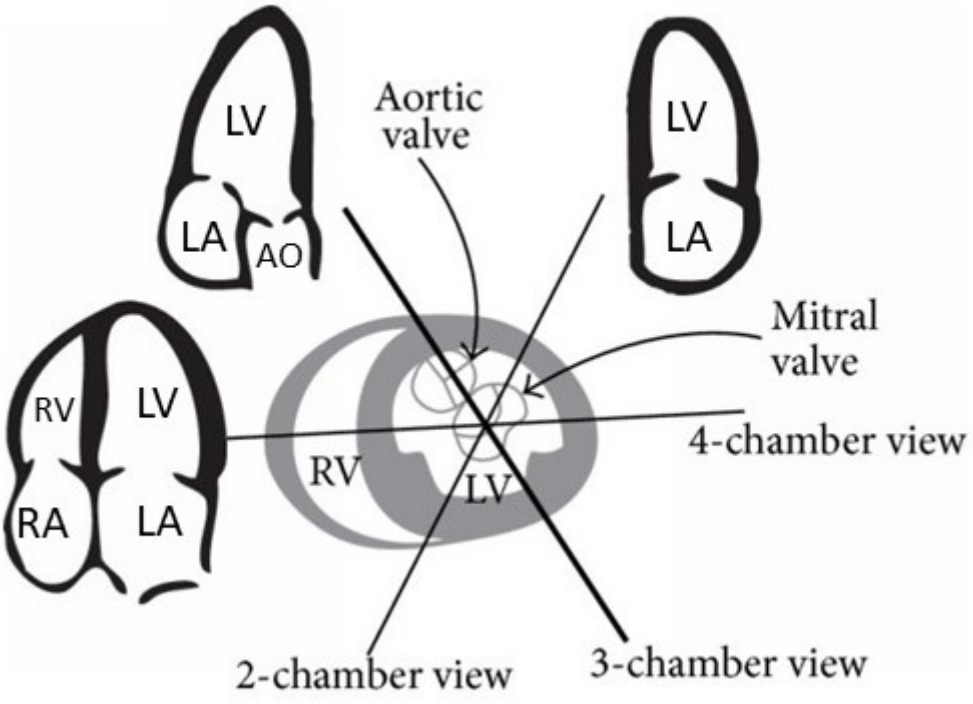

Figure 2.8: Apical window views.

2.5.4 Subcostal window

The subcostal window has the advantage of the absence of bone or lung tissue to obstruct the view of the heart This window includes the following views:

- Four-chamber view: structures to identify in this view are RV, LV, AL, PL, SL, LA, RA, IAS, pericardium PC, TV, MV, and anterolateral papillary muscle (ALPM).
- Short-axis view: structures to identify in this view are Hepatic Vein(HV), IVC, RV, TV, SL, and RA.
- Vena cava view: which shows Abd-Ao.

Figure 2.9 shows the different views of this window.

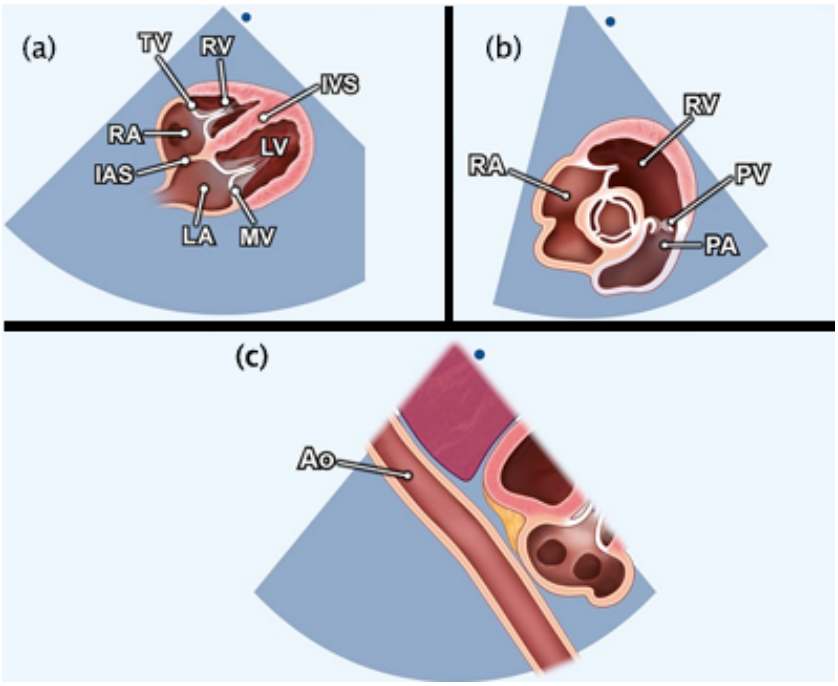

Figure 2.9: (a): Four-chamber, (b): Short axis, (c): Vena cava.

2.5.5 Suprasternal window

The suprasternal window is located in the jugulum right on top of the sternum (suprasternal notch), this window is rarely used by cardiologists, but it can be quite useful for specific situations, such as measuring the width of the aortic arch, looking for aortic dissection or coarctation, assessing retrograde flow in the DA (using color Doppler) or when quantifying aortic regurgitation. Figure 2.10 describes this window.

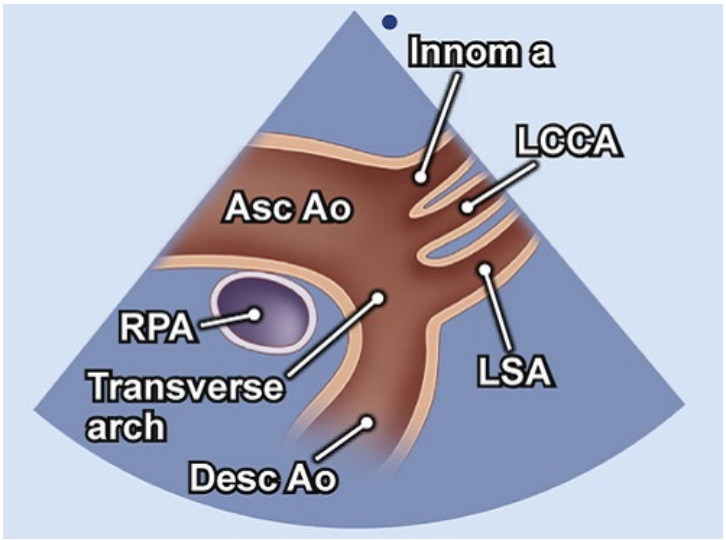

Figure 2.10: Suprasternal window.

Chapter 3

Machine learning and computer vison

3.1 Introduction

Digital processors deal with images as an array of values called pixels - a combination between the word picture and cell - *i.e.* a computer can't really see as an image as it is just a collection of numbers, this was the motivation behind developing a new field in computer science called computer vision, this field can be defined as an interdisciplinary scientific field that deals with how computers can gain high-level understanding from digital images or videos, from the perspective of engineering, it seeks to understand and automate tasks that the human visual system can do [18]. Much work has been done to develop this field and with the development of machine learning and deep learning the intersection between the two fields gave great tools, such as convolutional neural networks. Many subjects can be listed under computer vision, such as feature extraction, and image classification.

3.2 Image segmentation

An image is a collection measurement in 2D space or 3D space, these measurements can be acquired in the continuous domain (a medical example would be Xray films) or in discrete space which gives us digital images, the process in which the digital image can be divided into regions that share some characteristic such as intensity or texture is called image segmentation **i.e.** If an image is the set I the segmentation problem is to determine set S_k where it satisfy equation 3.1

$$I = \bigcup_{k=1}^{K} S_k \tag{3.1}$$

If the process is considered with only assigning a label (or class) to each pixel the problem is known as semantic segmentation, but if the problem is to separate labels for different instances of the same class, the problem becomes instance segmentation, figure 3.1[19] shows the differences between the two problems. In other words, semantic segmentation would identify the pixels which belong to the class of cells. However, instance segmentation will determine which pixels belong to each "cell" instance.

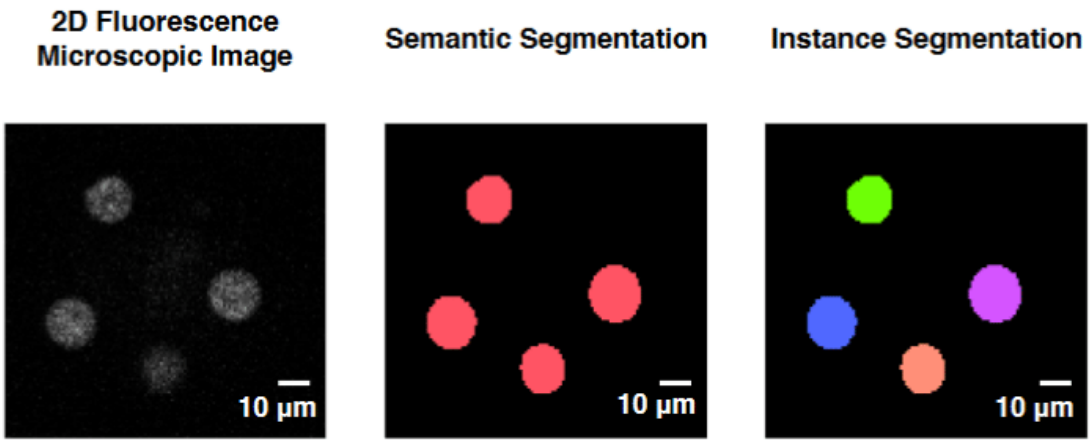

Figure 3.1: Instance segmentation and semantic segmentation.

3.3 Object detection

Object detection can be defined as detecting instances of semantic objects of a certain class, the main difference between it and image segmentation, is that finds bounding boxes around objects and classifies them. Figure 3.2 shows an object detection example in a CT scan to detect the liver and spine, it shows the bounding boxes around the detected regions.

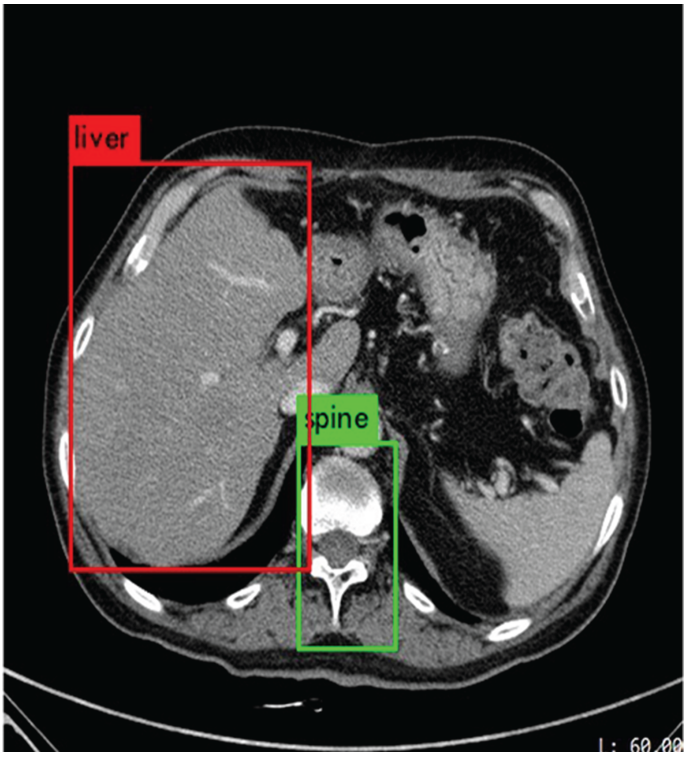

Figure 3.2: Object detection in CT image.

3.4 Evaluation metrics in image segmentation

When developing an algorithm to perform semantic image segmentation it is crucial to define quantitative metrics to evaluate the performance of the algorithm. The following describes some of these metrics:

3.4.1 Pixel accuracy

Pixel accuracy is described as the ratio of the truly predicted pixels to the number of all pixels, This metric is very problematic to use when there is a class unbalanced, an example is given in Figure 3.3, the problem is trying to segment the tumor in a CT scan image, as the tumor ratio is only 2% from the whole image, an algorithm which considers the whole image as background would still have an accuracy of 98%.

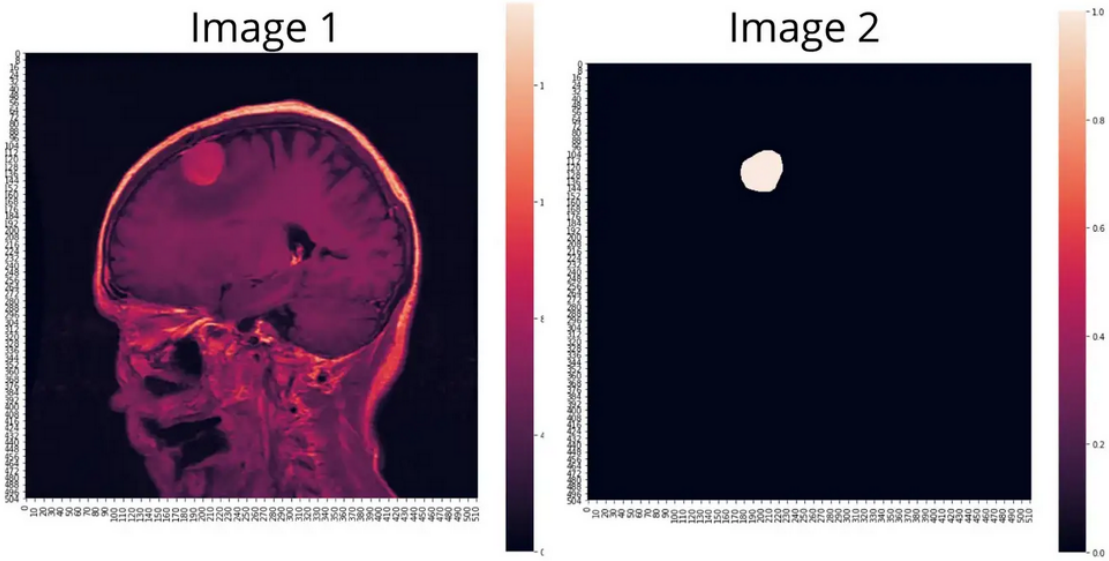

Figure 3.3: Brain tumor in CT image - Left shows the original CT image, right the ground truth mask for the tumor segmentation.

