Automatic Quality Assessment of Focused Cardiac Ultrasound Exams

Catarina Rodrigues

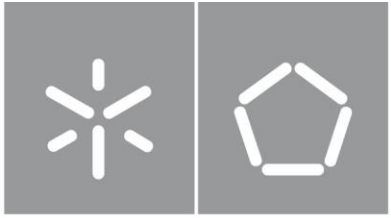
UMinho | 2022

Universidade do Minho
Escola de Engenharia

Catarina Rodrigues

Automatic Quality Assessment of Focused Cardiac Ultrasound Exams

novembro de 2022



Universidade do Minho
Escola de Engenharia

Catarina da Cunha Rodrigues

Automatic Quality Assessment of Focused Cardiac Ultrasound Exams

Dissertação de Mestrado
Mestrado Integrado em Engenharia Biomédica
Eletrónica Médica

Trabalho efetuado sob a orientação de
Professor Doutor Jaime Francisco Cruz Fonseca
Doutor Sandro Filipe Monteiro Queirós

novembro de 2022

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

AGRADECIMENTOS/ACKNOWLEDGEMENTS

Uma das mais belas e desafiantes etapas da minha vida termina com o desenvolvimento desta dissertação. Gostaria de expor o meu agradecimento a todas as pessoas que de algum modo, pequeno ou grande, contribuíram para a conclusão desta etapa.

Ao meu orientador, Professor Doutor Jaime Fonseca, agradeço a oportunidade, disponibilidade e auxílio. Ao meu coorientador, Doutor Sandro Queirós, quero agradecer pela dedicação e apoio diário ao longo deste trabalho. Obrigada pela confiança nas minhas capacidades e pela partilha dos conhecimentos e ferramentas que foram inestimáveis. Quero também deixar um agradecimento à Dra. Ana Oliveira pelo acesso aos dados utilizados no presente trabalho e por toda a ajuda.

Aos que partilharam comigo estes 5 anos, obrigada pela partilha de cada momento, pelo crescimento e apoio mútuo, e por cada riso (quer de puro divertimento, ou de puro desespero). São memórias que ficarão sempre comigo. Quero também agradecer aos meus amigos de sempre, por todo o incentivo e paciência.

À minha família, principalmente o meu pai, mãe e irmã, tenho de agradecer por todo o encorajamento e pela compreensão por todas as horas em que o estudo e trabalho me afastou. Obrigada por estarem ao meu lado a cada passo neste caminho.

Ao Gonçalo, por todas as vezes que ouviste as minhas frustrações e desabafos e respondeste com compreensão, amor e ajuda, por acreditares em mim, por toda a força que me deste. Foste sempre um porto de abrigo, um obrigada não é suficiente.

The work presented in this project was performed in the Life and Health Sciences Research Institute (ICVS, School of Medicine) and in Centro ALGORITMI (School of Engineering), University of Minho. Financial support was provided by National funds, through the Foundation for Science and Technology (FCT) - project PTDC/EMD-EMD/1140/2020 and scholarship UMINHO/BIM/2021/64. It is also acknowledged the donation of a RTX A6000 GPU by NVIDIA (USA).

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

RESUMO

Avaliação Automática da Qualidade de Exames Cardíacos de Ultrassom *Point-of-Care*

O ecocardiograma dirigido realizado à cabeceira do doente (do inglês *focused cardiac ultrasound*, FoCUS) refere-se à utilização de imagens de ecografia, obtidas de forma rápida pelos clínicos, para avaliar a estrutura e a função cardíaca. Nos últimos anos, o FoCUS tornou-se uma ferramenta de diagnóstico de primeira linha indispensável, complementando a examinação física tradicional e acelerando a avaliação dos doentes em contexto agudo. Este exame pode ser realizado por uma vasta gama de médicos, de várias especialidades e com diferentes níveis de experiência, os quais devem ser proficientes na aquisição e interpretação das imagens. Estando a sua acuidade clínica intrinsecamente dependente da competência do utilizador, é expectável que operadores menos experientes estejam sujeitos a adquirir imagens das diferentes janelas acústicas cardíacas com uma qualidade inadequada. Com o objetivo de ajudar os sonógrafos a adquirir vídeos de FoCUS com elevada qualidade, esta tese propõe uma *framework* de avaliação automática de qualidade em duas etapas.

A primeira etapa consiste na classificação de cada vídeo em uma das sete vistas de FoCUS. Para tal, propõe-se uma rede neuronal com arquitetura 3D baseada na ResNet-18, aliada a uma estratégia de *augmentation* que tira proveito das especificidades do ciclo cardíaco e a uma rotina de inferência à base de múltiplos clips. Esta metodologia e os seus componentes foram avaliados através de um conjunto extenso de testes, onde mostraram acurácia e robustez. Num conjunto independente de dados de teste, esta proposta obteve um MCC de 0.9569 e uma média de F1 de 0.9501.

Após separar os vídeos por vistas, um conjunto de modelos especificamente treinados para cada vista avalia os vários atributos de qualidade e dá uma nota geral à qualidade da aquisição. O *feedback* foca-se em elementos como o ganho e a profundidade da imagem, ou a presença dos referenciais anatómicos necessários em cada janela cardíaca. No presente trabalho, os modelos propostos focaram-se nas vistas subxifóide, apical quatro câmaras e veia cava inferior. Apesar de limitados pelo elevado desbalanceamento entre classes e pelo ruído nas anotações, os modelos propostos obtiveram um MCC médio de 0.6024 e um F1 médio de 0.7243 num conjunto independente de dados de teste.

Com esta proposta, pretende-se apoiar a formação de profissionais médicos em FoCUS, bem como a sua prática clínica, para desta forma melhorar o cuidado prestado aos pacientes.

Palavras-Chave: Análise de imagem médica, Avaliação de qualidade, *Deep learning*, Ultrassom Cardíaco.

ABSTRACT

Automatic Quality Assessment of Focused Cardiac Ultrasound Exams

Focused cardiac ultrasound (FoCUS) refers to the use of ultrasound imaging to evaluate cardiac structure and function at the bedside by a treating physician. In recent years, FoCUS has become an indispensable first-line diagnostic tool, complementing the traditional physical examination and accelerating patients' evaluation in acute care settings. FoCUS may be carried out by a wide range of medical professionals, with varied specialties and backgrounds, all of whom should be proficient in image acquisition and interpretation. With its clinical efficacy tightly dependent on the operator's skill, while experienced practitioners are expected to easily find and acquire each cardiac window, less trained technicians are prone to obtain images with suboptimal quality. Aiming to assist ultrasonography practitioners to acquire high quality FoCUS videos, this thesis proposes the development of a two-stage automatic quality assessment framework.

The first stage comprehends the classification of each video into one of seven FoCUS views. To do so, a 3D neural network architecture based on the ResNet-18 was proposed, along with a training strategy that leverages of domain knowledge into the augmentation scheme and a multi-clip inference routine. This pipeline and the blocks it entails were evaluated in an extensive set of experiments, showing its accuracy and robustness. In a held-out test set, the proposal achieved a MCC of 0.9569 and a macro-averaged F1-score of 0.9501.

Upon being separated by views, each video is then passed through view-specific models that assess a variety of quality attributes and provide an overall acquisition quality score. The quality feedback focuses on features such as image gain, acquisition depth, and the presence of the necessary anatomical references in each cardiac window. At this stage, the current work focused in the subxiphoid, apical four-chamber and inferior vena cava views. Despite affected by class imbalance and noisy labels, the proposed models achieved an average MCC of 0.6024 and an average F1-score of 0.7243 on the held-out test set.

With this proposal, one intends to support medical professionals performing FoCUS in clinical practice, allowing them to improve their technique, and, in this way, improve patients' care.

Keywords: Cardiac Ultrasound, Deep learning, Medical image analysis, Quality Assessment.

TABLE OF CONTENTS

List of Abbreviations and Acronyms

#

2D two-dimensional

3D three-dimensional

A

A4C apical four-chambers

AI artificial intelligence

ANN artificial neural network

AUC area under the curve

AV aortic valve

B

BN batch normalisation

BPM beats per minute

C

CMR cardiac magnetic resonance

CNN convolutional neural network

CT computed tomography

D

DL deep learning

E

ECG electrocardiogram

ECR Ethics Committee for Research

F

FC fully connected

FN false negative

FoCUS focused cardiac ultrasound

FOV field of view

FP false positive

FPS frames per second

H

HCPA *Hospital de Clínicas de Porto Alegre*

G

GSPECT gated single-photon emission computed tomography

I

IVC inferior vena cava

L

LA left atrium

LSTM long short-term memory

LV left ventricle

M

MCC Matthews correlation coefficient

MHSA multi-head self-attention

ML machine learning

MV mitral valve

P

PET positron emission tomography

PM papillary muscles

POCUS point-of-care ultrasound

PSLA parasternal long-axis

PSSA parasternal short-axis

R

RA right atrium

ReLU rectified linear unit

RNN recurrent neural network

ROC receiver operating characteristic

RV right ventricle

S

SA sinoatrial

SPECT single-photon emission computed tomography

SX subxiphoid

T

TN true negatives

TP true positives

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 THE CARDIOVASCULAR SYSTEM

Blood must be continually pumped through the blood vessels of the body so that it can reach the cells and exchange nutrients with them. The cardiovascular system performs this task and consists of a pump (the heart), a series of distributing and collecting tubes (blood vessels), and an extensive system of thin vessels (capillaries) that allow rapid exchange between tissues and vascular channels. To achieve its goal, the heart beats about two and a half billion times in a lifetime, continuously recycling about 5 litres of blood [1–3].

1.1.1 ANATOMY AND PHYSIOLOGY

In the thoracic cavity, the heart can be found between the lungs and the mediastinum. It is a hollow muscular organ that is relatively small (about the size of a fist). The base, which is the widest part of the heart, is superior to its tip (called apex), which rests on the diaphragm. The base faces the right shoulder, while the apex points to the left hip. It is responsible for separating deoxygenated blood from the oxygenated one, keeping blood flowing in one direction, moving blood through the system by creating blood pressure and, finally, regulating the blood supply according to the body's needs [3].

Enclosing the heart, there is a two-layered serous membrane, the pericardium (Figure 1.1). It consists of a loosely fitting sac that protects the heart. The pericardial cavity contains a few millilitres of

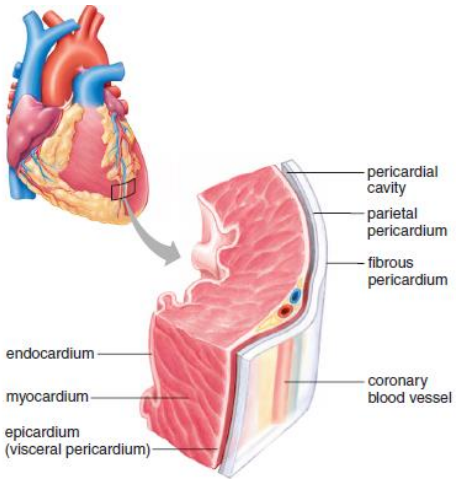


Figure 1.1 - Pericardium and heart wall [3].

pericardial fluid, a lubricating serous fluid that reduces friction as the heart beats [2]. This covering layer of the heart confines it to its allotted space while still allowing it to perform its function [3].