3.4.2 Jaccard's similarity coefficient

Jaccard's similarity coefficient (JSC), which is also known as Intersection over Union or (IoU), in statistics compares members for two sets to see which members are shared and which are distinct. It's a measure of similarity for the two sets of data, its value is between 0 and 1, equation 3.2 describes how to calculate this index for two sets of numbers (A) and (B):

$$JSC = \frac{|A \cap B|}{|A \cup B|} \tag{3.2}$$

In image segmentation, JSC is the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth, for binary classification (There are only two classes) JSC can be described by equation 3.3

$$JSC = \frac{TruePositive}{TruePositive + FalsePositive + TrueNegative} \tag{3.3}$$

3.4.3 Dice similarity coefficient

Simply put, the dice similarity coefficient (DSC) is very similar to JSC it only doubles the counts for intersection (true positive), as described in equation 3.4.

$$DSC = \frac{2 * TruePositive}{2 * TruePositive + FalsePositive + TrueNegative} \tag{3.4}$$

3.4.4 Hausdorff distance

The average Hausdorff distance (HD) between two finite point sets S and L is defined in equation 3.5.

$$HD_{95th} = max\{K^{th}_{s \in S} \min_{g \in G} \parallel S - L \parallel, K^{th}_{s \in S} \min_{g \in G} \parallel L - S \parallel\} \tag{3.5}$$

3.4.5 Mean absolute distance

Let Y_i and X_i denote the i^{th} contour point from the segmented contour and the ground truth, respectively, after equally spaced sampling. The Mean absolute distance (MAD) is defined as in equation 3.6.

$$MAD = \frac{1}{n} \sum_{i=1}^{n} (||X_i - Y_i||) \tag{3.6}$$

3.4.6 Center of mass distance

The center of mass distance (CMD) is the euclidean distance between the center of the mass of the ground truth mask and the center of mass of the predicted mask.

3.5 Deep learning

Deep learning methods are learning methods with multiple levels of representation, achieved by assembling simple but non-linear modules, each of which converts the representation at one level (starting from the raw input) into a representation at a higher, slightly more abstract level, by assembling a sufficient number of such transformations, very complex functions can be learned [20]. Deep learning can be "supervised", "semi-supervised", and "unsupervised", it can be used to solve regression, classification problems, clustering and other problems,

3.5.1 Artificial neuron

An artificial neuron is the base structure for an artificial neural network (ANN), loosely modeled by the neurons in a biological brain. The simplest neuron is modeled after McCulloch-Pitts neurons, which was published in 1943 [21], the neuron has inputs and an output, and it consists of a linear part, followed by a nonlinearity, as shown in Figure 3.4.

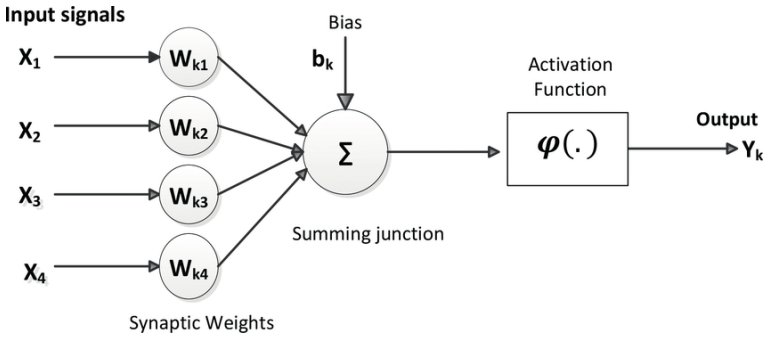

Figure 3.4: McCulloch-Pitts Neuron.

The input is processed in the neuron as shown in equation 3.7.

$$Y_k = \phi(\sum_n x_n * W_{kn} + b_k) \tag{3.7}$$

W_{kn} is called weights, and the function ϕ is called the activation function.

Activation functions

There are many choices of the activation function used by artificial neural networks:

- Step function: This was one of the earliest functions used in ANN, equation 3.8 describes this function, and it is useful to use in binary classification problems. However, the problem is the derivative of the step function is 0, which is a problem with training algorithms, Figure 3.5 shows the plot of the function.

$$f(x) = \begin{cases} 0 & x < 0 \\ 1 & x > 1 \end{cases} \tag{3.8}$$

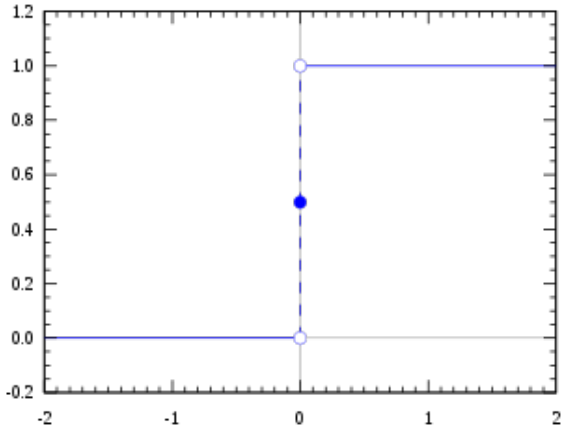

Figure 3.5: Step function.

- Sigmoid function: A sigmoid function is a bounded, differentiable, real function that is defined for all real input values and has a non-negative derivative at each point, equation 3.9 describes this function where e is Euler's number, and Figure 3.6 shows the plot of the function.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.9}$$

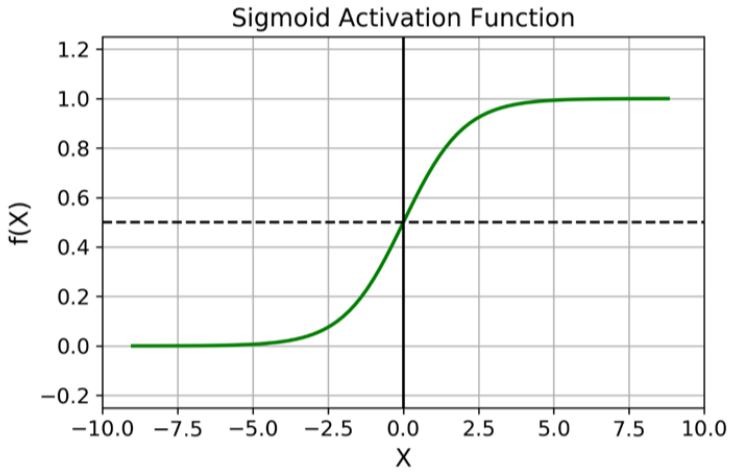

Figure 3.6: sigmoid function.

- Softmax function: This function converts a vector of real numbers into a probability distribution of possible outcomes, very useful in the case of multi-class classification. equation describes this function.

$$f(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{3.10}$$

- Rectified Linear Unit (ReLU) function: This is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero. It has become the default activation function for many types of neural networks because a model that uses it is easier to train and often achieves better performance, equation 3.11 describes this function, and Figure 3.7 shows the plot of the function.

$$f(x) = max(0, x) \tag{3.11}$$

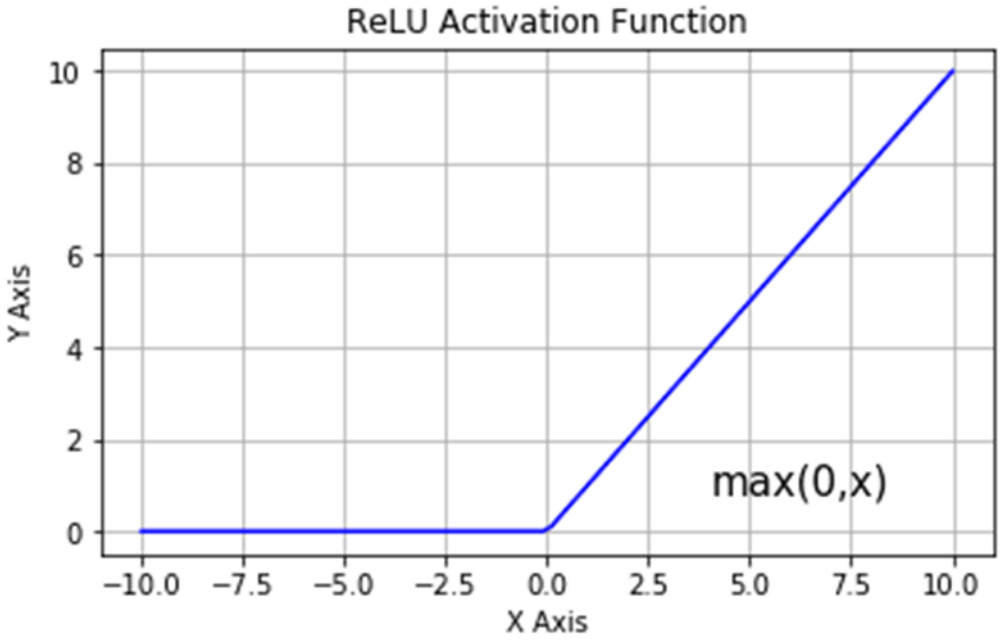

Figure 3.7: Rectified linear unit function.

- MISH function: It was proposed in 2020 [22], and it aimed on solving some of the problems which ReLU had, such as Dying ReLU, which is experienced through a gradient information loss caused by collapsing the negative, mathematically described in equation 3.12, while Figure 3.8 shows its plot.

$$f(x) = x.\tanh((ln(1 + e^x))) \tag{3.12}$$

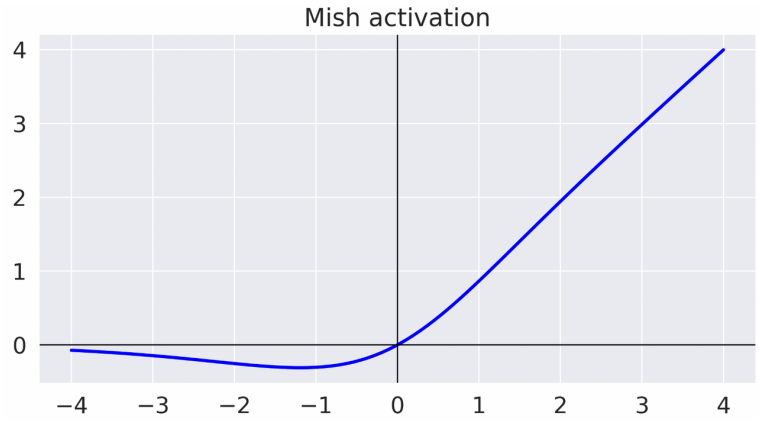

Figure 3.8: MISH function

3.5.2 Artificial neural network

Artificial Neural network (ANN) is a computation system inspired by the human brain to solve complex problems, it consists of connected Artificial Neurons, and traditional ANN can be modeled as shown in Figure 3.9, as it has an input layer, an output layer and hidden layers between them. Cybenko et. al (1989) [23] proved that a multilayer ANN with 1 hidden layer is a universal approximator of any continuous function defined on a compact subset of R^P. This is a useful theorem but it does not explain how many units are needed nor how should the weights be chosen.

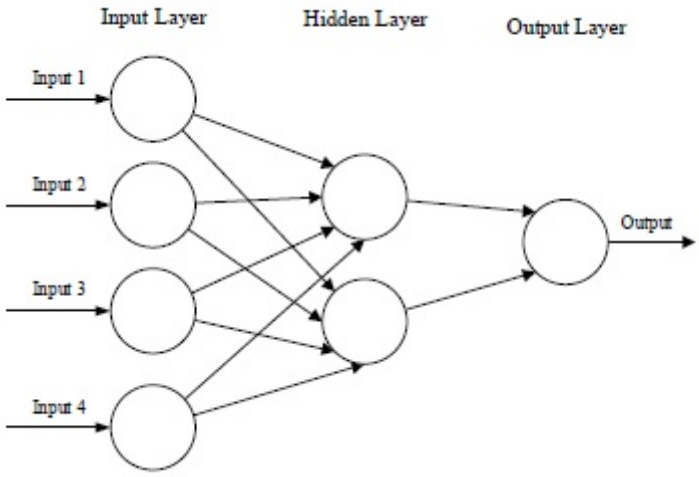

Figure 3.9: Traditional neural network.

This structure is quite useful for many applications, but when it comes to dealing with raw images as input, the computational complexity becomes very high. As an alternative to this structure, in 1980 Fukushima et. al [24] proposed a structure to extract features from the images, and in 1998 Lecun et. al [25] proposed the name "Convolutional neural network", and in 2012, the winner of ImageNet challenge used CNN, known as AlexNet [26], which reduced the error from 25.8% to 16.4%, finally

in 2015 CNN exceeded human accuracy in a classification where it is estimated that human performance has around 5.1% of error while CNN achieved an error rate of 3.57% [27]. Usually, CNN architecture has multiple different layers, that has different working principle.

Convolutions layers

Convolutions Layers are the most important layer, as they are the main building block of a CNN, it was inspired by that of the visual nervous system of vertebrates, the layer has a set of filters (or kernels) with a fixed size, and trainable weights, and this filter scans the entire image using a step called "stride". Assuming the network received an image $A^{(m-1)}$ with K_m channels as an input, the output $A^{(m)}$ will have O_m channels, where O_m is equal to the number of kernels the layer has, equation 3.13 explain it mathematically [26].

$$A_o^{(m)} = g_m \sum_k W_o^{(m)} k * A_k^{m-1} + b_o^{(m)} \tag{3.13}$$

where $W_{ok}^{(m)}$ is matrix of shape $P_m \ddot{O} Q_m$ and $b_o^{(m)} \in R$. The matrix $W_{ok}^{(m)}$ parameterizes a spatial filter that the layer can use to detect or enhance some feature in the incoming image. Figure 3.10 shows an example of a kernel of size 3X3 when applied to an image.

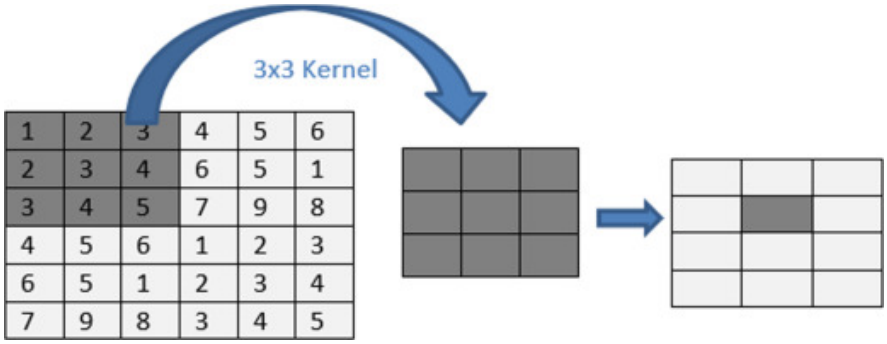

Figure 3.10: Convolutions Layer with a kernel size 3x3

Pooling layer

Pooling layers of a CNN implement a spatial dimensionality reduction operation designed to reduce the number of trainable parameters for the next layers and allow them to focus on larger areas of the input pattern. This reduction could be carried out by using functions such as max, or averages. Pooling layers' parameters are not trainable.

Dropout layer

Dropout was proposed by Hinton et al. (2012)[26] as a form of regularization for fully connected neural network layers. Each element of a layer's output is kept with probability p, otherwise is set to 0 with probability $(1 - p)$. Extensive experiments show that dropout improves the network's generalization ability, giving improved test performance.

Batch normalization layers

Batch normalization layers are built upon the idea that for every neuron (activation) in a particular layer, we can force the pre-activations to have zero mean and unit standard deviation, this can be achieved by subtracting the mean from each of the input features across the mini-batch and dividing by the standard deviation. This layer has two trainable parameters, an offset factor α and a scaling factor γ.

Transposed convolutional layer

Transposed convolutional layers are supposed to reverse the operation of a convolution layer, *i.e.* it will expand the size of the feature map, they are different from upsampling because they have trainable weights, and they work in a manner very similar to the convolutional neural networks.

3.5.3 Training artificial neural network

Training a network is the process to adjust the weights in order to minimize the error between the predicted output and the true output. Many algorithms were developed to achieve this purpose, in which the error is measured by what we call a loss function, there are many available loss functions that have different advantages depends on the problem and the nature of the data.

Dice Loss function

It depends on the dice similarity coefficients, mathematically it can be expressed by equation 3.14.

$$L = 1 - D \tag{3.14}$$

Where D is the dice similarity coefficient described in equation 3.4.

Cross categorical entropy

In binary classification, where the number of classes M equals 2, Binary Cross-Entropy (BCE) can be calculated as described in equation 3.15

$$BCE = -(y \log(p) + (1 - y) \log(1 - p)) \tag{3.15}$$

If $M > 2$ (**i.e.** multiclass classification), we calculate a separate loss for each class label per observation and sum the result, as explained in equation 3.16

$$MCCE = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{3.16}$$

Where M represents the number of classes, log is the natural log, y is a binary indicator (0 or 1) if class label c is the correct classification for observation o, and p represents predicted probability observation, and o is of class c.

3.5.4 Over-fitting

One of the problems of training any supervised machine learning model is overfitting, in which the Model does not generalize well from observed data to unseen data, in another word, the model becomes specifically designed for the data used in the training procedure. The causes of this phenomenon might be complicated, generally, we can categorize them into three kinds:

- noise learning on the training set: when the training set is too small in size or has less representative data or too many noises.
- hypothesis complexity: the trade-off in complexity, a key concept in statistic and machining learning, is a compromise between variance and bias.
- multiple comparisons procedures which are ubiquitous in induction algorithms, as well as in other artificial intelligence (AI) algorithms.

One method to solve this problem is to identify when to stop training, this is usually done by dividing the training set into two sets, the training set that is used to train the model and the validation set which is used to track the performance of the model during training when the model tends to fit the training *i.e.* the loss function for the training set starts to decay while the loss function for the validation set is increasing, the training should be stopped by that moment, which is known as "early stop"[26], figure 3.11 shows the position of this point, stopping before it is underfitting, and after it is overfitting.

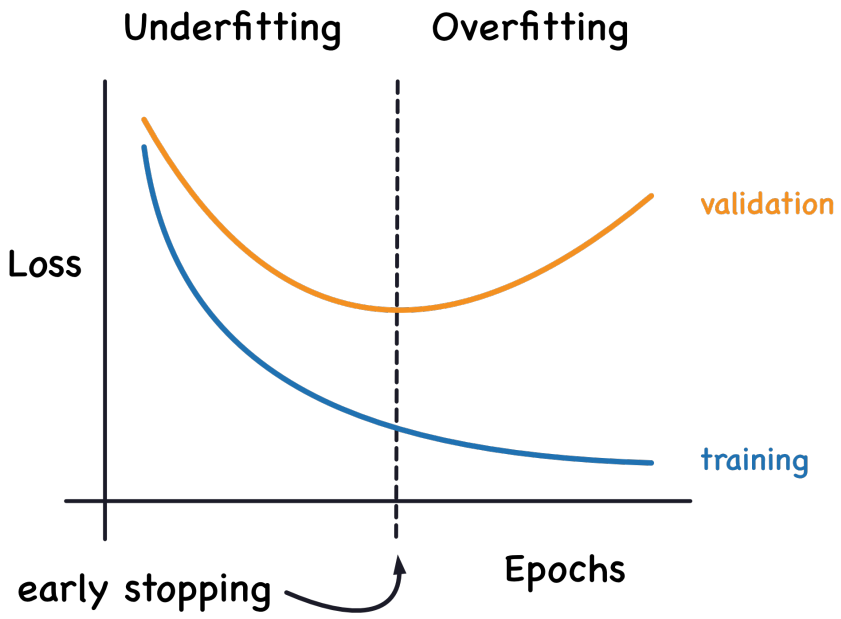

Figure 3.11: Overfitting and early stop point.

3.5.5 Fully convolutional network

A fully Convolutional Network (FCN), is a class of artificial neural network architecture, mainly used for semantic segmentation, as it takes an image with an arbitrary size and produces an image with the same size. Usually, FCN consists of a downsampling path, used to extract and interpret the context, and an upsampling path, which allows for localization [28]. Figure 3.12 shows an FCN archetechture.

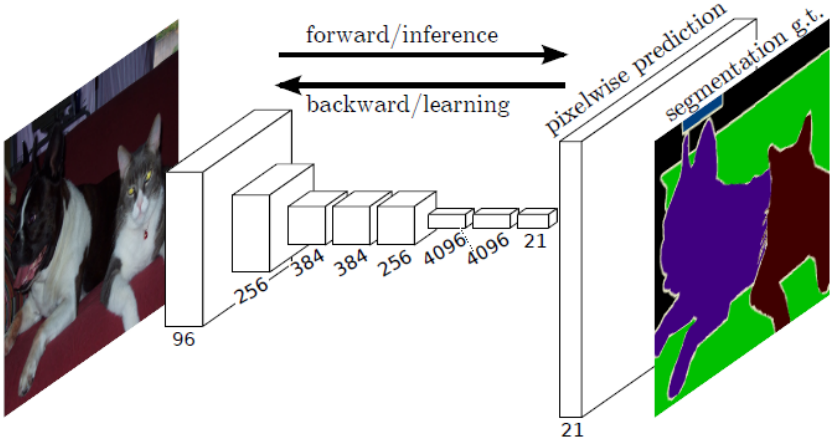

Figure 3.12: Fully connected network.

Later the concept was developed with the idea of skip connection, which as the

name suggests skips some of the layers in the neural network and feeds the output of one layer as the input to the next layers, which gave the architectures known as FCN8, FCN16, and FCN32. The three architectures share the same down-sampling path (known as encoder), but a different upsampling path, as shown in Figure 3.13 FCN-32s: Upsamples at stride 32, predictions back to pixels in a single step (Basic layer without any skip connections) FCN-16s: Combines predictions from both the final layer and the pool4 layer with stride 16, finer details than FCN-32s. FCN-8s: Adds predictions from pool3 at stride 8, providing even further precise boundaries.

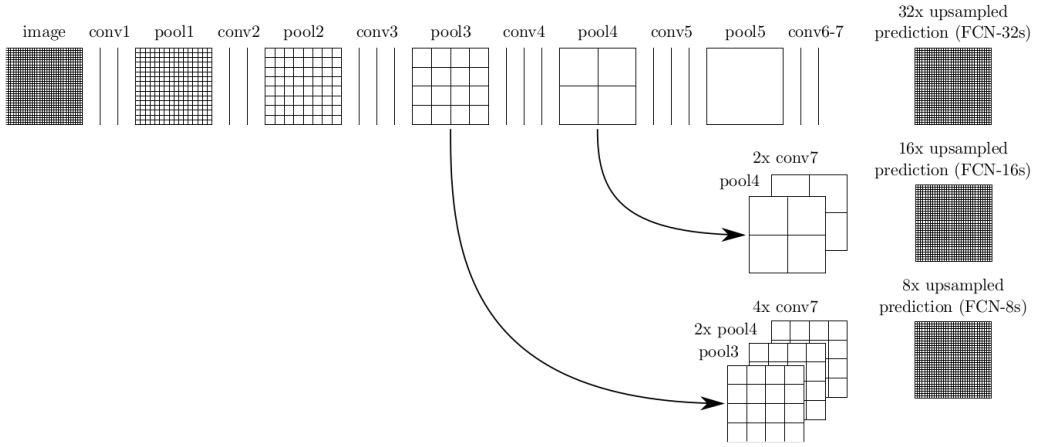

Figure 3.13: Fully connected Network 8, 16, and 32

3.5.6 UNET architecture

UNET structure was proposed by Krizhevsky et. al (2015) [26] for biomedical image segmentation, this model was built upon the before mentioned FCN model using an encoder and a symmetric upsampling path (or decoder), this decoder has the same shape as the encoder but in reverse, which give the network a shape like the letter U, hence the name, the skip connections are used after every pooling layer except the last one. Figure 3.14 shows the architecture of the proposed UNET by Krizhevsky et. al (2015) [26].

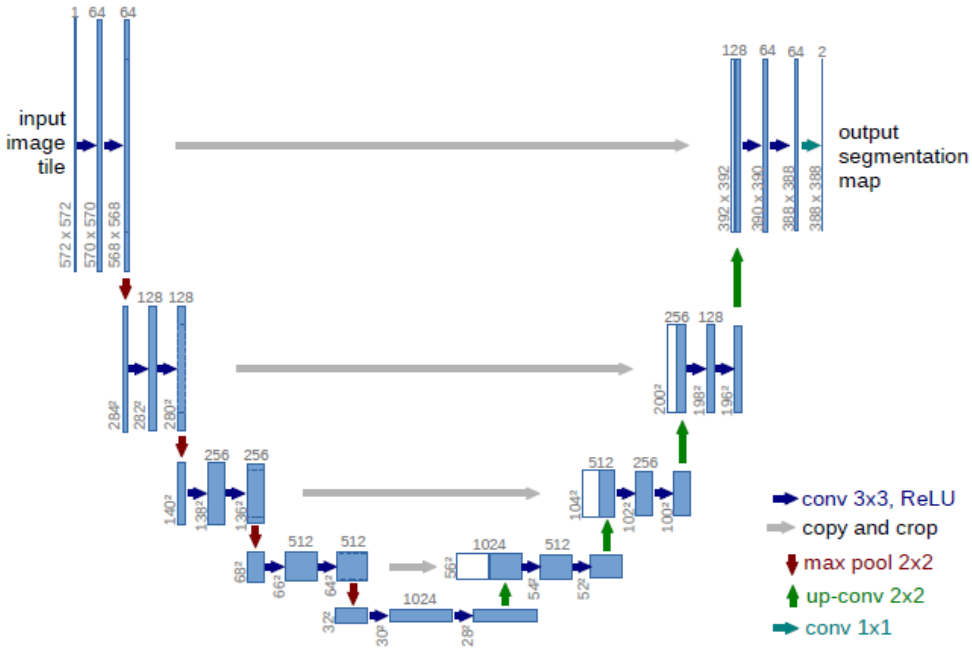

Figure 3.14: U-net architecture (example for 32x32 pixels in the lowest resolution)

This architecture has advantages over FCN, it performs better even with a small training set, and the skip connections help to further localize the higher features.

3.5.7 Object detection algorithm

You Only Look Once (YOLO) algorithm proposes the use of an end-to-end neural network that makes predictions of bounding boxes and class probabilities all at once. YOLO algorithm employs CNN to detect objects in real-time. As the name suggests, the algorithm requires only a single forward propagation through a neural network to detect objects, which makes it faster than other algorithms. YOLO has three important features most important is speed, this algorithm improves the speed of detection because it can predict objects in real time, high accuracy where YOLO is a predictive technique that provides accurate results with minimal background errors, and learning capabilities because The algorithm has excellent learning capabilities that enable it to learn the representations of objects and apply them in object detection.

Working principle

YOLO algorithm aims to predict four different values about the class:

- Coordinates of the center of the bounding box (b_x, by).
- High of the bounding box.

- Width of the bounding box.
- Class of the Object c.
- The probability of the prediction P_C.

Because of this, YOLO expects the data to be prepared during training in the shape of images for the input and XML files that contain the required information about the bounding boxes in that image, where the coordination should define the center coordinates, high and width, and the class of that bounding box, all these values should be scaled to be between 0 and 1 considering the upper left corner as the origin, as shown in Figure 3.15.

Figure 3.15: YOLO bounding box annotation system.