The wall of the heart consists of three layers: the epicardium, the myocardium, and the endocardium, as shown in Figure 1.1. The myocardium is composed of cardiac muscle and makes up 95% of the heart [2, 3].

The heart consists of a double pump with four chambers (Figure 1.2) [4]: two upper atria which are separated by the interatrial septum, and two lower ventricles that, similarly, are separated by the interventricular septum. Thus, the heart is longitudinally divided into two functional halves. The thickness of the myocardium of a chamber depends on its function. The atria have thin (2-3 mm) walls because they only conduct blood into the adjoining ventricles. The ventricles, on the other hand, are thicker and pump blood into the blood vessels that supply the body. The left ventricle (LV) has a thicker wall than the right ventricle (RV) (10-15 mm and 4-5 mm, respectively), which is explained by the fact that the right side of the heart pumps blood through the vessels of the nearby pulmonary circuit, while the left side of the heart pumps blood through the vessels of the systemic circuit, *i.e.* the rest of the body [1–3]. The following description portraits the sequence of the cardiac cycle, which includes both systemic and pulmonary circulations. These circuits, represented in Figure 1.3, function in a series arrangement as the output of one is the input of the other [2, 5].

The right atrium (RA) receives deoxygenated blood from the body through three veins: the superior vena cava, the coronary sinus, and the inferior vena cava. Venous blood flows from the RA

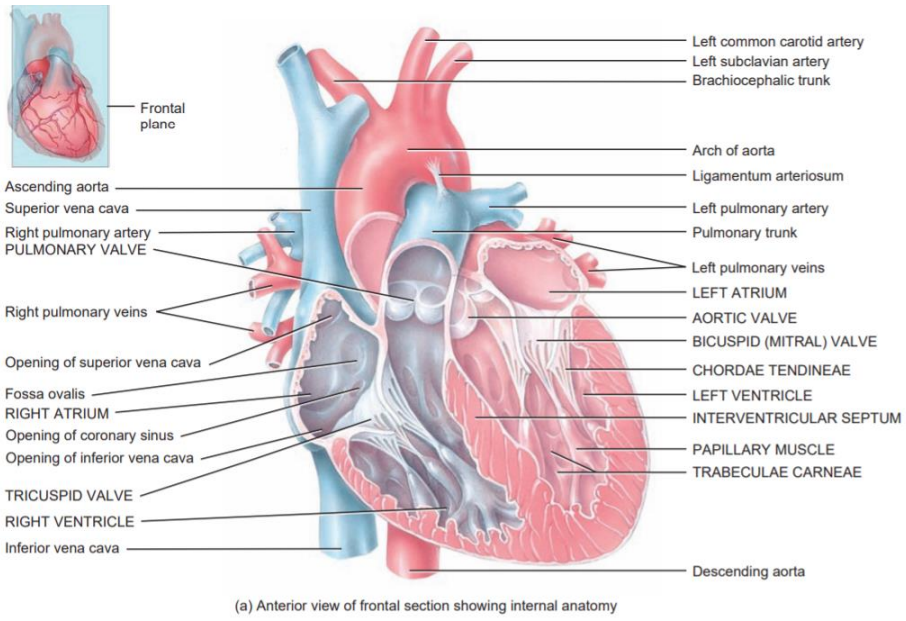


Figure 1.2 - Internal heart anatomy [2].

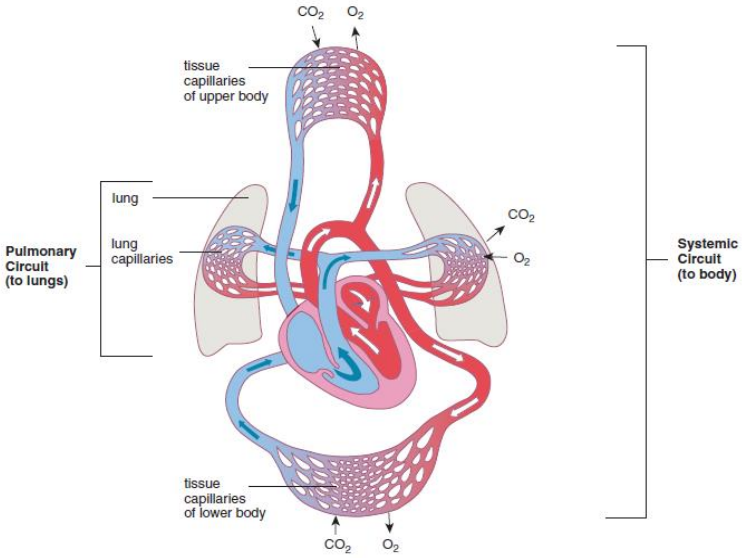


Figure 1.3 - Series arrangement of pulmonary and systemic circulations [3].

through an atrioventricular valve into the RV. This valve is called tricuspid, and like the other heart valves, directs the blood flow and prevents backflow [3]. In the RV, the cusps of the tricuspid valve are connected to fibrous cords, called chordae tendinea, which in turn are connected to the papillary muscles (PM), which are cone-shaped extensions of the heart muscle (Figure 1.2). Blood from the RV flows through a semilunar valve into the pulmonary trunk. This valve is called the pulmonary valve and prevents blood from flowing back into the RV [3, 4]. Subsequently, gas exchange occurs as the blood flows through the capillaries of the lungs, enriching it with oxygen.

The left atrium (LA) then receives oxygenated blood from four pulmonary veins. Blood flows from the LA through an atrioventricular valve, the mitral valve (MV), into the LV [3]. The PM in the LV are quite large and the chordae tendineae attached to the MV valve are thicker and stronger than the ones in the RV. Blood is ejected from the LV through a semilunar valve into the ascending part of the aorta. This semilunar valve is appropriately called the aortic valve (AV). The semilunar cusps of this valve are larger and thicker than those of the pulmonary valve. The aorta distributes blood to the rest of the body, and gas and nutrients exchanges occur as the blood flows through the tissue capillaries [3].

This cardiac cycle is enabled by the conduction system of the heart. It is a pathway of specialised cardiac muscle fibres that initiate contraction of the atria and ventricles. The conduction system is considered intrinsic because the heart beats without the need for external stimulation [4]. The heartbeat is controlled by nodal tissue sending an impulse at a given rate. The pulses spread to the atria, making them contract simultaneously. When the atrioventricular node receives the impulses, there is a slight delay before transmitting the signal which allows the atria to finish the contraction before the ventricles begin theirs [3, 6]. Sympathetic stimulation speeds up the heart rate when needed, while parasympathetic

stimulation does the opposite. A normal heart rate for adults lays between 60 to 100 beats per minute (bpm) [4].

The cardiovascular system includes the heart and an estimated 100 thousand kilometres of vessels spread across our body through which blood flows [2, 4]. Blood vessels form a tubular network and belong to one of three main types: arteries, capillaries, and veins (Figure 1.4) [3, 4].

Arteries are responsible for transporting blood away from the heart and act as pressure reservoirs to maintain blood flow during ventricular relaxation [6]. Arterioles are small arteries visible to the naked eye that are less elastic than the larger arteries. Arterioles branch into capillaries, which are extremely narrow blood vessels (7 to 10 μm) whose wall consists of only one layer of endothelial cells. Capillaries are an important part of the cardiovascular system because nutrients, waste products and oxygen molecules are exchanged only through their thin walls [1, 3, 4]. Veins and smaller vessels called venules collect capillary blood and carry it back to the heart. These vessels sometimes do not have enough pressure to transport blood back to the heart, so they present valves that prevent blood from flowing back [4]. At any given time, more than half of the total blood volume is in the veins and venules, and thus veins act as a blood reservoir [1, 3].

Apart from the lungs, veins from the entire body converge in the venae cavae. The superior vena cava receives blood from the upper extremities. The inferior vena cava (IVC), which has the largest diameter of any vessel, follows the aorta through the abdominal cavity until it enters the RA and drains the lower extremities, as well as hepatic, renal and iliac veins [2–4]. The IVC is located to the right of the abdominal aorta and is sensitive to changes in right atrial pressure and volume status. In normal situations, during inspiration, the intrathoracic pressure becomes negative, and the normal response is a decrease in IVC diameter. However, in ventilated patients, the change in IVC during inspiration reverses and causes IVC distention [7].

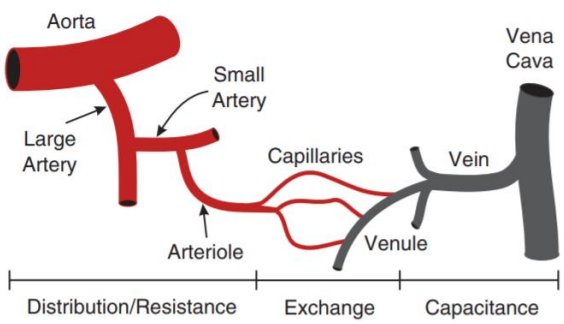


Figure 1.4 - Main types of blood vessels [5].

1.1.2 Cardiac Imaging

Imaging allows visualisation and assessment of internal structures, exploiting different physical principles to produce visual representations, or images, in a non-invasive approach [8–10]. Imaging data accounts for about 90% of all health data and are therefore one of the most important sources of information for clinical analysis and medical intervention.

Echocardiography is the cardiac subspecialty of ultrasonography, a technique that allows visualisation of both superficial and deep structures of the body by registering ultrasound pulses reflected from the tissues. Due to its non-invasive nature and cost effectiveness, echocardiography is a widely used technique for cardiac imaging. Current echo machines can be portable and can record information in real time as single images or short videos [9, 10]. The technique can be performed virtually anywhere, in the emergency room, at the bedside or even during surgery. A transducer comes into contact with the skin and generates high-frequency sound waves that propagate through the body and are reflected at the interface of different tissues. The echoes are reflected throughout the body until they are picked up by the transducer and converted into electrical energy. The electrical signals are then recorded and displayed on a monitor in the form of a cross-sectional image [9]. Real-time two-dimensional (2D; or B-mode) echocardiography is used to assess heart function, valvular disease, and congenital defects. In its turn, M-mode is an echocardiographic mode that describes the movement of heart structures located in a single beam over time, which allows for more detailed tracking of structural dynamics [10]. The main disadvantage of any ultrasound mode is that it is operator-dependent, requiring in-depth knowledge [11].