First, the image is divided into various grids, each grid has a dimension of SxS.The algorithm need to generate a vector for each cell in the form described in 3.17.

$$C_{n,m} = (P_C^1, B_x^1, B_y^1, B_w^1, B_h^1, P_C^2, B_x^2, B_y^2, B_w^2, B_h^2, C_1, C_2) \tag{3.17}$$

Where C_1 and C_2 are binary values referes to the class of the image, the vector can have as many C_n values as the classes in the dataset. Training this architecture requires a loss function, which we define as if the cell i predicts class probabilities $\hat{p}_i(\text{aeroplane}), \hat{p}_i(\text{bicycle})...$ and the bounding box $\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i$ then our full loss function for an example is given by equation 3.18:

$$\sum_{i=0}^{n} \left(\lambda \mathbb{1}_i^{\text{obj}} ((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2) + \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \right) \tag{3.18}$$

Where $\mathbb{1}_i^{\text{obj}}$ encodes whether any object appears in cell i. Note that if there is no object in a cell the algorithm does not consider any loss from the bounding box coordinates predicted by that cell. In this case, there is no ground truth bounding box so we only penalize the associated probabilities with that region. One issue that

might happen is when the algorithm predicts several bounding boxes for one class, a solution could be selecting only one box per class, that has the highest probability, but what if there are more objects of one class on the image, because of that, YOLO uses a non-max suppression algorithm, where it takes the boxes with the maximum probabilities, and compares the box with all other boxes of that particular class using IoU, if the IoU is higher than the predefined threshold (for example 0.5), then the box with a smaller probability is suppressed or excluded. It means that two boxes with high IoU values probably indicate the same object on the image, so it excludes the box with a lower probability. This process is repeated until all boxes are taken as object prediction or excluded. The CNN used in the original YOLO is shown in Figure 3.16

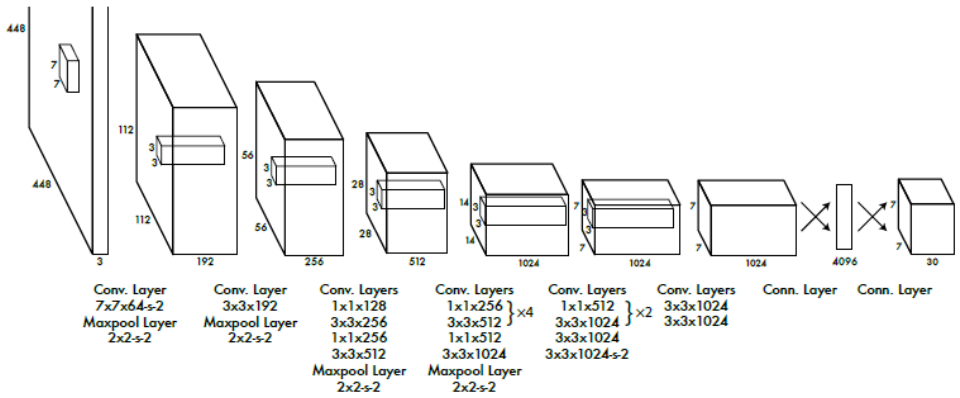

Figure 3.16: YOLO architecture.

This was the first version of YOLO presented by Nekrasov et. al (2016) [28] known as YOLOv1, There were many improvements over the years on the algorithm, in YOLOv2 for example instead of making arbitrary guesses on the boundary boxes, in YOLO v2 authors defined 5 anchor boxes with predefined width and height. To identify the most appropriate dimension of the boxes, k-means clustering is used on the dimensions of bounding boxes from the data set, with distance metric based on IoU [28], the later versions kept improving the accuracy and the performance.

Chapter 4

Literature review

4.1 Introduction

Researchers have been trying to perform semantic segmentation on echocardiography images, whether using conventional image processing methods, like edge detection or predefined contours or using deep learning techniques like CNN.

4.2 Method

PubMed search engine was used to carry out the search for related research, the keywords for this task were: "Semantic Segmentation", "2D Echocardiography", and "deep learning" we also searched for "Ejection fraction estimation" with "Deep Learning" or "Convectional Neural Networks" as these researches also would include segmentation procedure mainly for the Left ventricle, we excluded researchers that have been done on the fetal heart during pregnancy, as they were dealing with a different problem, finally, although we excluded researches that were done on cardiac images from sources other than ultrasound, we included studies that used multi-images-sources as long as it includes Echocardiography.

4.3 Results

The way of searching described above gave us 10 researches that have been published between 2019 and 2023.

4.3.1 Leclerc et al. (2019)

This paper [29] was to introduce the Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) dataset. Although the introduction of the dataset was the main purpose, they discussed the problem using deep learning, and compare multiple methods, they only segmented the Left Ventricular endocardium and Left Ventricular myocardium using all but poor image quality. However, they showed how encoder-decoder-based architectures outperform state-of-the-art non-deep learning methods their work paved the way for future research to be done using the same dataset.

4.3.2 Moradi et al. (2019)

The work of Moradi et al. [30] is one of the earliest works that used UNET to segment the left ventricle which we could find, it modified the UNET architecture to improve its performance, the proposed architecture - named MFP-Unet) adds extra convolution layers for extracting feature maps from all levels of the expansion path in order to be included in the segmentation process in the last layer. This inclusion is promised by a feature pyramid network. While they got better results than the work of Leclerc et al. [29]. However, the main limitation was the limited data used for training, also they only segmented the LV. Figure 4.1 describes the proposed MFP-Unet architecture.

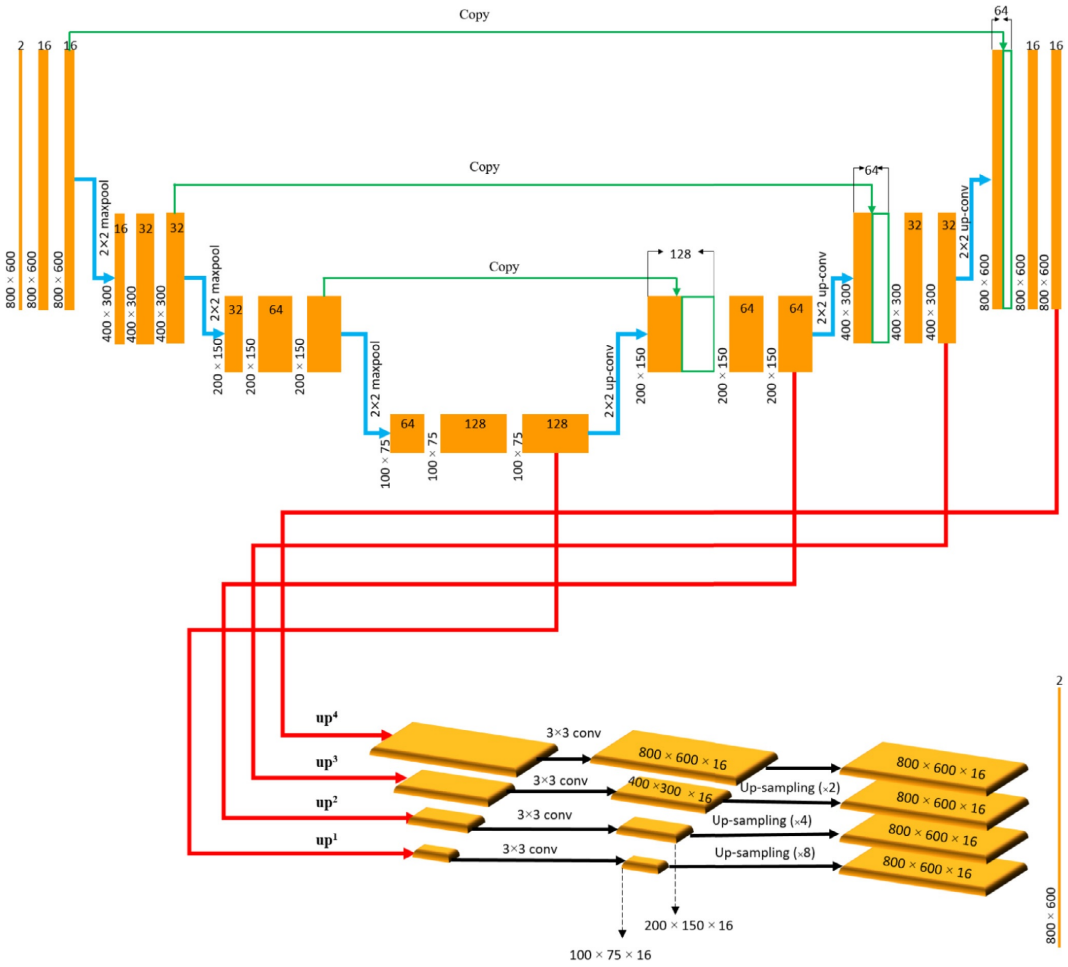

Figure 4.1: MFP-Unet architecture used by Moradi et al.(2019)

4.3.3 Kim et al. (2021):

The work of Kim et al. [31] aimed to segment the Left Ventricle endocardium and Left Ventricle myocardium regions from porcine images, and then using learning transfer techniques they tested it on the human dataset, this study uses six different views, both Apical views, and base views, and used post-processing on the output

of the Neural Network which seemed to improve the metrics, but upon further inspection using statistical analyze it turns out this was not statistically significant (P>0.05). The main limitation of this study was using images which were obtained from open-chest pigs, which has better quality than human echo, using transfer learning may prove that the concept works on human but only two views (Apical 4CH and Apical 2CH) were available to test. Figure 4.2 describes the proposed neural network in this study.

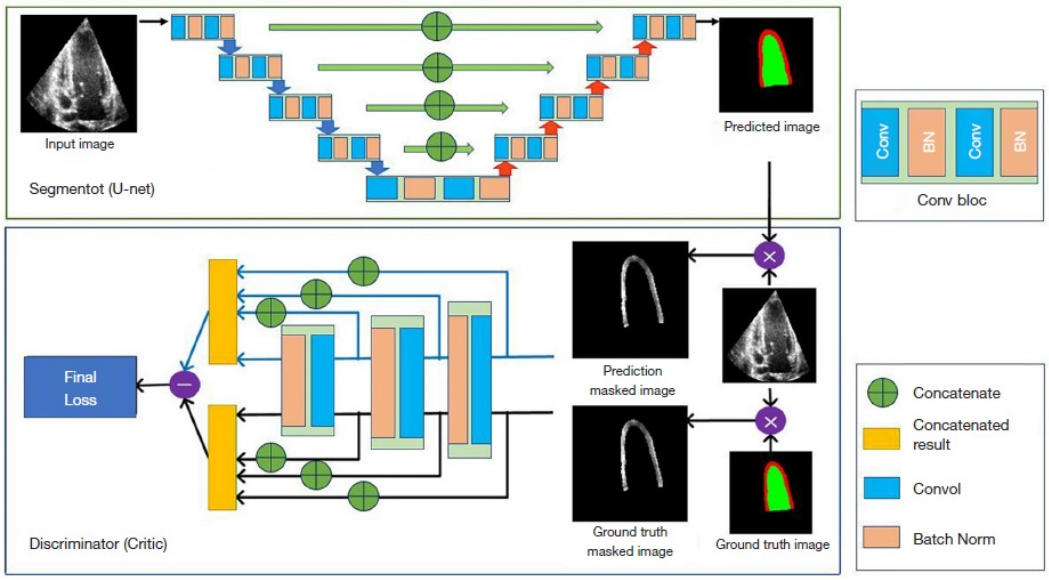

Figure 4.2: segAN architecture used by Kim et al.(2021)

4.3.4 Zhuang et al. (2021):

YOLOv3 (You only look once) algorithm was used by Zhuang et al. [32] to detect three points of the left ventricle -Apex and bottom (septal wall base and lateral wall base) - and the overall bounding box of the left ventricle, then using iterated conditional model (ICM) they performed initial segmentation of the Myocardium, using the three points location as a restraining condition Myocardium's left and right part are located, finally using B-spline method to smooth the edges and morphological filter to reduce speckle noise. This research combines the CNN methods to detect the LV region and its points and other image processing methods, which are promising methods, especially the YOLO algorithm is known to be fast and convenient for real-time detection. In terms of accuracy indices (*i.e.*, SDC, MAD and HD) this model did not achieve better results than other research, and this paper does not give detail about the dataset used or the ratio of (training/validation/test). Figure4.3 shows the YOLOv3-based architecture proposed used in this research.

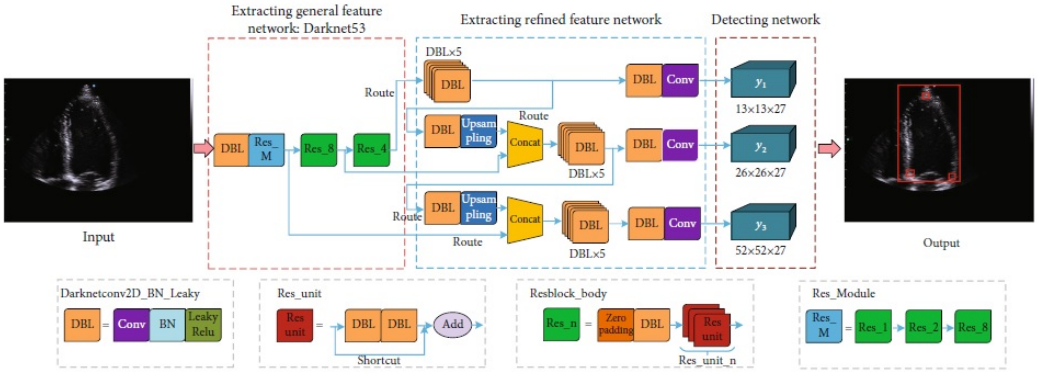

Figure 4.3: proposed architecture used by Zhuang et al.(2021)

4.3.5 Girum et al. (2021)

Girum et al. [33] in this paper came up with the idea of combining modified UNET architecture for forward learning with FCN encoder for a feedback loop, this idea is supposed to improve the high-level feature extraction and allow the system to learn from its mistakes by providing a second chance for the forward system's decoder network to look back on its predicted output. This network is tested on four different datasets, one of the CAMUS dataset, and it actually segmented the heart into three different classes (Not including the background), with good, promising results. Figure4.4 shows the UNET with FCN feedback (which they named LFB-Net) architecture proposed used in this research.

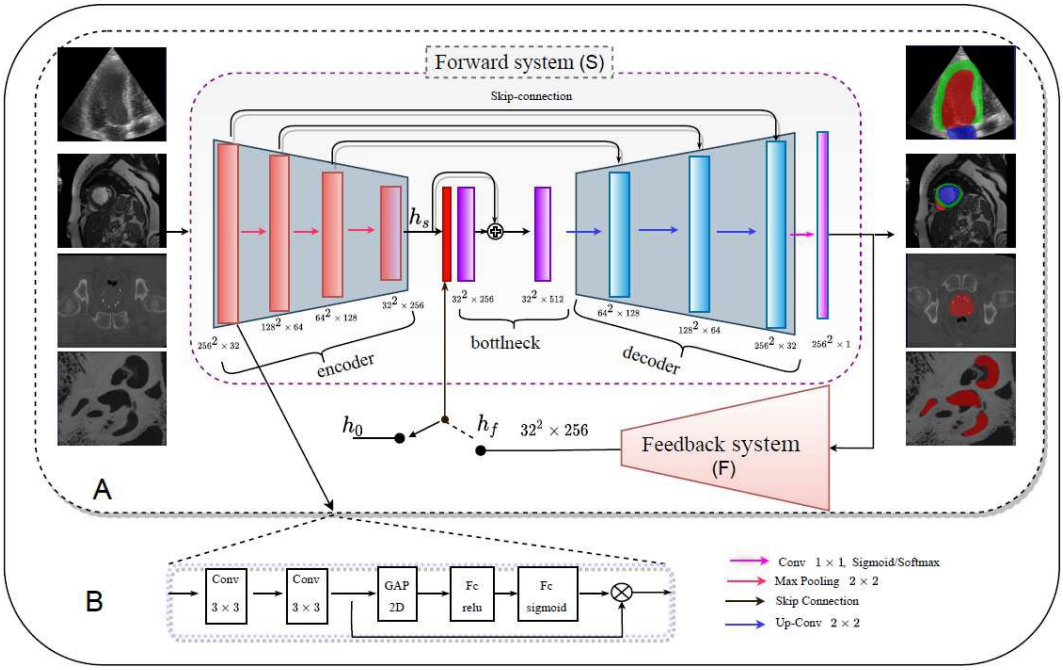

Figure 4.4: LFB-Net architecture used by Girum et al.(2021)

4.3.6 Liu et al. (2021)

The work of Liu et al. [34], addressed two problems, first the low contrast between myocardium tissues and edge dropout, which is related to the fact that 2D echography has low SNR, secondly that the current deep image segmentation technique assigns a prediction to each pixel alone without taking its neighbors' values into consideration, for that they proposed a deep learning model, called deep pyramid local attention neural network, this technique has been rarely used in semantic medical image segmentation models, and it is a little complicated, it uses BiSeNet Bilateral Segmentation Network to extract deep semantic features, then using pyramid local attention algorithm to enhance feature within the compact and sparse neighboring contexts, finally comes the novel Label coherence learning mechanism (LCL), which they claimed to solve the single pixel prediction problem. The author claimed that this method was also helpful in accurately locating the vertical axis, a parameter that is very important to compute the EF using the modified Simpson method. The main problem with this approach is the memory- consumption of the network training is large, which can cause a memory explosion in the training phase. It also only segments the LV endocardium and Left Ventricular myocardium. Figure4.5 shows the PLANet architecture proposed used in this research.

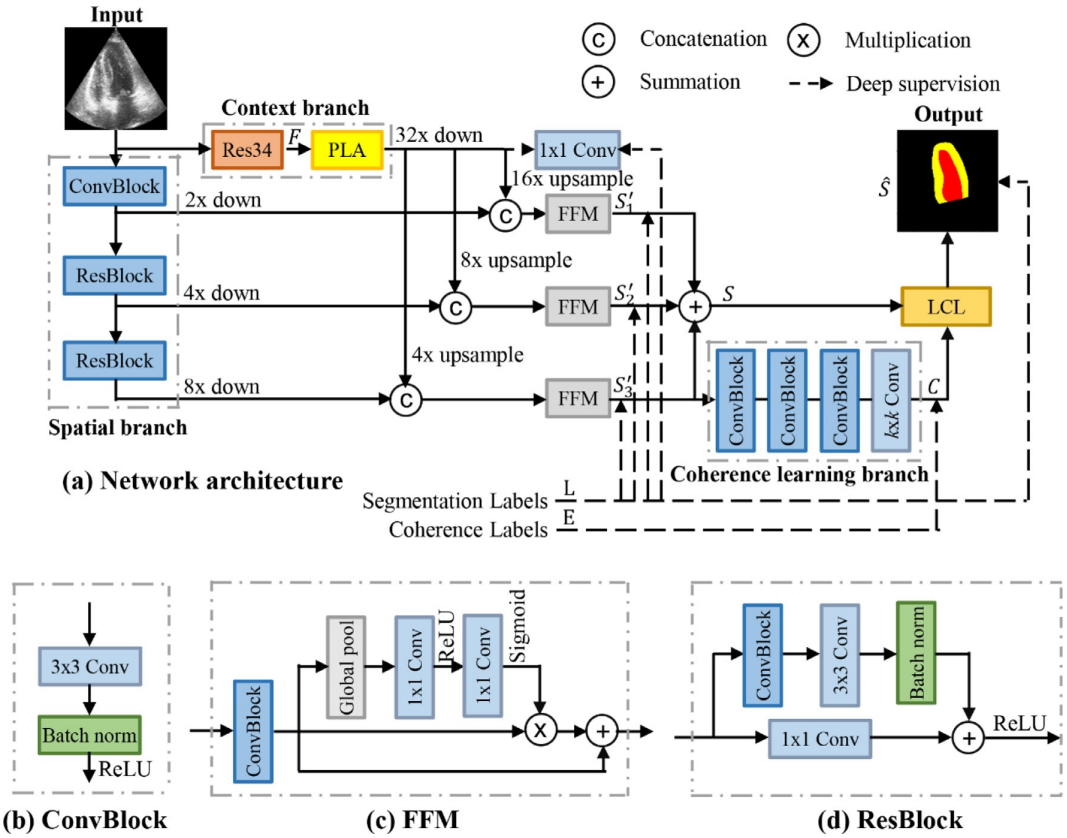

Figure 4.5: PLANet architecture used by Zhuang et al.(2021)

4.3.7 Lei et al. (2021)

Lei et al. [35] proposed a system (named Cardiac-SegNet), consisting of three sub-networks, the first one is a UNET to perform higher feature extraction, the second one is a Fully Convolutional OneState object detector (FCOS), which will segment the image into three ROI, after that the ROI will rescale and used in Mask Head network which will perform the segmentation. This system can segment three regions (Left Ventricle endocardium, Left Atrium, and Left Ventricle myocardium), while this system performs well in the segmentation process, and the idea of using the bonding box and center of mass to improve the segmentation is promising. However, the paper reports that the time of segmentation can be 0.5 seconds.

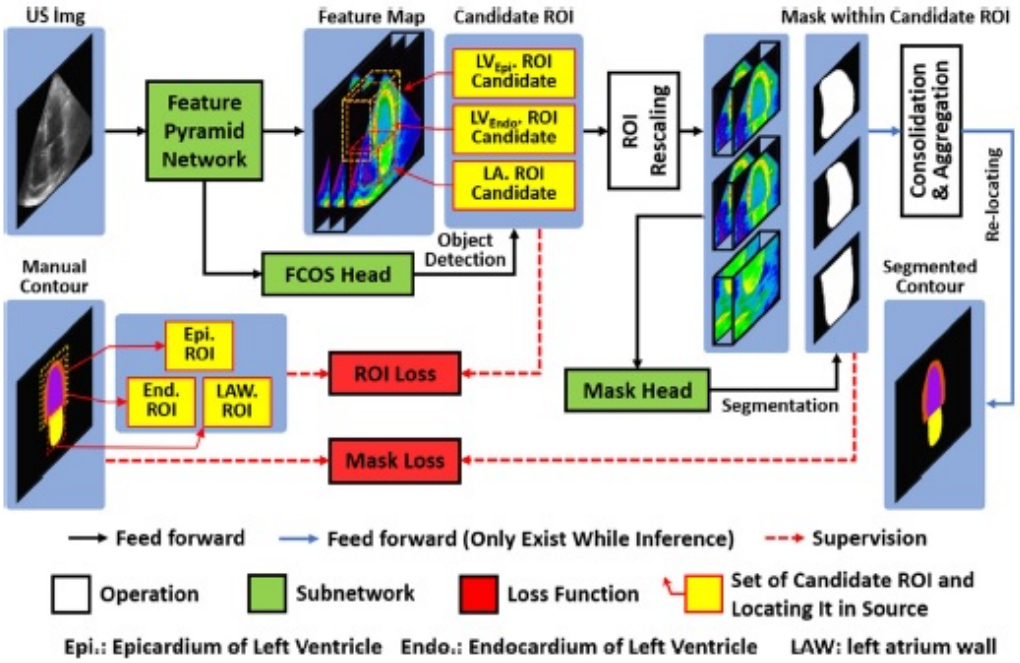

Figure 4.6: UNET architecture used by Alam et al.(2022)

4.3.8 Alam et al. (2022):

Alam et al. [36] proposed a two parallel pipeline for each of the ES frame and ED frame, this procedure used DeepResUNet for segmentation (alongside UNET and ResUNET for comparison), and finally, they used the segmented masks to estimate ejection fraction using Simpson's method, since the main goal was to calculate EF the two pipelines were suggested. The study had a relatively low DSC

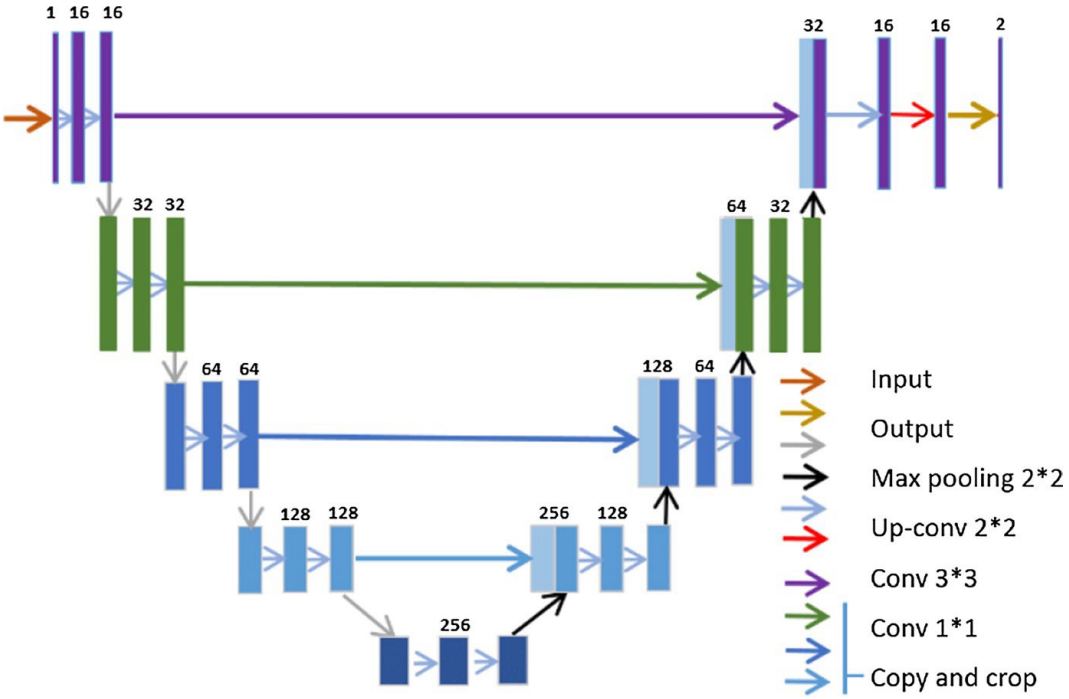

Figure 4.7: UNET architecture used by Alam et al.(2022)

4.3.9 Saeed et al. (2022)

Saeed et al. [37] tried to use a self-supervised algorithm to segment the left ventricle from four apical chambers view, using DeepLapV3 and SimCLR pre-trained backbone, also they tried to use BOYL and U-NET which did not perform better, this research tried to overcome the lack of labeled data by using the self-supervised technique, they do not provide detailed explanation about the loss function used in training the network, and they only segment the LV. Finally, both SimCLR and BYOL are sensitive to batch sizes and require very large batch sizes for optimal performance.

4.3.10 Zeng et al. (2023)

Zeng et al. [38] proposed a system to estimate the Left ventricle ejection fraction from Apical four chambers 2D cardiography, using a modified Simpson's method, this system takes a video as an input and passes it through Multi-attention efficient feature fusion that will give out a binary mask for the left ventricle, the network has an encoder-decoder structure with skip connections, the Network took advantage of the Dual-Attention layers and they introduced a new concept for up-sampling under the name of (pixel shuffling layers), a concept they claimed it overpass the low resolution of the images. The paper achieved good results in terms of DICE coefficients but because of the low resolution of some of the images in the EchoNet dataset, they had to downsample all frames to 112x112 which is a very low resolution. Figure4.8 describes the architecture used by this study.

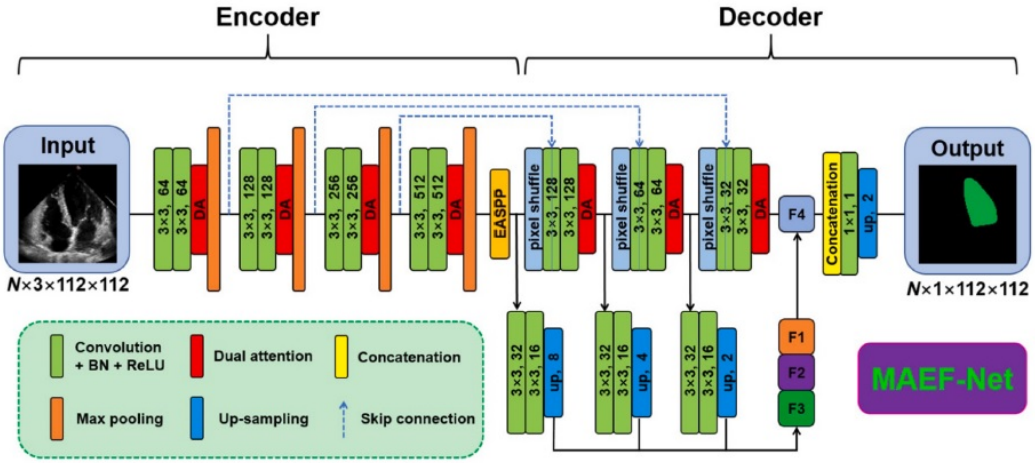

Figure 4.8: The architecture of the MAEF-Net proposed by Zeng et al.(2023):

4.4 Comparison tables and discussion

Clearly, not all studies focused on LA segmentation since the application of LV endocardium included the calculation of EF all studies perform its segmentation, while LV myocardium segmentation was present more than LA but not as many as LV endocardium since it has some application in measuring some features like strain. We provided in this section comparison among the studies we used in our review, Table 4.4 provides a general comparison, where In this Table: A4C: Apical four Chamber view, A2C: Apical two Chamber view, LV endo: Left Ventricle endocardium, LVMayo: left ventricular myocardium, DSC: Dice similarity coefficients, JSC: Jaccard distance, AER: Aera error ration, HD: Hausdorff distance, MAD: Mean absolute distance, CMD: Center of Mass distance, P: Precision, S: Sensitivity. Table 4.1 describes the results of the LV endocardium segmentation, Table 4.2 describes the results of the LV myocardium segmentation, and 4.3 describes the results of the Left atrium segmentation.

Table 4.1: Comparison of LV endocardium segmentation results among studies.