Cardiac computed tomography (CT) allows visualisation of cardiovascular anatomy and is the modality of choice to assess the coronary circulation and great vessels [10–12]. In this technique, X-rays are sent through the body while the X-ray tube and detector rotate around the patient. Multiple energy absorptions are measured, recorded, and compared in a computer to determine the density of each element of the selected plane and produce cross-sectional images [9, 10]. Hereto, the use of contrast agents can enhance certain structures. Synchronisation of the image with the electrocardiogram (ECG), a technique known as ECG gating, can be used to minimise artefacts due to the heart movement [10]. CT can achieve submillimetre resolution, which is one of its main advantages as it results in high quality images. On the other hand, the use of radiation and the limited temporal resolution are its primary limitations.

Cardiac magnetic resonance (CMR) is another widely used imaging modality and is considered the gold standard for quantifying myocardial volume and function. It can provide relevant information

about most aspects of cardiac structure, valve function, flow patterns, myocardial perfusion, and coronary anatomy [10–12]. For this purpose, the patient is placed in a large scanner with a strong magnetic field that forces the protons in the body to align. Then, the body is exposed to radiofrequency waves that change the alignment of the protons. When the radiofrequency signals are turned off, protons realign with the magnetic field and send out signals, which differ based on the proton density of each tissue. These signals are recorded and used to reconstruct different images of the body. CMR computers can reconstruct tissues in any plane from the collected data: sagittal, axial, coronal, or even arbitrary planes. They can also create three-dimensional (3D) reconstructions [9, 11]. These machines can also record temporal information in the format of a CMR cine sequence, which despite having lower temporal resolution than echocardiographic images have higher quality. In addition, tissue differentiation is better than with CT images, with these scanners producing good images of soft tissues without ionising radiation [9, 10, 12]. Like CT, CMR may also benefit from the administration of a contrast agent [10]. Its main contraindications relate to the presence of metal in the patient. Patients with permanent pacemakers, defibrillators, and other implanted devices should also not undergo this exam [12]. Furthermore, it is a long examination, with expensive equipment and low availability. Overall, CMR is considered a complementary study to echocardiography, being the first choice for the diagnosis and follow-up of multiple cardiac diseases when the latter is inconclusive for a particular pathology or due to technical difficulties [11].

Nuclear medicine provides information about distribution or concentration of small amounts of radioactive substances introduced into the body [9]. Radionuclide imaging of the heart is well established for clinical assessment of myocardial perfusion and metabolism and is considered robust, accurate, and reliable [11, 12]. This modality differs from the above ones as it evaluates organ function rather than anatomy. The main techniques are single-photon emission computed tomography (SPECT), which uses radioisotopes that emit gamma radiation as tracers, and positron emission tomography (PET), which uses radioisotopes that emit positrons. Between the two, SPECT is cheaper, more readily available and allows a larger imaging time (its tracer has a half-life of up to 6 hours, compared to PET's 75 seconds). On the other hand, it has longer scan times, is susceptible to artifacts and produces images with lower resolution. In addition, only PET allows quantification of blood flow. Gated SPECT (GSPECT) is an ECG-gated technique that allows combined assessment of myocardial perfusion and left ventricular function, playing an important role in diagnosing coronary artery disease, three-vessel disease and assessing the severity of ischaemia. Theses diagnoses are based on the detection of myocardial regions with abnormal perfusion due to damaged vessels [10].

1.2 Focused Cardiac Ultrasound

The value of ultrasound as a diagnostic cardiac modality is unsurpassed in many respects, making it the first-line imaging modality for most cardiac studies. As mentioned earlier, it is more portable and cost-effective than other imaging modalities, while providing the opportunity for real-time imaging of cardiac structure and function. Unlike methods that expose patients to radiation, diagnostic ultrasound has no known adverse effects, which allows for safe, serial examination of patients. In this way, cardiac ultrasound can provide valuable information in critical situations and emergencies [12–15].

Due to technology advancements, traditional cart-mounted ultrasound machines are being replaced by portable point-of-care ultrasound (POCUS) machines. These are often packaged in the form of a laptop or as an ultrasound probe with a wired (or wireless) connection to a mobile device (Figure 1.5). This easy accessibility of POCUS devices has paved the way for new clinical use cases. Particularly, in emergency or critical care settings, physicians can use a portable device to quickly perform the patient's initial assessment and make time-critical and maybe life-saving diagnostic decisions [16–18].

POCUS devices have significantly fewer features and capabilities, making them easier to use. The simplified operation, along with the much smaller size and lower cost, has led to their use by non-traditional users of cardiac ultrasound. Cardiologists and sonographers are not always available during medical emergencies, so clinicians from a variety of fields are interested in exploiting the diagnostic value of cardiac ultrasound in their practice. These include emergency physicians, intensive care specialists, anaesthesiologists, sonographers/cardio-physiologists, and fellows in training. Despite their diverse medical backgrounds, provided they have the appropriate training, they can perform cardiac ultrasound and identify important findings to obtain important answers in an emergency [13, 15]. This is the concept of the focused use of cardiac ultrasound, whose hypothesis is that those who are not trained in echocardiography can learn to obtain and interpret cardiac ultrasound images as an addition to their physical examination [13].

Figure 1.5 - Example of a POCUS device [18].

Focused cardiac ultrasound (FoCUS) is defined as a point-of-care cardiac ultrasound examination to assess cardiac pathophysiology by the treating physician. FoCUS differs from echocardiography in terms of where it is performed, the providers who perform the exam, the equipment used, and, most importantly, the scope of the exam. FoCUS is mainly used in intensive care units and emergency departments to evaluate patients in shock, with symptoms of dyspnea and chest pain, among others [13, 15, 17]. 'Focused' refers to a narrowed, specific question and scope of expertise [13]. The purpose is to find answers to specific questions which have clinical implications. The provider conducting the FoCUS study is looking for a yes or no answer, and he/she is not responsible for any incidental findings on a stored clip [17]. FoCUS is limited by several factors, including time constraints, a limited image acquisition protocol, the experience of the examiner, and the technical capabilities of the available equipment [15].

The primary role of FoCUS is the timely assessment of symptomatic patients. This evaluation includes assessment of pericardial effusion, or evaluation of relative chamber size, global cardiac function, and volume status of the patient. For example, the latter may be evaluated by looking at left ventricular size, ventricular function, and inferior vena cava's size and respiratory change (mostly qualitative). In addition, FoCUS is used to guide urgent invasive procedures or to assess the position of a transvenous pacemaker [17, 19, 20]. Performing FoCUS in emergencies has improved outcomes by reducing the time to diagnose and treat traumatic cardiac and thoracic injuries. Studies have shown not only that morbidity has decreased with the inclusion of FoCUS in trauma diagnosis, but also that the use of FoCUS in penetrating trauma has a mortality benefit. Because FoCUS can provide important information in cardiac arrest that can directly alter management, it is also currently being incorporated into the advanced cardiovascular life support algorithm [15].

1.2.1 Ultrasound Principles

Although ultrasound equipment and image quality have improved with time and technological advances, modern ultrasound machines are still based on the same original physical principles from centuries ago. This section provides an overview of the basic principles of ultrasound physics and imaging modes [21, 22].

Sound is composed of waves. A wave represents the propagation of energy produced by the motion of a particular entity. Sound is a longitudinal, mechanical wave that propagates in a straight line and as such requires a particle medium. Like any waveform, sound is defined by several parameters, including frequency, period, amplitude, intensity, wavelength, and velocity. The first four are defined by the sound source. The number of waves within a given unit of time is called the frequency and is measured

in hertz (Hz or cycles per second). The period of a wave is the reciprocal of its frequency, and it represents the time it takes to complete one cycle. Amplitude is the maximum height that occurs in a wave minus its mean value. Amplitude can also be expressed in decibels, which corresponds to a logarithmic scale. Intensity is a magnitude divided by a unit of area. The strength of a sound beam is described by both amplitude and intensity. Wavelength is defined as the distance between corresponding points on two successive waveforms. Velocity is the speed at which sound travels through a medium and is the product between wavelength and frequency. Both velocity and attenuation depend on the nature of the medium and the wave's inherent properties (amplitude and frequency) [12, 21, 23, 24].

The speed of propagation of sound waves in human tissues is 1,540 m/s, while in air it is 330 m/s. Humans can hear sound waves with frequencies between 20 Hz and 20 kHz. Thus, ultrasound is, predicably, defined as sound with frequencies above 20 kHz. Transducers with frequencies between 1 and 20 MHz are used for diagnostic medical ultrasound [21, 24].

In imaging, ultrasound waves are generated by transducers equipped with piezoelectric crystals. These crystals change shape when electric current is passed through them, and they produce electrical signals when they are mechanically compressed [21–23]. The individual crystals lie side by side in an 'array' and are electrically connected to each other. Vibration and consecutively ultrasound emissions are produced by applying a fast, alternating current to the crystals. The transducer acts as both a transmitter, sending out an ultrasound pulse, and a receiver, receiving the ultrasound signals reflected from the internal tissue interfaces, which compresses the crystals, generating electrical signals. The 'transmitting phase' is very short (only 0.5 to 3 μs), while the 'receiving phase' is much longer (up to 1 ms), as it needs time to detect all echoes from different depths. The pulse repetition period is defined by the combined duration of the transmitting and receiving phases. The interval between emission of the signal and reception of the echo is determined by the depth of the structure in analysis and the speed of sound in the tissues. Imaging is based on this interval, so structures represented at higher depth, had longer flight times [22–24].

The electrical pulse generated in the crystals by the echoes is then sent to the computer/display, that calculates the time it took for the electrical pulse to travel into the body and back, and determines where on the display (*i.e.* at which depth) a dot is projected. The shade of gray (from light to dark) is determined by the intensity of the reflected echo, amplified by a value of gain that can be adjusted by the user. A two-dimensional image is built up by firing a beam vertically, waiting for the returning echoes, receiving all echoes along the beam, maintaining the image along the beam, and a new beam is sent out in the adjacent region. The ultrasound beam may be steered either mechanically or electronically. By

retaining the information from all returned beams, a complete image is formed. With today's technology, ultrasound machines can produce an image with sufficient depth and resolution and with good temporal resolution for 2D visualisation of normal cardiac activity [23, 24].