	DSC%	**JSC%**	**HD[mm]**	**MAD[mm]**
Zeng et. al (2023) (Private)	92.81 ± 2.85	-	-	-
Zeng et. al (2023) (EchoNet)	93.10 ± 2.22	87.21 ± 3.85	2.17 ± 1.37	-
Saeed et. al (2022) (EchoNet)	92.52 ± 0.0476]	-	-	-
Saeed et. al (2022) (CAMUS)	93.11 ± 0.0424	-	-	-
Alam et. al (2022) (ES)	82.1 ± 0.8	66.9 ± 6.4	23.8 ± 0.1	-
Alam et. al (2022) (ED)	86.5 ± 1.1	63.7 ± 9.6	19.7 ± 0.2	
Liu et. al (2021) (EchoNet) (ES)	91.8 ± 3.4		5.4 ± 2.6	1.6 ± 0.7
Liu et. al (2021) (EchoNet) (ED)	94.2 ± 2.1		5.0 ± 2.2	1.4 ± 0.6
Liu et. al (2021) (CAMUS) (ES)	93.1 ± 3.2	-	4.3 ± 1.5	1.4 ± 0.6
Liu et. al (2021) (CAMUS)(ED)	95.1 ± 1.8		4.2 ± 1.4	1.3 ± 0.5
Girum et. al (2021) (2CH)	94.00 ± 3.0	-	5.6 ± 3.22	-
Girum et. al (2021) (4CH)	94.00 ± 3.0	-	5.0 ± 2.83	-
Lei et. al (2021) (ES)	92.7 ± 4.3	-	2.247 ± 2.274	1.893 ± 1.785
Lei et. al (2021) (ED)	94.8 ± 2.4		2.288 ± 1.784	1.887 ± 1.530
Kim et. al (2021)	91.7 ± 0.071	-	5.14 ± 1.71	-
Leclerc et. al (2019) (ES)	91.6 ± 6.1	-	5.5 ± 3.8	1.6 ± 1.6
Leclerc et. al (2019) (ED)	93.9 ± 4.3	-	5.3 ± 3.6	1.6 ± 1.3
Moradi et. al (2019) (CAMUS)	95.3 ± 1.9	-	3.49 ± 0.95	1.32 ± 0.53
Moradi et. al (2019) (costume)	94.5 ± 1.2	98.0 ± 1.0	1.62 ± 0.05	1.12 ± 0.11

Table 4.2: Comparison of LV myocardium segmentation results among studies.

	DSC%	HD[mm]	MAD[mm]
Liu et al. (2021) (ES)	95.6 ± 1.4	4.6 ± 1.4	1.6 ± 0.6
Liu et al. (2021) (ED)	96.2 ± 1.2	4.6 ± 1.5	1.5 ± 0.5
Girum et al. (2021) (2CH)	88.0 ± 4.0	7.1 ± 3.86	-
Girum et al. (2021) (4CH)	86.0 ± 6.0	6.7±3.04	-
Zhuang et al. (2021)	93.5 ± 1.9	6:68 ± 1:78	2:57 ± 0:89
Lei et al. (2021) (EchoNet) (ES)	94.3 ± 1.9	5.5 ± 2.1	1.8 ± 0.6
Lei et al. (2021) (EchoNet) (ED)	95.1 ± 1.7	5.5 ± 2.1	1.7 ± 0.7
Lei et al. (2021) (CAMUS) (ES)	95.3 ± 2.2	2.755 ± 2.157	2.746 ± 2.329
Lei et al. (2021) (CAMUS) (ED)	96.0 ± 1.6	2.946 ± 2.125	2.369 ± 2.029
Kim et al. (2021)	85.9 ± 6.4	6.18 ± 1.17	1.9 ± 1.2
Leclerc et al. (2019) (ES)	94.5 ± 3.9	6.1 ± 4.6	-
Leclerc et al. (2019) (ED)	95.4 ± 2.3	6.0 ± 3.4	-

Table 4.3: Comparison of LA segmentation results among studies.

	DSC%	HD[mm]	MAD[mm]	CMD[mm]
Girum et. al 2021	92.0 ± 4.0	5.2 ± 3.48	-	-
Lei et. al 2021 (ES)	92.2 ± 5.5	2.65 ± 3.453	1.703 ± 1.677	1.352 ± 1.639
Lei et. al 2021 (ED)	89.5 ± 8.5	2.214 ± 4.107	1.696 ± 1.750	1.645 ± 2.148

Table 4.4: Comparison among different studies

Title	Year	Dataset	Number of Patients in Datasets	Number of images or videos	Dataset Split	Used View	Type of Data	Size of frame used in model	Segmented Region	Artifitial Neural Network used	Evaluation metrics for Segmentation	Loss Function	Other tasks performed by the system
MAEF-Net: Multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography	2023	- EchoNet-Dynamic - Private clinical dataset	10,019//22	10,019 video 2129 frames	- Train = 7465, Validation = 1288, Test = 1226 (video)- Test = 2129 (frame)	A4C	Video	112 x 112	LV	Multi-attention efficient feature fusion network(MAEF-Net)	DSC% JSC AER HD	Combination between cross-entropy loss and soft Dice loss	- ES and ED frames detection. - Ejection fraction estimation
Contrastive Pretraining for Echocardiography Segmentation with Limited Data	2022	- EchoNet-Dynamic - CAMUS	10,024//400	20,048 frames800 frames	- Train = 14920, Valiidation = 2576, Test = 2552 (frame)- Train = 600, Validation = 100, Test = 100 (frame)	A4C	Images	224 x 224	LV	DeepLabV3with a SimCLR pretrained backbone	DSC	-Not mentioned	- None
Ejection Fraction estimation using deep semanticsegmentation neural network	2022	- costume dataset	380	380 frames	-Train = 308*2, Test = 72*2 (frame)	A4C	Images	256 x 256	LV	Deep ResUNet	DSC HD IoU	-Not mentioned	- Ejection fraction estimation
Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography	2021	- sub-EchoNet-Dynamic -Dynamic- CAMUS	2500// 500	5000 frames2000 frames	- Train = 1600, Validation = 400, Test = 500 (patients)- Test = 2000 (frame)	A4CA2C	Images	512 x 512	- LV endo- LV MYO,	Deep pyramid local attention neural network(PLANet)	DSC HD MAD	combination of cross-entropy loss functions and binary cross-entropy loss function	- Real Time segmentation
Learning With Context Feedback Loop forRobust Medical Image Segmentation	2021	- CAMUS	450	1800 frames	- Train = 396, Validation = 54, (patients)	A4CA2C	Images	256 x 256	- LV endo- LV MYO- LA	Learning with context FeedBack system (LFB-Net)	DSC HD	average of binary crossentropyand Dice coefficient loss functions as:	- Prostate segmentation in radiotherapy- Inner ear segmentation- Cardiac cine-MRI segmentation
Automatic Segmentation of Left Ventricle in Echocardiography Based on YOLOv3 Model to Achieve Constraint and Positioning	2021	Costume dataset	-	-	-	-	video	416x416 *	LV MYO,	YOLOv3 (Darknet53)	FPS DSC MAD HD	- YOLOv3 loss function, sum squared error for: - location of the centroid of the anchor- boxwidth and height of the anchor - boxconfidence score of whether there is an object or not-classification loss	- Real time segmentation
Echocardiographic image multi-structure segmentation using Cardiac-SegNet	2021	CAMUS	450	1800 images	Cross validation with 5 flods,	- A4C- A2C	Images	640x960	- LV endo- LV MYO- LA	Cardiac-SegNet(It has three substages: 1. Res2D2. FCOS3. deep attentionU-Net	DSC HD MAD CMD]	(for first subnetwork):'- binary cross entropy loss(for second subnetwork):- Sigmoid cross entropy loss for Center of mass (COM) location Sigmoid cross-entropy loss for the class of the COM pixel-level intersection-over-union (IoU) loss for bounding box	- None
Automatic segmentation of the left ventricle in echocardiographic images using convolutional neural networks	2021	- in-house dataset - CAMUS	6 to 8 different pigs//450 human patients	1,649 porcine images1800 human images	-	- A3C- A4C- A2C- short-axis: base- short-axis: mid- short-axis: apex.	Images	128×128	- LV endo- LV MYO,	SegAN	DSC HD P S	multi-scale feature loss function based on the mean absolute error (MAE) and dice loss	- 3D reconstruction of LV - Ejection Fraction, Strock Volume and Heart rate estimation.
Deep Learning for Segmentation using an OpenLarge-Scale Dataset in 2D Echocardiography	2019	- CAMUS	406	1624 frames	Cross-validation with 10 folds,	- A4C- A2C	Images	256x256	- LV endo- LV MYO,	U-Net	DSC MAD HD	Cross-entropy + weight decay	- None
MFP-Unet: A novel deep Learning based approach for left ventricle segmentation in echocardiography	2019	- CAMUS- Prepared custom dataset	500//137	-	5-fold cross-validation	- A4C	Images	800 x 600	LV	MFP-Unet	DSC JSC MAD HD	- Not mentioned	- None

Chapter 5

Multi-structure semantic segmentation of Echocardiography images

5.1 Introduction

The human heart is a complex multi-structured organ with a main pumping function and the US is considered a powerful tool that can image these structures. In order to make important measurements that can quantify the heart functions and early diagnosis of heart failure, semantic segmentation must be performed. This segmentation can be done manually by an expert sonographer. However, many studies have been done to automate this process, using conventional or deep learning methods.

5.2 Materials and methodology

5.2.1 Dataset

The dataset which was used in this research is part of the "Cardiac acquisitions for multi-structure ultrasound segmentation" (or CAMUS) challenge. The data was collected from 500 patients, using GE Vivid E95 ultrasound scanners (GE Vingmed Ultrasound, Horten Norway), with a GE M5S probe (GE Healthcare, US) at the University Hospital of St.Etienne in France. In order to make the data realistic regarding clinical cases, no data selection was performed, some cases did not give high-quality images and yet were included as part of the dataset. For each patient, there are two sequences taken from the apical window each showing one heart cycle, one is four chambers view (4CH), and the other is two chambers view (2CH), within the full sequence, two frames were specified from each sequence, one showed end-systolic (ES) frame while the other showed end-diastolic (ED) frames, these frames were chosen following the recommendations of American society of echocardiography and the European association of cardiovascular imaging [39], where the ED frame is the frame appearing immediately after the mitral valve is closed, or the frame in which the left ventricle has the largest volume, and the ES frame is the frame appearing immediately after the aortic valve is closed, or the frame in which the left ventricle has the smallest volume, due the lack of reliable electrocardiography

data the volume method were chosen. The data contained segmented masks for 450 patients, and the authors chooses to not publish the mask for the remaining 50 patients as a part of the challenge, the masks were only provided for the end-systolic and end-diastolic frames, which gave each patient four annotated frames. The masks were generated manually by three experts, they were asked to segment: The LV endocardium, the LV epicardium, and the LA. Figure 5.1 shows images from the dataset with their corresponding masks, the black regions in the mask are considered the background (this will include the heart's other structures like the right ventricle, right atrium, ... etc), and white is the left atrium, light gray is the left ventricle epicardium, and dark gray is The Left ventricle endocardium.

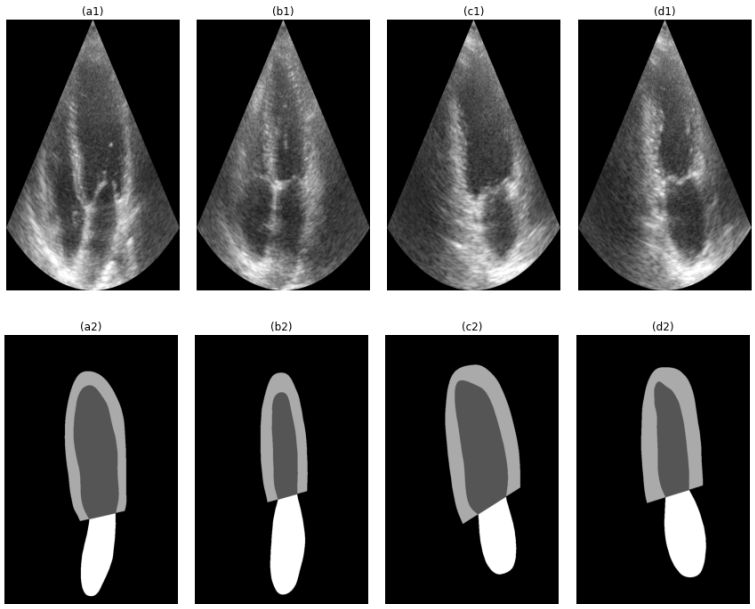

Figure 5.1: - first-row Ultrasound image, second row the masks, (a) four chambers at end-systole (b) four chambers at end-diastole (c) two chambers at end-systole (d) two chambers at end diastole

Images have different sizes in the dataset ranging from $584x354$ up to $1945x1181$. The images are saved with "mhd" and "raw" extensions, and the mask is provided as an image with four possible values for each pixel, as the following: 0 indicates that the pixel is background, 1 indicates that the pixel belongs to LV endocardium, 2 indicates that the pixel belongs to LV epicardium, and 3 indicates that the pixel is from LA. The data is available online[1].

5.2.2 Proposed model

We proposed a deep learning model to perform semantic segmentation to extract The LV endocardium, the LV epicardium, and the LA regions from two-dimensional

[1]https://www.creatis.insa-lyon.fr/Challenge/camus/databases.html

ultrasound images, Apical two and four chambers views, this model consists of two steps:

1. Extract the region of interest (ROI): Using YOLOv7 [40] algorithm to perform extraction of the regions of the left chambers. YOLOv7 is the latest version of the YOLO algorithm at the time this research was done. It has better accuracy than the older version. However, as it uses more floating points operation it is a little slower.

2. Performing Semantic segmentation: Using UNET architecture, this step used the output of the first step as its input, as this step will have to perform semantic segmentation on the ROI detected by the last step. Figure 5.2 shows the proposed unit for this research. It consists of five stages encoder, and a similar five stages decoder, each step in the encoder path has two convolutional layers, with the "MISH" function as their activation function, one batch normalization layer, and max pooling layer, with the final stage having a dropout layer with the rate of 0.5 to prevent overfitting. The decoder used transpose layers for upsampling (or deconvolution), with the final layer having softmax as an activation function since this is a multi-class problem.