The resolution of ultrasound imaging comprises temporal and spatial resolution. The latter is further divided into axial, lateral, and elevational resolution, as shown in Figure 1.6. Axial resolution consists of the ability to distinguish structures aligned along the beam and it gets better with higher frequency. The ability to discern objects located on the perpendicular axis to the beam is the lateral resolution. The most important determinant of lateral resolution is beam width, being best at shallow depths and for narrow beams, and worse at deeper imaging and for wide beams [22–24]. Elevational resolution indicates the extent to which an ultrasound system can resolve objects within an axis that is perpendicular to the plane formed by the axial and lateral dimensions, and thinning the beam improves the elevational resolution. In the near field, the beam loosely maintains the transducer's diameter. However, the beam diverges and widens as it becomes further away from the transducer. This area is the far field. The focus position, *i.e.* the point where the beam is finest and the divergence is smallest, is adjustable by the operator and is one of the most important steps in image optimisation. Beam width is a function of size, shape, frequency, and focus of the transducer [24]. Temporal resolution is the ability to detect that an object has moved over time; it is described by the frame rate (in Hz or frames/second). The frame rate depends on the time required to create a single line, and the number of lines that make up each image. Temporal resolution is improved if the sweep speed of the beam is increased, being limited by the speed of sound. If the desired depth is decreased, the time from sending to receiving the pulse is shortened, and the next pulse (for the next beam) can be sent sooner, increasing the sweep

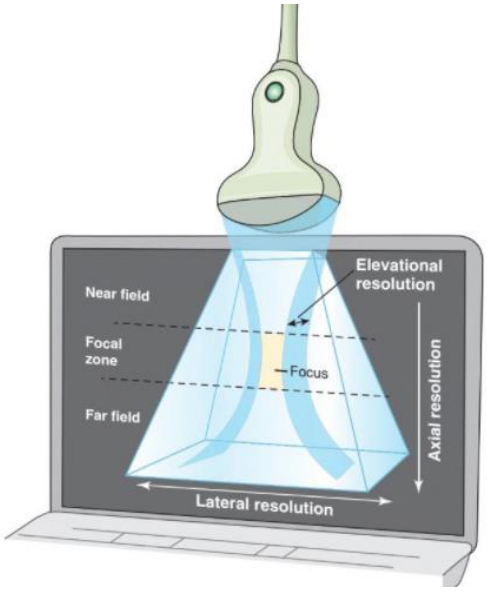


Figure 1.6 - Types of resolution of ultrasound imaging [21].

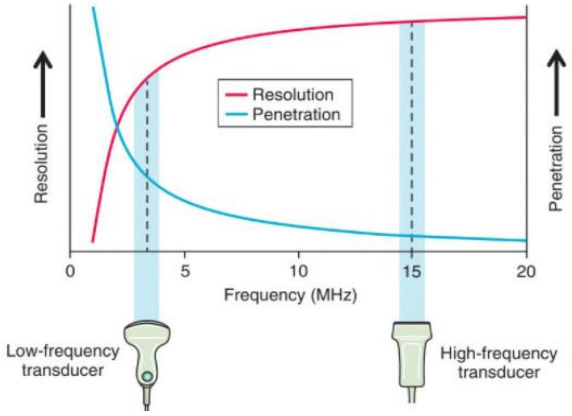


Figure 1.7 - Resolution and penetration as a function of the transducer's emission frequency [21].

speed and frame rate. Thus, frame rate is determined by the sector size (width and depth) and the line density (which also affects the lateral resolution) [22, 23].

There are a variety of transducers, each with a different width, emission frequency, and focal characteristics. The transducer's width is related to the ultrasound emission frequency, and it is important to consider the relationship between resolution and penetration to select the frequency range appropriate for a particular examination. Ultrasound penetration measures the ability of the ultrasound beam to pass through various cardiac structures. An increase in ultrasound emission frequency results in an increase in image resolution. However, if attenuation also increases, the ultrasound penetration decreases. This trade-off is illustrated in Figure 1.7 [12, 23]. In echocardiography, phased array transducers are commonly used because their small footprint allows imaging through small intercostal windows [17, 22].

Images are formed by interactions between the emitted ultrasound waves and the tissues. These interactions can be of different types (Figure 1.8). Reflection happens when an ultrasound beam hits the interface between two different tissues and part of the signal is reflected to the probe. The extent of reflection depends on the acoustic impedance of the two tissues, and on the angle between the tissue boundary and the ultrasound beam. When the tissue border is orthogonal to the ultrasound beam,

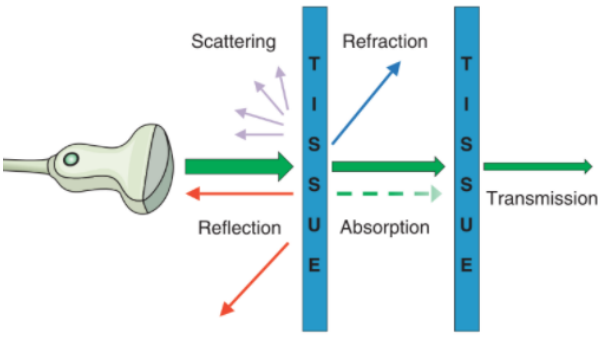


Figure 1.8 - Types of interactions between ultrasound waves and the tissues [21].

maximum reflection is achieved. When an ultrasound beam encounters a boundary consisting of small structures (smaller than the sound's wavelength), the ultrasound beam is scattered, resulting in reflection of the beam in all directions and a disordered return signal. This effect is the cause of the loss of most of the signal. Refraction refers to the bending of the ultrasound beam when it enters a medium in which its velocity of propagation is different. The degree of refraction depends on the angle between the beam and the surface and on the different propagation velocities of the tissues at the interface. As the ultrasound propagates through the tissue, some of the energy is lost due to absorption and scattering, which is a process called attenuation. This results in a weaker signal intensity from structures farther from the probe. The higher the frequency, the greater the attenuation and the shallower the penetration depth. Modern scanners use automatic 'time-gain compensation' to mitigate this problem [22]. Time-gain compensation proportionally amplifies echoes based on the time interval since the initial pulse (*i.e.* the depth of the structure). Since attenuation varies from person to person, time-gain compensation can be adjusted by the user. Usually, the near-field gain is set to a lower value, while the far-field gain is gradually increased to achieve better image quality [24].

1.2.2 Cardiac Views

The views acquired during a FoCUS examination are familiar to any echocardiologist. For all providers, this is a skill like any other, and one that takes much more time to master than to learn [17]. Echocardiographic examination usually requires a frequency of at least 2.0 MHz, although this may change depending on the characteristics of the patient's chest. In a child or thin adult, a 3–5 MHz transducer with good resolution and penetration can be used. A 7–7.5 MHz transducer provides better quality images in a neonate, while a 2–2.5 MHz transducer is optimal for an obese adult [12, 20]. The echocardiographic examination is usually performed in the left lateral decubitus position with the chest flexed at 30° to position the heart closer to the anterior chest wall (Figure 1.9 A). The patient's left arm

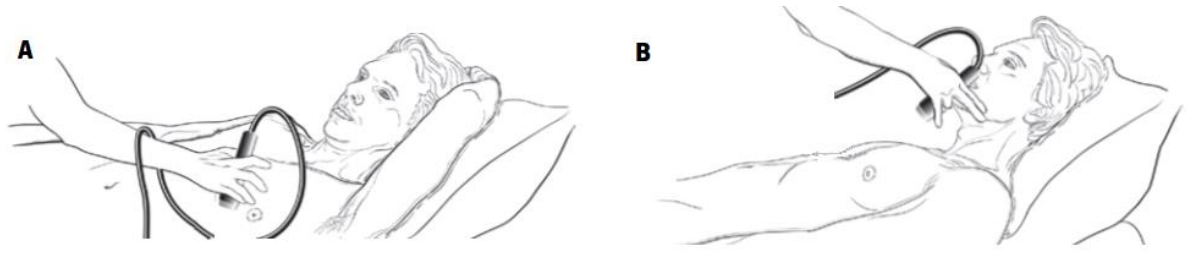


Figure 1.9 - Patient positioning (adapted from [12]): A) decubitus position and B) supine position.

is placed under the head to widen the intercostal spaces. However, some views require repositioning to supine position (Figure 1.9 B) [12].

For each acquisition, there is a marker in the image that indicates the probe's orientation relative to the patient's head. In standard cardiology, the probe marker is on the right of the screen, whereas in general ultrasound, the marker is placed on the left side of the screen [17]. Because FoCUS is performed by physicians with different training backgrounds, the latter method is used as convention.

During an examination, the heart is imaged from several windows. Each window is defined by the position of the transducer (parasternal, apical or subcostal) and the orientation of the plane through the heart (*e.g.*, long-axis, short-axis, four-chamber, two-chamber, five-chamber) [23].

In the parasternal long-axis view (PSLA, Figure 1.10), the transducer is placed on the 3rd or 4th left intercostal space adjacent to the sternum to obtain a long-axis view of the heart bisecting the aortic and mitral valves [23]. In this view, the ultrasound plane intersects an imaginary line drawn from the right shoulder to the left hip, representing a long-axis slice through the LV. Due to variations in patient anatomy, the transducer may have to be slightly repositioned to optimise the image. The RV is in the region closest to the transducer. The LA can be seen at the image's far-right region. In some cases, it is possible to see a round structure representing the left inferior pulmonary vein immediately posterior to the LA. The proximal structures of the ascending aorta and the right and noncoronary cusps of the AV can be visualised. The mitral leaflets are also readily seen, as are the chordae and their fusion with the PM. The left ventricular outflow tract is located between the interventricular septum and the anterior mitral leaflet, which can be seen in the anterior and posterior portions of the image, respectively. The LV is in the left portion of the image when the interventricular septum and posterolateral wall are placed proximal and distal to the transducer, respectively. The pericardium appears at the bottom of the image, where the descending thoracic aorta is sometimes visible [12, 23].

The parasternal short-axis view (PSSA, Figure 1.11) is obtained by rotating the transducer clockwise 70° to 110° from its original position. The ultrasound plane now intersects an imaginary line

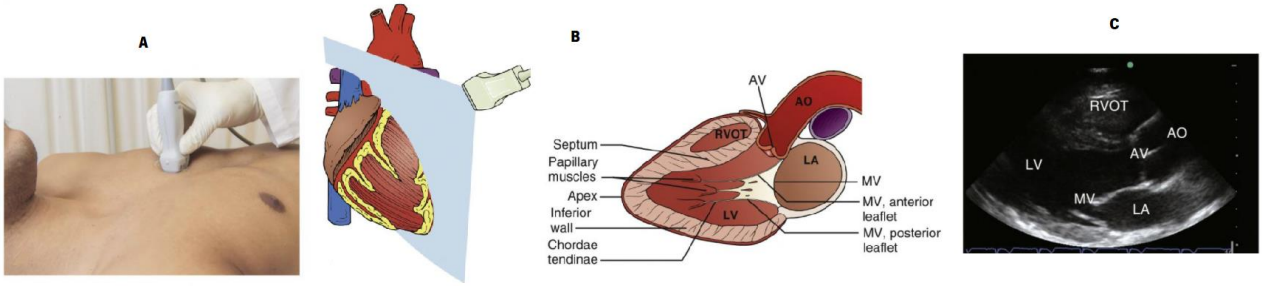


Figure 1.10 - Paraesternal long-axis view [21]: A) probe position, B) scan plane and C) ultrasound image.
AO – Aorta; AV – Aortic Valve; LA – Left Atrium; LV – Left Ventricle; MV – Mitral Valve; RVOT – Right Ventricular Outflow Tract.

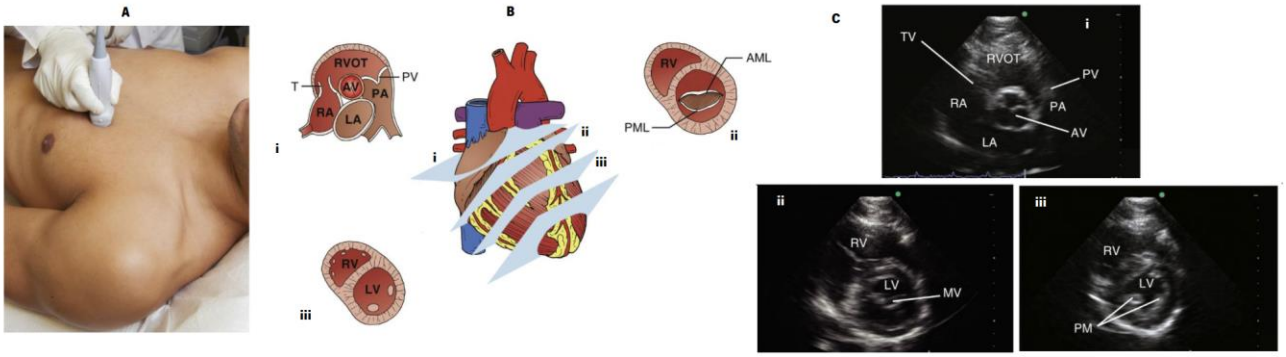


Figure 1.11 - Paraesternal short-axis view (adapted from [21]): A) probe position, B) scan planes and C) ultrasound images, for i) aortic level, ii) mitral Level and iii) papillary muscles level.

AML – Anterior Mitral Leaflet; AO – Aorta; AV – Aortic Valve; IVS – Interventricular Septum; LA – Left Atrium; LV – Left Ventricle; MV – Mitral Valve; PA – Pulmonary Artery; PM – Papillary Muscle; PML – Posterior Mitral Leaflet; PV – Pulmonary Valve; RA – Right Atrium; RV – Right Ventricle; RVOT – Right Ventricular Outflow Tract; TV – Tricuspid Valve.

that runs from the left shoulder to the right hip. Tilting the transducer slightly allows three different imaging planes: aortic valve, mitral valve, and papillary muscles [12, 23].

The base of the heart can be visualised by tilting the transducer in the direction of the right shoulder in the so-called PSSA plane at the level of the aorta. The AV is at the centre of the scan, encircled by both atria, the interatrial septum, two leaflets of the tricuspid valve, the wall of the RV, the right ventricular outflow tract, the pulmonary valve, and the main pulmonary artery. Anterior to the LA and posterior to the RV, all three cusps of the AV in its 'Y' configuration are perceptible in the moment of ventricular diastole. The LA appears posterior to the AV and is separated from the RA by the atrial septum [12, 23].

A slight tilt of the transducer downward and to the right results in a PSSA view at the level of the mitral valve. The septal leaflet of the MV is seen in the anterior position and its posterior leaflet in the lower part of the image. The mitral orifice has an appearance similar to a fish mouth. The RV can also be seen [12, 23].

When tilting the transducer more parallel to the direction of blood flow (slightly more downward), the angulation permits the visualization of the contracting papillary muscles in the LV (at the 3 and 8 o'clock positions) [12].

In the apical window, the ultrasound plane intersects an imaginary line running from the left median axillary line superiorly and medially to the patient's right scapula. The transducer is placed tangentially in the fifth intercostal space along the median axillary line at the apical level. The apical four-chamber view (A4C, Figure 1.12) shows all four chambers simultaneously, interventricular and interatrial septa, the mitral and tricuspid valves, and the crux of the heart. The apex of the heart and the atria can be seen at the top and bottom of the image, respectively. The right and left cavities of the heart along

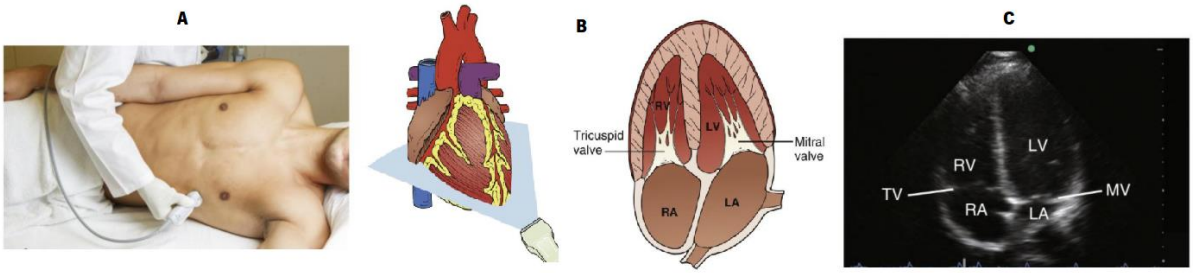


Figure 1.12 - Apical four-chamber view [21]: A) probe position, B) scan plane and C) ultrasound image.

LA – Left Atrium; LV – Left Ventricle; MV – Mitral Valve; RA – Right Atrium; Right Ventricle; TV – Tricuspid Valve.

with their respective atrioventricular valves are, respectively, located in the left and right regions of the image. The anterior mitral leaflet appears medially, with the posterior leaflet laterally. Regarding the tricuspid valve, both septal and lateral leaflets are seen. The MV is usually at a slightly higher position than the tricuspid valve [12].

The sub-costal, or subxiphoid, view (SX, Figure 1.13) provides an assessment of both right and left sides of the heart. Hereto, the transducer is placed in the middle of the epigastrium and tilted down along an imaginary line that runs to the patient's left shoulder. This allows visualisation of the RV, inferior interventricular septum, and anterolateral left ventricular walls. The interatrial septum is nearly perpendicular to the ultrasound beam. The liver is located at the top of the image, with the RV below and its apex directed to the right. The contractility of the inferior and lateral walls and the apex of the RV, as well as the presence of pericardial effusion, can be analysed [12, 23].

From the sub-costal view, a 90° rotation about the axis of the RA provides a long-axis view of the IVC through the liver (Figure 1.14) [17]. An ideal view of the IVC shows this vein come into the RA and, simultaneously, part of the hepatic vein opening into the IVC. The IVC typically exhibits calibre variations throughout the respiratory cycle [7]. To avoid confusing the IVC with the adjacent abdominal aorta, it helps to identify the junction between the RA and the IVC. It is also essential to capture the IVC with the transducer centred on the longitudinal axis to accurately determine its true diameter. Off-centre imaging will result in a falsely reduced diameter, the so-called 'cylinder effect' [7].

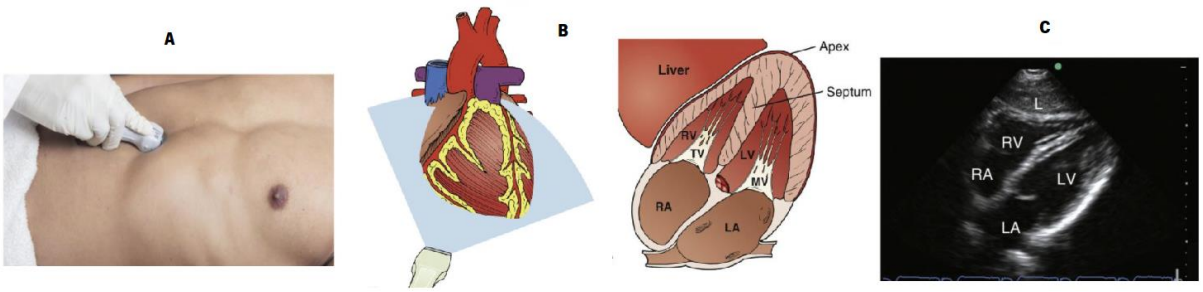


Figure 1.13 - Sub-costal view [21]: A) probe position, B) scan plane and C) ultrasound image.

L – Liver; LA – Left Atrium; LV – Left Ventricle; RA – Right Atrium; RV – Right Ventricle.

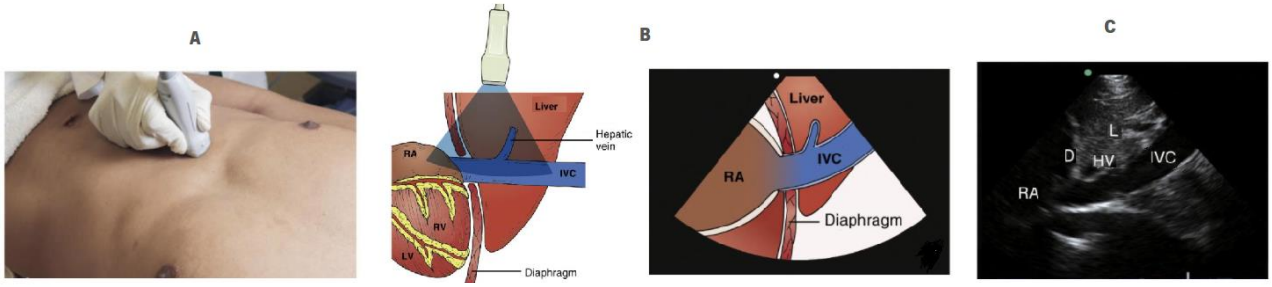


Figure 1.14 - Inferior vena cava view (adapted from [21]): A) probe position, B) scan plane and C) ultrasound image. D – Diaphragm; HA – Hepatic Artery; IVC – Inferior Vena Cava; L – Liver; RA – Right Atrium.

There is no question that certain windows are easier to learn than others. Studies of FoCUS training have shown that parasternal views are generally easier to learn. The landmarks for these windows tend to be more reliable. Images from the parasternal window are easier to keep stable and consistently provide more interpretable images than apical ones. The parasternal view is also less dependent on patient's positioning and is subject to less interference from patient's body habitus [13]. In turn, the apical and subcostal views are more difficult to obtain, and more training appears to be required to optimise these views to patient's position, body habitus and respiratory cycle [19, 25]. Notably, an off-axis imaging and, consequently, foreshortening can negatively impact image interpretation, revealing the criticalness of optimising each cardiac view to guarantee adequate diagnosis and clinical management [13].