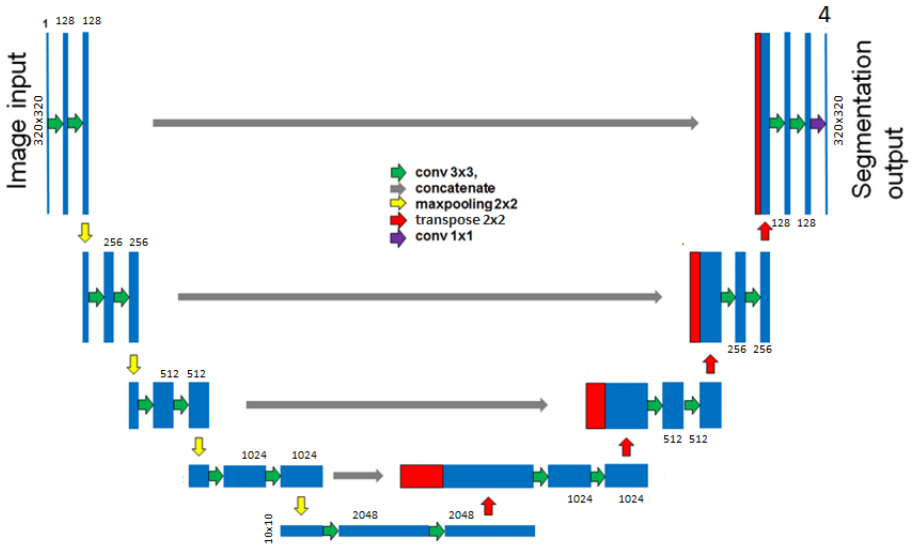

Figure 5.2: UNET proposed in this research

After each max pooling layer, the feature size will go down by half, which means the size of the high feature array will go to 2^5 its original size. The output layer will have four filters, this is known as one hot encoding, each layer is a binary array representing the probability of each pixel to be from this class. We need to perform argmax operation on each pixel to determine its class.

Figure 5.3 shows the flow chart of the proposed system.

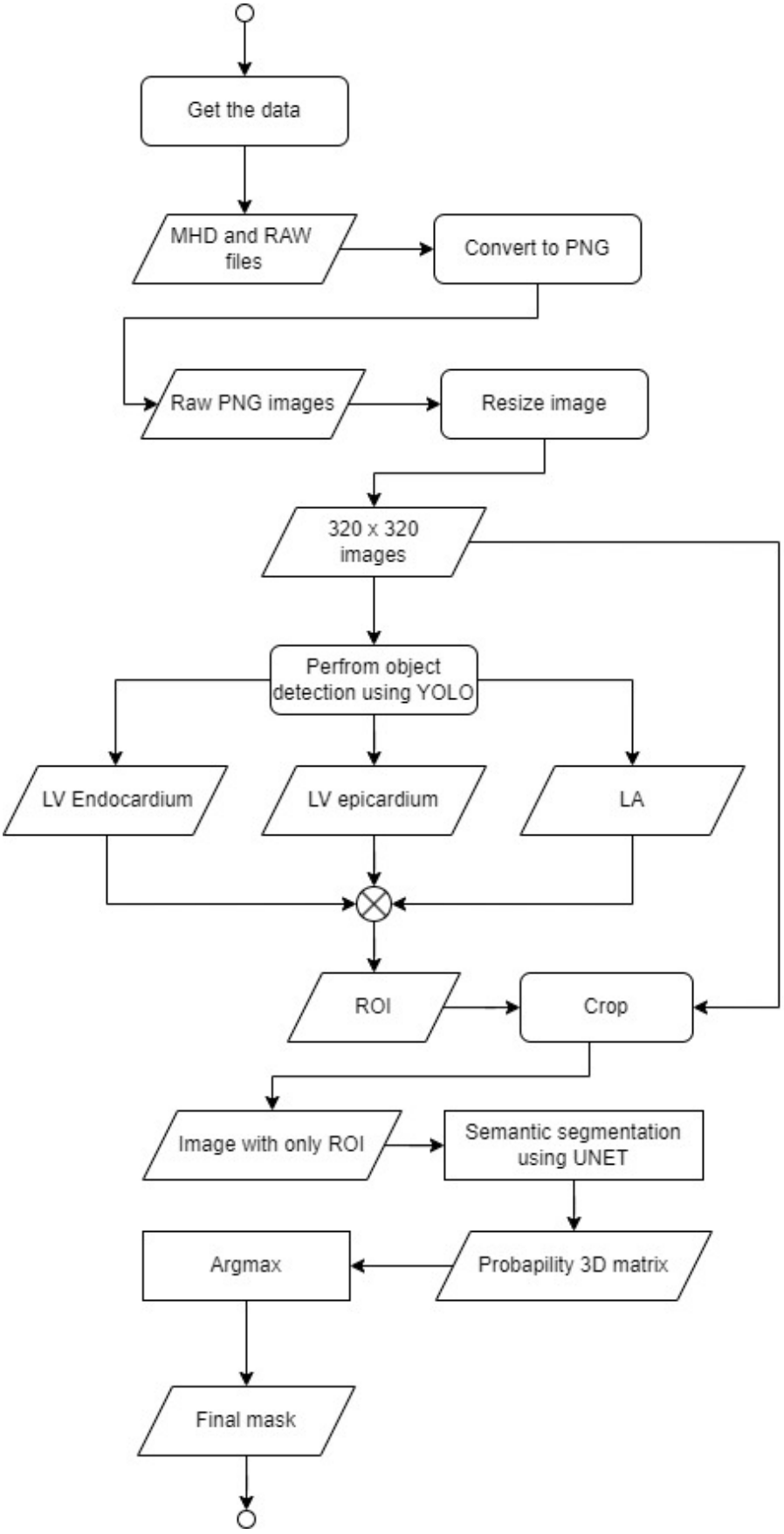

Figure 5.3: Proposed system flow chart.

5.2.3 Training strategy

We implemented the thesis in a Python environment for all steps.

Preprocessing

The dataset as mentioned before is provided using "MHD" and "RAW" extensions, which are common extensions in the medical field. However, they are not well compatible with python packages like Tensorflow, hence we converted these images to PNG, using the SimpleITK python package.
The PNG images had different sizes which would cause a problem in building patches for training, for that we had to resize the images.
The chosen size was 320x320 pixels, two factors supported this choice, one is that the smallest dimension in the dataset was 354, and since this number is not divisible by 32, a crucial point in our encoder-decoder architecture, the closest factor was chosen *i.e.* 320.
Resizing was carried out using the OpenCV python package, and we made sure to use the suitable interpolation for the resizing, in order to make sure that no values besides the allowed one appear in the masks (0,1,2 and 3).
Although the masks were made available no mask bounding boxes data were available, and training of YOLO requires these data, we created the corresponding these boxes by following a simple algorithm; in which we scanned the images from the top row first, and then from the bottom row going up searching in both cases for the first instance of each class, saving the corresponding row number, then repeating the procedure with the columns, right to left the left to right, till we got the four coordinators, which then we used to get the center coordinators according to equation 5.1, were X_{center}, Y_{center} represent the row and column of center, and the height and width of the box using equation 5.2

$$X_{center} = \frac{row_{top} + row_{botom}}{2}; Y_{center} = \frac{column_{left} + column_{right}}{2} \tag{5.1}$$

$$X_{center} = row_{top} - row_{botom}; Y_{center} = column_{left} - column_{right} \tag{5.2}$$

These values were scaled to be between 1 and 0 and used to create the XML files.
Finally, the dataset were divided into three sets: the training set which contained 60% of the data (1080 frames), these frames were used to train the system, the validation set which contained 10% of the data (180 frames) for the sole purpose of determining when to stop training to avoid overfitting the training set, and the test set containing 30% of the data (540 frames) this data were kept away from the system during training and were used only to evaluate the system, as shown in Figure 5.4

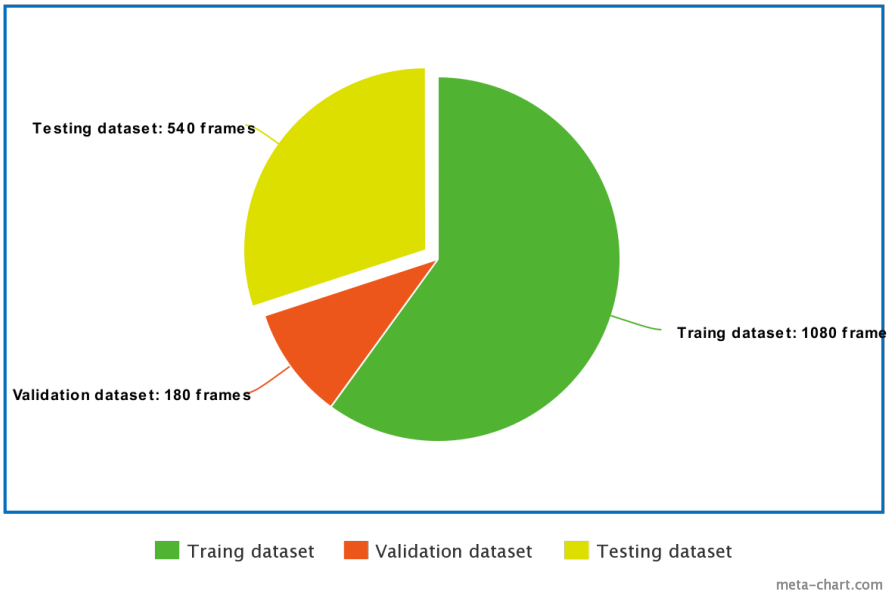

Figure 5.4: Splitting the dataset.

Loss function

The loss function to train the UNET is given by equation 5.3

$$L_{total} = \frac{1}{2} \times (L_{entropy} + L_{Dice}) \tag{5.3}$$

Where $L_{entropy}$ and L_{Dice} are given using equation 5.4 and 5.5 respectively:

$$L_{entropy} = -\sum_{k=1}^{c}\sum_{i=1}^{I} log\left[\frac{e^{s(k,i)}}{\sum_i s(k,i)}\right] \tag{5.4}$$

$$L_{Dice} = 1 - \frac{1}{\sum_k \alpha_k k}\left[\sum_k \alpha_k k \frac{2 \times \sum_{i \in I} u_i^k \mu_i^k}{\sum_{i \in I} u_i^k + \sum_{i \in I} \mu_i^k}\right] \tag{5.5}$$

Where c: is a number of classes (3 in our case ignoring the background), I: is the number of pixels in each image, $s(k,i)$ is the probabilistic feature maps at a pixel $i \in I$ belonging to the pixel class k. α_k: A unique weight given to each class.

Cloud computational

Training a model requires a huge amount of computational power, specifically a powerful graphic process unit (GPU), there are some services that allow users to run their codes on clouds server, Google provides one of these services under the name google collab, which provides a Jupyter notebook to run python codes. The paid version can provide: 32GB of Random access memory (RAM), and NVIDIA Tesla P100 or T4 GPU, adding the fact that it can access cloud google cloud storage service (Google Drive), which made this service very suitable for this application. In

the collab environment, we cloned the YOLOv7 files from [2] which is built using the PyTorch package, while the UNET we built using the TensorFlow package.

5.2.4 Evaluation Metric

The metric which was used to evaluate our model: are the dice similarity coefficient described in equation 5.6, Jaccard's similarity coefficient (JSC) provided by equation 5.7, and Hausdorff distance (HD) defined in equation 5.8

$$DSC = \frac{2 * TruePositive}{2 * TruePositive + FalsePositive + TrueNegative} \tag{5.6}$$

$$JSC = \frac{|A \cap B|}{|A \cup B|} \tag{5.7}$$

$$HD_{95th} = max\{K^{th}_{s \in S} \min_{g \in G} \parallel S - L \parallel, K^{th}_{s \in S} \min_{g \in G} \parallel L - S \parallel\} \tag{5.8}$$

5.3 Results

The first step was trained for 100 epochs, with an early stop at epoch number 80, while the second step was trained for 50 epochs and early stops at epoch number 42. Figure 5.5 shows the output of the YOLO algorithm using a four chambers image, while Figure 5.6 shows the output of the YOLO algorithm using a two chambers image, each bounding box has a value representing the confidence factor.

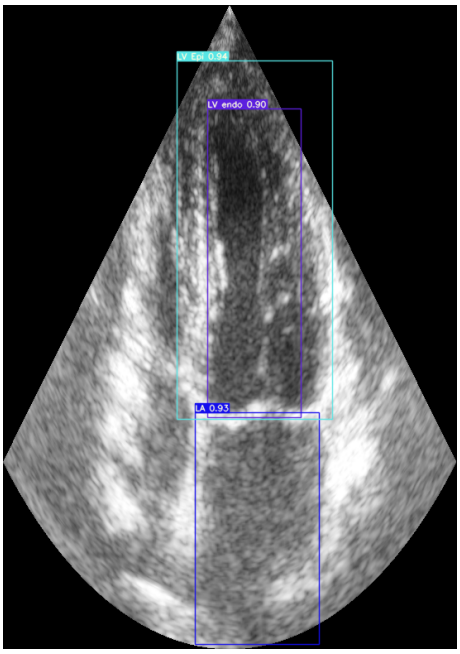

Figure 5.5: YOLO output, four chambers view.

[2] https://github.com/WongKinYiu/yolov7

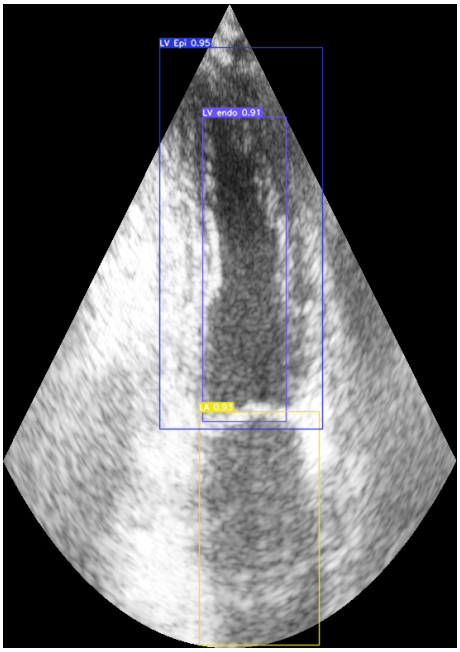

Figure 5.6: YOLO output, two chambers view.