1.3 Motivation

The use of FoCUS in daily clinical practice is on the rise, being a great tool in critically ill patients and acute situations. However, image quality is highly dependent on the equipment used, the operator's experience and even on the patient's characteristics. Although clinical guidelines lay down the criteria for a correct acquisition of each view, it is not straightforward to follow them, especially for inexperienced sonographers. Furthermore, the clinician's ability to interpret an echo study is highly dependent on image quality, so suboptimal images can compromise interpretation and adversely affect patient care. Therefore, video quality assessment is an important issue in ultrasound imaging to ensure that the acquired videos are suitable for health assessment. This would allow automatic selection of videos to be stored in the patient's record, keeping only those that can be interpreted for diagnosis. It would also allow the elimination of videos without the required anatomical features from further automatic processing.

Additionally, as POCUS equipment becomes increasingly available, less experienced clinicians are using FoCUS. Yet, it still presents a lack of educational opportunities and insufficient volume of trained personnel for assistance and feedback during student training on this technique. Studies have shown that a hands-on-approach is paramount to achieve proficiency in FoCUS [25]. This not only allows practical

training by the trainees, but also keeps them motivated. As psychology's self-determination theory claims, students are more motivated to learn when their instructor encourages autonomous work [26], which also suggests that a practical approach is not sufficient alone. Feedback on how to improve the acquisition plays a fundamental part.

Some artificial intelligence-based FoCUS-oriented solutions have already been developed. Some authors have focused on assessing the acquired images' overall quality by assigning a global score [27–29]. However, this approach does not allow to know what aspects of the acquisition fall short from ideal. Others have proposed the integration of this type of algorithms in the equipment itself, guiding the movement of the probe to the correct view [30]. However, it limits the acquisition to specific equipment(s), not broadly available and that may be (cost-wise) unfeasible.

Altogether, this highlights the need to develop new strategies for quality assessment of FoCUS videos.

1.4 Aims and contributions

Aiming to assist ultrasonography practitioners to acquire high quality FoCUS videos, this thesis proposes the development of methods that allow automatic quality assessment and feedback that can serve both trainees and day-to-day users of this imaging technique in the clinical practice. Videos from their routine cardiac exams are given feedback on the overall acquisition quality and which attributes are sub-optimal. Feedback should cover features such as image gain, acquisition depth, the presence of the necessary anatomical references and others, as well as an overall quality score. As each of the assessed attributes is specific to each cardiac view, the solution was to create specific deep learning classifiers for each view. This requires the development of two main blocks: a view classification module and a view-specific quality assessment module. The main contributions of the present work are thus as follow:

1. Study of the state-of-the-art methods on view classification and quality assessment of ultrasound videos, namely from echocardiography and FoCUS, along with the challenges and difficulties present in these images;
2. Preparation and curation of a FoCUS dataset annotated for multiple tasks;
3. Development of a deep learning-based classifier that automates view identification of FoCUS videos;
4. Development of a set of deep learning-based classifiers that automate a view-specific quality assessment of multiple attributes;
5. Validation of the developed algorithms.

We expect this proposal to be a first step towards helping practitioners perform higher-quality exams and, ultimately, improve patients' care.

1.5 Thesis Overview

In the present chapter, the clinical context of the present work was introduced. First, a description of the cardiovascular anatomy and the various imaging modalities useful for its visualisation and assessment were given. Then, an overview of what FoCUS is, and its clinical significance, was provided, along with a description of how ultrasound works and the type of cardiac windows generally acquired in these examinations. Finally, the motivation and aims of this work were presented. The following chapters are dedicated to the further development of the presented topics.

The second chapter aims to present the state-of-the-art on this topic. First, deep learning is introduced, and its basic concepts explained. Then, two specific classes of deep learning models are described, together with their most common applications. Next, previous works from the literature related to the dissertation topic and based on deep learning techniques are presented. Finally, the evaluation metrics used to assess the performance of the developed methods are introduced.

In the third chapter, the view classification framework is described. This includes the presentation of the dataset utilised, the developed methods, a study of different methods to leverage of the temporal information within FoCUS videos, and the validation of the proposed view classification algorithm.

The fourth chapter is dedicated to the quality assessment framework. The dataset and respective annotations are presented, followed by the methods implemented and the results achieved. Some experiments which try to mitigate the dataset's limitations are also presented.

Finally, the fifth chapter concludes the present thesis and discusses possible future developments.

2. STATE-OF-THE-ART

This chapter presents some background about deep learning and its basic concepts, and then summarises previous literature works relevant to the dissertation topic, as well as relevant evaluation metrics.

2.1 DEEP LEARNING

Artificial intelligence (AI) is a technique that allows a machine to mimic human behaviour in the sense that it observes its environment and makes decisions to achieve its goals. A branch of AI is the so-called machine learning (ML), which is a technique that achieves artificial intelligence through algorithms trained with data. Lastly, deep learning (DL) is a type of machine learning that maximises the success of the learning process through its ability to self-learn, with models' architectures inspired by the structure of the human brain (*i.e.* a network of neurons). DL has been successfully applied in many fields, such as computer vision, speech recognition or medical image analysis, being considered a state-of-the-art tool for automatic analysis tasks [31, 32].

Suppose the task at hand is to distinguish two objects. If machine learning would be used, the human would have to design the features by which the two can be distinguished and then feed them to the machine, which would learn how to distinguish the objects based on the provided features. In deep learning, on the other hand, the features are selected by the algorithm itself, *i.e.* it learns relevant features and the prediction model at the same time without human intervention, which is often referred to as end-to-end learning. This autonomy has the downside of requiring a much larger amount of data to effectively train the machine [31, 33, 34]. Furthermore, DL-based methods are usually evaluated, and proven effective, on larger datasets compared to traditional ML-based methods, which indicates better generalisability and robustness to data variations.

Deep learning has a wide range of applications, but again there are some limitations. The first, as mentioned above, is the huge amount of data required to train these algorithms. Even assuming access to the required amount of data, processing this is not within the capability of every machine. This describes the second limitation: computational power. Indeed, training these networks requires expensive graphical processing units with thousands of cores. Another limitation comes from the training time. This time increases with the amount of data and the number of parameters in the model, potentially taking hours or even days. Finally, while traditional ML methods use hand-crafted features, which makes them

more comprehensible, deep learning models suffer from lack of interpretability and transparency, as they generate complex and abstract features [32].

To assess the performance of these models, it is recommended to divide the data into training, validation, and test sets. The training set is used to make the model learn about the data. The validation set is used to determine the reliability of the learning results and make optimisation decisions (such as selection of hyperparameters or other algorithmic choices). The test set assesses the generalisability of a trained model to data that the model has never seen. When the number of training samples is limited, k-fold cross validation approaches are used in which the data are divided into random groups of equal size. The training process is run k times, each with one group used for performance assessment and the rest for training. Then, the results over all folds are aggregated into a final performance metric.

The training process consists of repeatedly performing the task and adjusting the model at each repetition to improve the result. This process is often performed in a supervised manner, *i.e.* it involves ground truth labels for all input data and a loss function which is minimised iteratively over the training samples. Supervised learning is the most common training approach but requires a laborious generation of labels, as an output is needed for every training example [32]. Supervised models can be divided into two types of problems: classification models are used to predict (or classify) discrete values, such as gender (male or female), while regression models predict continuous numerical values, such as a price or an age. Classification problems can be further divided into two types: multi-class, where the classes to predict are mutually exclusive, so each sample is assigned to only one label; and multi-label, where each label represents an independent task, so multiple or none of the labels may be assigned to one sample. At the heart of the most popular deep learning methods nowadays are artificial neural networks (ANNs), which are explained in the following section. The basics of these networks and the necessary concepts to be able to understand the technology are presented.

2.1.1 Artificial Neural Networks

Artificial neural networks (Figure 2.1) are structures consisting of multiple layers of interconnected neurons. These neurons are the core units of the network where the information processing takes place. Following biological neurons, all neurons are connected. First, there is the input layer, which receives the input, and at the end an output layer that predicts the final output. These two layers are called visible layers, and in between are the hidden layers that do most of the computation required by the network. When the input data is an image, each pixel is fed as input to a specific neuron in the first layer. The neurons of one layer are connected to the neurons of the next layer. Each connection

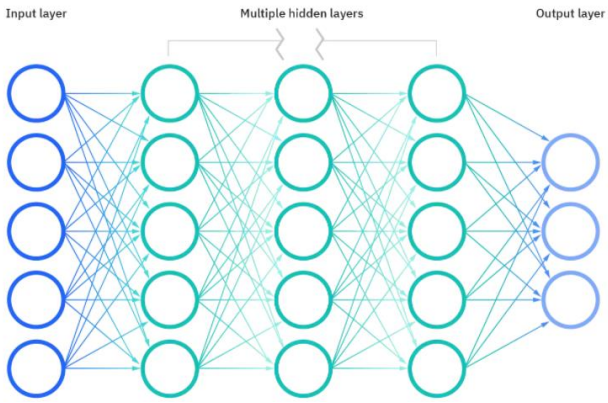


Figure 2.1 - Schematic drawing of an artificial neural network [35].

is assigned a numerical value known as weight. These weights are usually initialised randomly. Each of these neurons encompasses another numerical value called bias. The inputs are multiplied by the corresponding weights, and the bias is then added to the inputs' sum. The resultant value is then passed through a nonlinear function called the activation function, with the resulting activation value being transmitted to the connected neuron in the following layer.

In this way, the data is passed through the network, which is called forward propagation. In the output layer and using a classification problem as example, the values are basically a probability of the label. For multi-class problems the class with the higher probability is selected, whereas in multi-label a threshold is applied to each value to select the predicted labels. In the case of regression, the output is the value itself. Clearly, the neural network can make incorrect predictions, especially at the beginning of the training process. Thus, after each forward propagation, the predicted output is compared to the actual output to detect the error in the prediction through a loss function. This information is then transmitted backwards through the network. This is called backpropagation, and the weights and bias are adjusted to minimise the loss through an optimisation algorithm. The magnitude of this adjustment is controlled by a parameter called learning rate, which can also be changed iteratively by the optimiser. These updates are made after going through a fixed number of samples, called the batch size. This cycle of forward propagation and backpropagation is performed iteratively with multiple input samples. This process continues until the weights are assigned in such a way that the network can make a correct prediction for most samples, or until the learning algorithm has passed through the entire training dataset a predetermined number of times, called the number of epochs [32, 33, 35].

During the training process, many decisions must be made, including how to preprocess the data, which network (or architecture) to use, how to optimise the learning process, etc. Even when choosing the network, there are other decisions to be made, like the number of layers (network depth) or the number of neurons in each layer. These decisions are called hyperparameters and there is no

universal theoretical reasoning behind their selection; they are often determined by trial and error. For this reason, the evaluation of the algorithm plays a crucial role. The metric chosen must be relevant to the task at hand, so that it can be generalised to new data [33].

A challenge that is very common when using a limited number of training samples, which often happens in medical imaging, is that models can memorise the training set. This is called overfitting and can be detected by a near perfect accuracy in the training phase and a poor result in the validation (or test) set. In terms of the model, the simplest approach to reduce overfitting is to simplify the model. This works well because the deeper the model, *i.e.* the higher the number of weights, the easier it is for the model to memorise a low number of training samples. Another way to solve the problem is to use regularisation, which attempts to prevent the model from learning too complex relationships within the data. The idea is to add a term to the loss function that penalises the weights and biases. There are two main types of regularisations: L1 – penalises individual values of weights and biases, shrinking them towards 0 (effectively eliminating their significance in the output if equal to 0); and L2 – which instead of penalising the individual values, penalises the sum of the squared weights, forcing their values to be small (but not necessarily 0). Adding dropout layers can be considered another type of regularisation. These layers randomly deactivate neurons on each training iteration. Another approach to combat overfitting is to stop the training when the validation accuracy stops improving, which is called early stopping. Batch normalisation is also effective as it standardises the layer's inputs for each batch of samples [31, 33]. Despite this myriad of options, the simplest solution would be to gather more training data. Since this is often difficult, an alternative approach is to create new samples by slightly modifying the existing dataset. Using once more an image dataset as an example, one may flip, translate, rotate, or modify the intensity values of each sample randomly within a pre-defined reasonable interval. In this way, during training, the network faces an (artificially) enlarged dataset, with increased variability, which helps improve its generalisation to new data [33].

In this section, a simple type of deep neural network has been described. However, there are more complex architectures, which are often tailed towards certain problems or datasets. When dealing with images, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are two of the most popular architectures, and thus a brief overview is given in the next sections.

2.1.1.1 Convolutional Neural Networks

Nowadays, CNNs are the state-of-the-art method for image analysis, and an example is represented in Figure 2.2. This type of network has the advantage that the number of weights can be

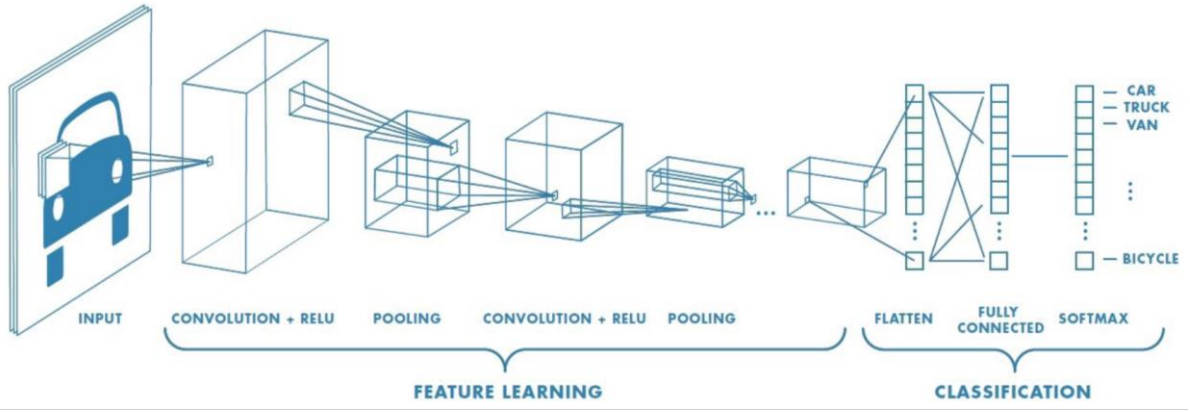


Figure 2.2 - Example of a convolutional neural network [37].

significantly reduced when compared to a conventional ANN, which in turn reduces the required computational power. This is achieved by applying a convolution operation that implies the sharing of the weights at each image pixel [27, 28, 31, 32]. The architecture is also inspired by the human brain, in particular the visual cortex, where neurons respond to signals from several overlapping regions that together cover the whole visual area [34]. There are five main types of layers: the convolutional layer, the pooling layer, the fully connected layer (FC), the activation layer, and the normalisation layer. The first two can be easily intercalated and repeated throughout the network, but the fully connected layer is usually located at the last layers. The activation layers allow the network to learn complex data relations by employing nonlinear functions. The normalisation layer, as the name implies, normalises its input so that the network can be unbiased to high value features; in this way, weights do not vary as much which allows faster optimisation.

The **convolutional layer** is the main block of the CNN, being responsible for feature extraction. For the computation of the convolution, a filter (or kernel) is needed. This filter represents the operation being applied to the area of the image that is being analysed at a given time. This filter is moved across the image to compute the respective feature values. The filter consists of a 2D array of weights that can vary in size. The filter is applied to the image by summing the product of its weights with the corresponding pixels in the image (Figure 2.3). The result is fed to the output and the process is repeated throughout the image until the filter has passed through all pixels and a feature map has been created. The weights of the filter do not change across the image and this is the reason why CNNs have fewer parameters. If the input image has multiple channels, like in RGB images (or at the middle layers of the CNN), then the filter has the same depth. The same principle is applied if the input is 3D-shapped [27, 31, 36].

For each convolutional block implemented, apart from the kernel size, three hyperparameters must be defined, namely the depth, the stride, and the padding. The first parameter corresponds to the desired depth of the output, which is controlled by the number of distinct filters applied. The stride controls

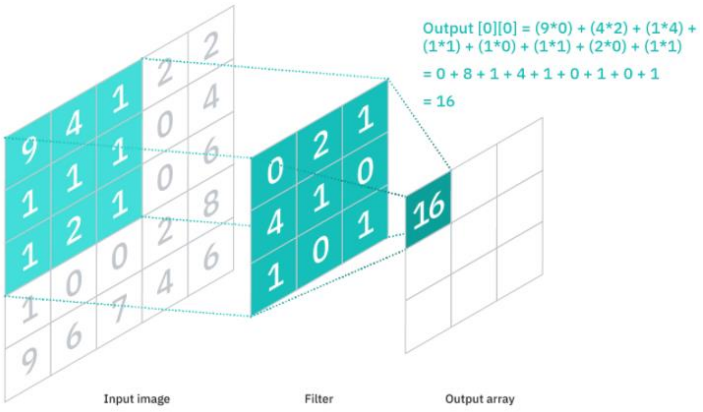


Figure 2.3 - Example computation of a convolution operation [36].

the number of pixels the filter skips as it moves across the input image. The higher the stride, the smaller the output size. Even at the smallest stride, the convolutional process results in a reduction of dimensions. The third hyperparameter, padding, addresses this issue. When the padding is set to *valid*, no action is taken. Alternatively, *same* padding sets elements around the input image to 0 to ensure that the computed output has the same size as the input. If *full* padding is selected, the zeroed elements are added around the resulting feature map, after reduction [31, 36].

The **pooling layer** is responsible for reducing the dimension of the feature map. This layer is similar to the previous one in that it sends its input through a 'filter' and is defined by the same hyperparameters. In this case, however, the filter has no weights, but it combines the values of the input within the range of the selected filter. If the pooling is set to *Max*, the output takes the maximum value of the input within the filter range. On the other hand, if set to *Average*, the average value within that range is used. This type of layer is the main responsible for the dimension reduction happening throughout the network and helps reduce the possibility of overfitting [31, 33, 36].

The **fully connected layer** is like a conventional ANN layer where the output of the previous layers is flattened and connected to each output neuron. The purpose of this layer is to combine all the information from the gathered features and prepare the final prediction [31, 33, 36, 37].

The **activation layer** transforms the weighted sum of the inputs into a value to be fed to the next layer, so in this way they define which neuron is fired of not. There are multiple possible functions with different characteristics. The ones relevant to the present work are present in Figure 2.4. Sigmoid constrains the values to a range of 0 to 1, where the larger the input the closer to 1 and vice versa, which expresses the probability of the class. Softmax is a combination of multiple sigmoids, allowing relative probabilities, *i.e.* it forces the sum of the probabilities of the output classes to be equal to one (which is necessary in multi-class problems). The Rectified Linear Unit (ReLU) follows a linear function but only if

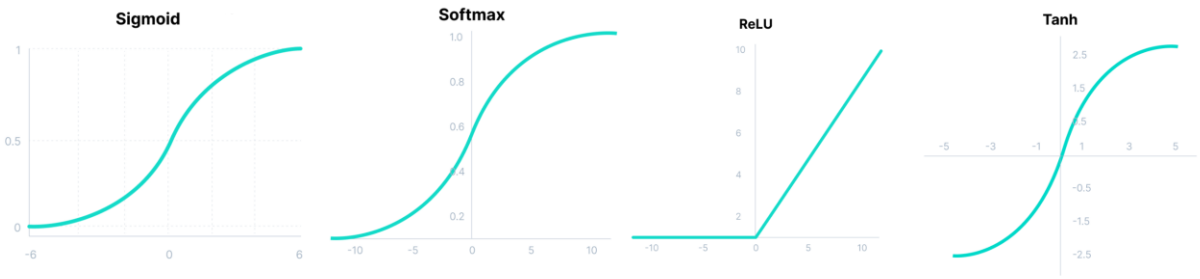


Figure 2.4 - Sigmoid, Softmax, ReLU and Tanh activation functions [38].

the input is higher than zero. With similar shape to the first two presented, Tanh is zero-centered: the larger the input the closer the output to 1, while smaller inputs will generate outputs closer to -1.0. [38].

The **normalisation layer**, as previously stated, scales the features to similar intervals to reduce the bias in the network, and in doing so it regularises the network, preventing overfitting. There are several techniques to apply normalisation, but the most common is Batch Normalisation (BN). BN computes the mean and variance of the features in the batch, and then subtracts the mean to each feature and divides it by the computed standard deviation. This stabilises the learning process allowing faster training [39].

Different arrangements of these layers have led to the proposal of different architectures over the years. There are various architectures that have proven successful in different image classification tasks, and they are constantly evolving. The trend is for models to become deeper and more complex. The best known are: AlexNet [40], which is one of the pioneer works in CNNs; VGGNet [41], a deeper network but with a simple arrangement of convolutional and pooling layers; ResNet [42], which proposed residual connections (also termed skip connections or shortcuts) between layers to prevent overfitting but also mitigate the vanishing of gradients during backpropagation; DenseNet [43], that introduced dense connections to improve feature propagation; and Inception [44], also known as GoogleNet, a complex model that incorporates different sizes of filters at the same layer [45].