Tables 5.1 details the DSC results for each class in each data set, while tables 5.2 shows the JSC, and table 5.3 provide the HD.

Table 5.1: DSC% results

	Training set	Validation set	Testing set
LV endocardium	93.00 ± 3.90	92.00 ± 3.90	91.00 ± 6.40
LV epicardium	88.00 ± 4.20	86.00 ± 4.30	85.00 ± 7.60
LA	91.00 ± 5.60	89.00 ± 9.50	88.00 ± 11.8

Table 5.2: JSC results

	Training set	Validation set	Testing set
LV endocardium	87.00 ± 6.30	86.00 ± 7.20	86.00 ± 7.30
LV epicardium	78.00 ± 6.40	77.00 ± 7.60	86.00 ± 7.30
LA	85.00 ± 7.70	83.00 ± 10.50	83.00 ± 10.30

Table 5.3: HD $results_{pixels}$

	Training set	Validation set	Testing set
LV endocardium	3.68 ± 0.94	3.87 ± 0.99	3.88 ± 0.95
LV epicardium	4.68 ± 0.78	4.86 ± 0.76	4.96 ± 0.86
LA	3.77 ± 0.83	4.06 ± 1.03	4.00 ± 0.98

Figure 5.7 shows an A2C view with its ground truth mask and our predicted mask,

while Figure 5.8 shows an A4C view.

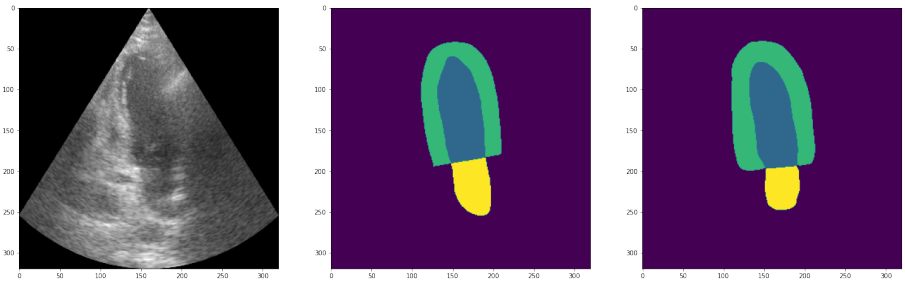

Figure 5.7: Two chamber view with ground truth and predicted mask - From left to right, input image, ground truth mask, predicted mask

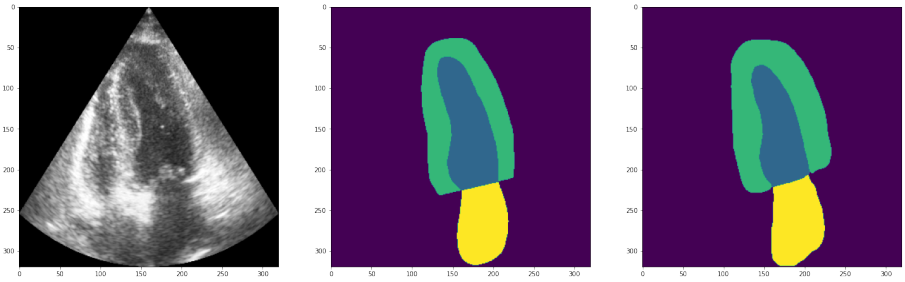

Figure 5.8: Four chamber view with ground truth and predicted mask - From left to right, input image, ground truth mask, predicted mask

5.4 Discussion

We proposed in this thesis a fully automated system that can segment structures from echocardiography images, in order to be used later in the evaluation of heart function and in diagnosing heart malfunctions, specifically the system was able to segment LV endocardium, LV epicardium, and LA from A4C and A2C images.
The proposed system depended on deep learning and CNN to perform the semantic segmentation, using two steps, first by performing object detection to extract the region of interest (ROI) using YOLOv7 algorithm, this is done by combining the three bounding boxes around the needed regions, and then semantic segmentation is performed on the ROI using U-net, and the system was trained and tested on the CAMUS dataset.
The combination between object detection and semantic segmentation allowed the system to perform better even when training the system using both views at the same time, *i.e.* there was no need for the training of two separated systems for each view, also unlike conventional image processing methods, our system could work on images with low resolution, as the images in the dataset had resolution categorize into (good, medium, and bad) and no image was excluded from this dataset. Furthermore, the images in the dataset belong to patients with different EF, as half of the patients

had EF less than 45%.
This kind of system has its problems and limitation, as it requires a huge amount of computational power to train, specifically in terms of GPU, we dealt with that problem by taking advantage of cloud computing services, another problem with these methods is that it requires a lot of data to train and test, finally because of the way the training procedure work we had to resize the images, which would affect other parameters like the EF.
Compared with the literature, only a handful of research have been done to segment all three regions, while most focused on the LV endocardium, others included LV epicardium and only two we could find that included LA, thus we tried to make our model generic and useful as possible.
While Zhuang et. al.(2021) tried to implement object detection in their module, they used in their work YOLOv3 a much older version than the one used in our work, and only segmented the LV epicardium, on the other hand, our work was the only one as far as we know that implemented the YOLOv7 algorithm in this way to get the ROI.
Also, we relied on the CAMUS dataset only, using data split into three sets, while other researchers had access to another dataset like the echo-dynamic dataset, for example, Liu et. al.(2021) was able to use the entire CAMUS dataset as a test, while the work of Kim et. al.(2021) used a huge dataset of porcine images and applied to learn transfer to test it on human images.
Our work did not split the images into two datasets either using views (A4C and A2C) or using events (ES and ED). Regarding the size of the frame, almost all literature resizes the entire dataset to a unified size. However, some went smaller like the Zeng et. al.(2023) in which the size of the frame was (112x112) pixels, since we did not use any pre-trained module, the size was chosen based on the dataset itself, and the way U-Net work. In terms of evaluation metrics, our system was able to achieve better HD than most of the work we found in the three classes.
While the system did not exceed the results achieved by the CAMUS team in the original papers, it was very close. However, it is worth mentioning that unlike the team we included all images, while they excluded the images with low quality, the class which has the lowest DSC score was the LV myocardium.

Conclusion

In this thesis, we provided a summary of the human heart, its structure and how it works, we explained the basics of computer vision and deep learning, and reviewed recent work in the field of semantic segmentation on echocardiography images, and finally we presented our full automated system to perform semantic segmentation on echocardiography images. While the results of our system are promising, there is room for improvement in terms of evaluation metrics, this improvement can be achieved by acquiring more data, and experiment more with CNN parameters. Next step should be finding a way to keep the original size of the images while keeping a reasonable batch size, also YOLO algorithm can be used to detect key points in LV that can help to measure EF. This thesis showed the ability of CNN in segmenting multiple structures of the echocardiography images, we were able to test this idea because there was an available annotated dataset. However, structures like the RV or RL also could be segmented providing the annotated data should be a priority in this field.

Bibliography

[1] A. J. Weinhaus and K. P. Roberts, *Anatomy of the human heart*, pp. 59–85. Humana Press, dec 2005.

[2] C. M. Bohun, J. E. Potts, B. M. Casey, and G. G. S. Sandor, "A population-based study of cardiac malformations and outcomes associated with dextrocardia.," *The American journal of cardiology*, vol. 100, pp. 305–309, jul 2007.

[3] S. Standring, H. Ellis, J. Healy, D. Johnson, A. Williams, P. Collins, and C. Wigley, "Gray's anatomy: the anatomical basis of clinical practice," *American journal of neuroradiology*, vol. 26, no. 10, pp. 997–1003, 2005.

[4] J. M. H. Wang, R. Rai, M. Carrasco, T. Sam-Odusina, S. Salandy, J. Gielecki, A. Zurada, and M. Loukas, "An anatomical review of the right ventricle," *Translational Research in Anatomy*, vol. 17, p. 100049, 2019.

[5] S. Whiteman, Y. Alimi, M. Carrasco, J. Gielecki, A. Zurada, and M. Loukas, "Anatomy of the cardiac chambers: A review of the left ventricle," *Translational Research in Anatomy*, vol. 23, p. 100095, 2021.

[6] S. Whiteman, E. Saker, V. Courant, S. Salandy, J. Gielecki, A. Zurada, and M. Loukas, "An anatomical review of the left atrium," *Translational Research in Anatomy*, vol. 17, p. 100052, 2019.

[7] A. Dahou, D. Levin, M. Reisman, and R. T. Hahn, "Anatomy and Physiology of the Tricuspid Valve," *JACC: Cardiovascular Imaging*, vol. 12, no. 3, pp. 458–468, 2019.

[8] E. Donal and V. Panis, "Interaction between mitral valve apparatus and left ventricle. Functional mitral regurgitation: A brief state-of-the-art overview," *Advances in Clinical and Experimental Medicine*, vol. 30, no. 10, pp. 991–997, 2021.

[9] J. Biga, L. M., Dawson, S., Harwell, A., Hopkins, R., Kaufmann, J., LeMaster, M., Matern, P., Morrison-Graham, K., Quick, D., Runyeon, "19.1 heart anatomy. Anatomy Physiology.," 2019.

[10] G. F. Tomaselli, M. Rubart, and D. P. Zipes, "Mechanisms of cardiac arrhythmias," *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine. 11th ed. Philadelphia, PA: Elsevier*, 2019.

[11] S. A. George, N. R. Faye, A. Murillo-Berlioz, K. B. Lee, G. D. Trachiotis, and I. R. Efimov, "At the Atrioventricular Crossroads: Dual Pathway Electrophysiology in the Atrioventricular Node and its underlying Heterogeneities," *Arrhythmia Electrophysiology Review 2017;6(4):179–85.*, 2017.

[12] A. Dupre, S. Vincent, and P. A. Iaizzo, "Basic ECG Theory, Recordings, and Interpretation BT - Handbook of Cardiac Anatomy, Physiology, and Devices," pp. 191–201, Totowa, NJ: Humana Press, 2005.

[13] B. E. Wright, G. L. Watson, and N. J. Selfridge, "The Wright table of the cardiac cycle: a stand-alone supplement to the Wiggers diagram," *Advances in Physiology Education*, vol. 44, no. 4, pp. 554–563, 2020.

[14] F. Chan-Dewar, "The cardiac cycle," *Anaesthesia Intensive Care Medicine*, vol. 13, no. 8, pp. 391–396, 2012.

[15] D. Lieu, "Ultrasound physics and instrumentation for pathologists," *Archives of Pathology and Laboratory Medicine*, vol. 134, no. 10, pp. 1541–1556, 2010.

[16] N. Seddon and T. Bearpark, "Observation of the inverse Doppler effect," *Science*, vol. 302, no. 5650, pp. 1537–1540, 2003.

[17] J. G. Webster, "Encyclopedia of Medical Devices and Instrumentation Second Edition Vol. 6," pp. 453 – 473, 2006.

[18] D. H. Ballard and C. M. Brown, *Computer Vision*. Prentice Hall Professional Technical Reference, 1st ed., 1982.

[19] F. M. Castelli, *3D CNN methods in biomedical image segmentation*. PhD thesis, 2019.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[21] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[22] D. Misra, "Mish: A Self Regularized Non-Monotonic Activation Function," 2019.

[23] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, 1989.

[24] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[25] Y. Lecun, L. Bottou, Y. Bengio, and P. Ha, "LeNet," *Proceedings of the IEEE*, no. November, pp. 1–46, 1998.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[28] V. Nekrasov, J. Ju, and J. Choi, "Global deconvolutional networks for semantic segmentation," *British Machine Vision Conference 2016, BMVC 2016*, vol. 2016-Septe, pp. 124.1–124.14, 2016.

[29] S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. M. Jodoin, T. Grenier, C. Lartizien, J. Dhooge, L. Lovstakken, and O. Bernard, "Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[30] S. Moradi, M. G. Oghli, A. Alizadehasl, I. Shiri, N. Oveisi, M. Oveisi, M. Maleki, and J. Dhooge, "MFP-Unet: A novel deep learning based approach for left ventricle segmentation in echocardiography," *Physica Medica*, vol. 67, no. October, pp. 58–69, 2019.

[31] T. Kim, M. Hedayat, V. V. Vaitkus, M. Belohlavek, V. Krishnamurthy, and I. Borazjani, "Automatic segmentation of the left ventricle in echocardiographic images using convolutional neural networks," *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 5, pp. 1763–1781, 2021.

[32] Z. Zhuang, P. Jin, A. N. Joseph Raj, Y. Yuan, and S. Zhuang, "Automatic Segmentation of Left Ventricle in Echocardiography Based on YOLOv3 Model to Achieve Constraint and Positioning," *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021.

[33] K. B. Girum, G. Crehange, and A. Lalande, "Learning with Context Feedback Loop for Robust Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1542–1554, 2021.

[34] F. Liu, K. Wang, D. Liu, X. Yang, and J. Tian, "Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography," *Medical Image Analysis*, vol. 67, p. 101873, 2021.

[35] Y. Lei, Y. Fu, J. Roper, K. Higgins, J. D. Bradley, W. J. Curran, T. Liu, and X. Yang, "Echocardiographic image multi-structure segmentation using Cardiac-SegNet," *Medical Physics*, vol. 48, no. 5, pp. 2426–2437, 2021.

[36] M. G. R. Alam, A. M. Khan, M. F. Shejuty, S. I. Zubayear, M. N. Shariar, M. Altaf, M. M. Hassan, S. A. AlQahtani, and A. Alsanad, "Ejection Fraction estimation using deep semantic segmentation neural network," *Journal of Supercomputing*, no. 0123456789, 2022.

[37] M. Saeed, R. Muhtaseb, and M. Yaqub, "Is Contrastive Learning Suitable for Left Ventricular Segmentation in Echocardiographic Images?," pp. 1–12, 2022.

[38] Y. Zeng, P. H. Tsui, K. Pang, G. Bin, J. Li, K. Lv, X. Wu, S. Wu, and Z. Zhou, "MAEF-Net: Multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography," *Ultrasonics*, vol. 127, no. March 2022, 2023.

[39] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, P. Lancellotti, D. Muraru, M. H. Picard, E. R. Rietzschel, L. Rudski, K. T. Spencer, W. Tsang, and J.-U. Voigt, "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging.," *Journal of the American Society of Echocardiography : official publication of the American Society of Echocardiography*, vol. 28, pp. 1–39.e14, jan 2015.

[40] C.-Y. Wang, A. Bochkovskiy, and H.-y. Liao, *YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.* jul 2022.