ResNet

ResNet stands for Residual Network, and it is a CNN first introduced in 2015 that gained popularity as it came first in several competitions in that same year. The proposal was to use skip connections, *i.e.* to connect activations of a layer to another while skipping one or more layers, forming residual blocks as shown in Figure 2.5. ResNets are made by stacking blocks like this together, and there are multiple variants of the network with different number of layers and blocks. Each variation is determined by the number of layers, which is added to the main name; for example, ResNet-18 has 18 layers. The architectures of the more popular ResNets are summarised in Table 2.1. ResNet-34 was the

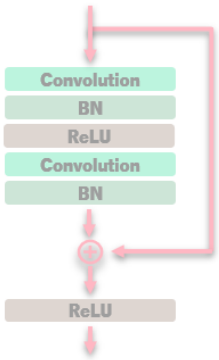


Figure 2.5 - Residual Block.

Table 2.1 - Architectures of ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-101 and ResNet-152 [42]

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				

first one to be proposed and is based on the VGG network. In deeper versions, the 2-layer residual block is replaced with a 3-layer version in a bottleneck design, reducing training time [42, 46].

2.1.1.2 Recurrent Neural Networks

RNNs are suitable for sequential or time series data. Their main feature is their capability of memory, as they take information from prior inputs combining it with the current input to influence the output. However, as the length of the sequential data grows, and with it the gap between past relevant information and the current output, RNN become unable to learn. Long Short-Term Memory (LSTM) networks were developed to address this problem. The difference lies in the presence of gates. This mechanism regulates the flow of information and learns which data in a sequence is important to keep or forget. In this way, it learns to use only the information relevant to make predictions [29, 31, 33].

2.2 Cardiac View Classification

View classification is the attribution of cardiac view labels to echo images, being an important step in any automated echocardiographic image analysis pipeline [47, 48]. The more numerous classes of views, the more difficult it is for a model to distinguish them. This happens because, as explained in Section 1.2.2, the difference between some cardiac views is a result of slight changes in the transducer pose, which makes them very similar in appearance. Because of that similarity, some researchers prefer to overlook their particularities and group them into the same super-class, easing their classification. Another challenge lies in the large intra-class variability and low inter-class variability inherent to the different ultrasound cardiac views [49, 50]. Previous studies have shown that DL methods have a higher versatility of training that represents a substantial gain over ML methods. Therefore, this section focuses on works that use deep neural networks [50]. In this analysis, state-of-the-art methods on cardiac view identification have been separated on those that classify only individual frames of a video and those that

consider the full video and its inherent temporal information. Moreover, unless explicitly stated, all described methods target conventional echocardiography.

2.2.1 Frame-Based View Classification

Most of the existing works use CNNs because they are extremely effective in learning patterns and features from images. Most of them have implemented well-established networks, or variations of them, as they have proven to be successful in multiple image classification tasks.

The network most often used, among the papers in study, is VGG-16. Blaivas *et al.* [51] compiled 750 thousand images of FoCUS examinations and trained and tested several networks, namely AlexNet, VGG-16, ResNet, DenseNet201, and Inception V4, for the task of classifying 5 views. The results showed a tendency for more modern and deeper models to perform worse than older and shallower ones, while VGG-16 showed the best performance. Similarly, Madani *et al.* [50] used a VGG-based method to discriminate 15 different echocardiographic views (Figure 2.6). They reported an average 91.7% accuracy in classifying single images, compared with 79.4% for echocardiographers classifying a subset of the same images. However, the experts were presented with the same down-sampled images that were fed into the CNN model, which partly explains the high discrepancy. Later, they presented an improved classification pipeline that achieved an accuracy of 93.64% by first applying a segmentation step in which the scan section was isolated from the text annotations in the images using a U-Net model before it was fed into the classification model [52]. This step of isolating the scan has proven successful in several applications, but most of them use simpler image processing methods (like thresholding) to do so [47]. Vaseli *et al.* [53] introduced a lightweight classifier for distinguishing 12 echocardiographic views. Based on three CNN architectures (VGG-16, DenseNet, and ResNet), several lightweight deep learning models were built and trained using knowledge distillation methods. These models were trained and evaluated on 807 thousand images that were resized to 80x80 and randomly augmented using translation, rotation, and up-scaling. Since the dataset had an imbalanced distribution of classes, the algorithm was created so that the randomly selected images in each batch represent a uniform distribution. By combining the output of the three lightweight models (a technique named ensemble modelling), an average accuracy of

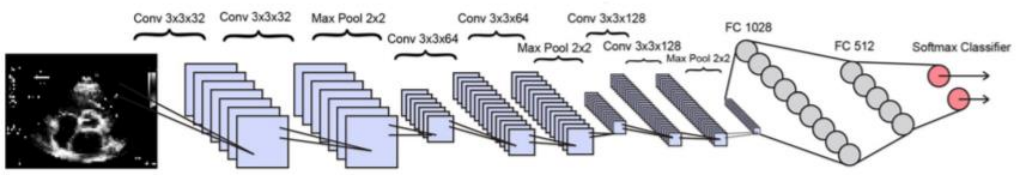


Figure 2.6 - Schematic of the network proposed in [50].

88.1% was achieved, with VGG-16 performing the best when evaluated alone. The lightweight version falls just 1% short of the full version, with only 1% of the parameters.

Ostvik *et al.* [54] proposed a cardiac view classification architecture to distinguish 7 views. Their preprocessing routine includes image normalisation and resizing to 128x128 pixels. The input image is then sent to an AlexNet and Inception-based CNN, trained with 265,649 images. Like in the above work, the dataset was imbalanced. Hereto, the authors opted to select images so that in each epoch the distribution of each class was similar, achieving 98.3% accuracy. To tackle the same issue, Chartsias *et al.* [55] implemented a contrastive learning approach for training a 5-block VGG-like CNN with the goal of classifying 13 cardiac views. Data augmentation was also applied, including brightness changes, contrast variations, 30° rotations, and translations. Preprocessing included resizing to 192x192 pixels, isolating the scan sector, and intensity rescaling to [0,1]. This contrastive learning technique proved beneficial against its baseline, and against undersampling the dataset to make all classes equally represented.

A classifier for 10 cardiac views was created by Gungor *et al.* [48] using the InceptionV3 network, which achieved an overall accuracy of 97.62%. The dataset consisted of 11 thousand videos, in a total 160 thousand images with an imbalanced distribution of classes. The novelty here was the transformation applied to each image. Using the metadata stored in the DICOM, the coordinates of the cone shaped scan were transformed from Cartesian coordinates to polar ones. Then, all images were downsampled to 256x256 pixels, the mean value subtracted, and the intensities normalised to [0,1]. The applied augmentations were scaling (up to 15%), shear (up to 3%), translation (up to 15%) and rotation (up to 10°) transformations, plus contrast changes (from -100 to 40).

Kusunose *et al.* [56] tested two types of input techniques in a 5-layer CNN to classify 5 cardiac views. The dataset was composed of 17 thousand images from echocardiograms, resulting of extracting 10 equally spaced images per video, with consideration of the different frame rates and heart rates. The two methods tested were: averaging the 10 frames of a video and provide it as input to the network; or input each one of the 10 frames individually, and average their predictions. The latter method proved more effective, achieving an overall accuracy of 98.1%.

2.2.2 Video-Based View Classification

The works presented previously fail to take advantage of the temporal information contained within the echocardiographic videos, namely regarding moving structures during the cardiac cycle. Spatio-temporal networks have been shown to have significant performance improvements over spatial-feature-based baselines for the task at hand [34, 57]. Here, some of these architectures are presented, most of

which are inspired by work in the field of human action recognition, where both spatial and temporal information contained in videos is accounted for [57, 58].

Howard *et al.* [57] created a dataset by extracting 40 frames per video from 8 thousand sequences and tested different classification models. These authors compared a frame-based model with various video-based ones, using different techniques such as: passing the frames sequentially through a CNN, and the resulting feature maps through a RNN or other neural network to extract temporal features; use a 3D CNN; or employ a two-stream CNN network, where one stream receives the image and the other processes the video's optical flow (Figure 2.7). They proved the usefulness of including spatio-temporal features, with the two-stream network giving the best result, followed by the sequential technique. Ye *et al.* [58] conducted a similar study to prove the benefit of taking advantage of the temporal information of an echocardiogram video to classify, in their case, 9 possible views. With a dataset of 2,693 videos, and upon isolating the scan sector, they first determined Xception to be the best CNN to extract spatial features. Then, they tested two methods of utilising the information of the moving frames. The first was to use sequentially the Xception network and a bidirectional LSTM. The second was to use a two-stream version of the first method, where one receives the original frames and the other the optical flow's result. The first method achieved the best performance, with an overall accuracy of 94.3%.

In Gao *et al.* [34], two CNNs were combined along the two directions of space and time to classify eight different views. The spatial CNN takes a 227 × 227 image as input and extracts spatial features. The temporal CNN takes as input the acceleration image, resulting from applying the optical flow twice, and extracts feature maps from it. The final classification is obtained by combining the two sets of features. They used a dataset of 432 image sequences and achieved an accuracy of 92.1%.

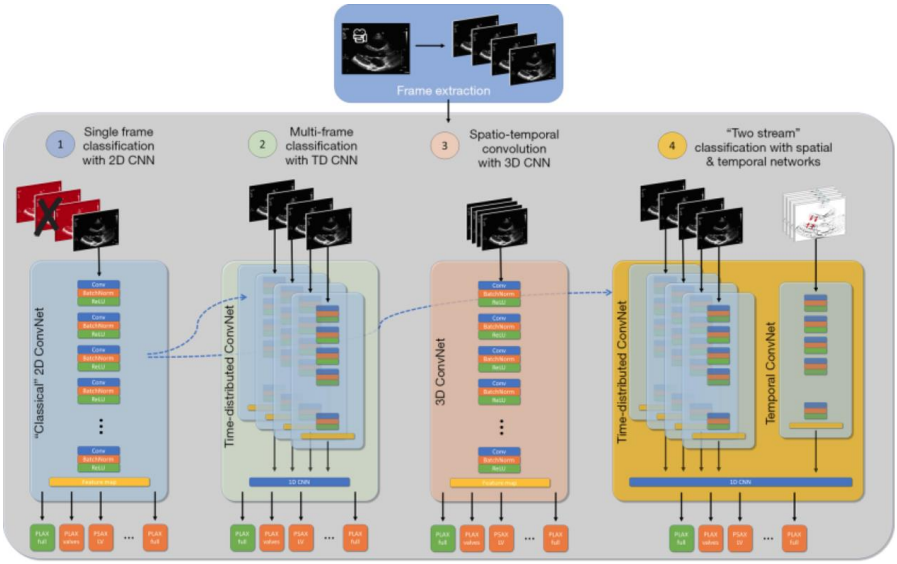


Figure 2.7 - Schematics of the networks used in [57].

Similarly, Shanin *et al.* [59] also proposed a two-stream model, employing a pretrained ResNet model to extract spatial features and extracting temporal features based on neutrosophic set theory. These features are then concatenated and fed into a trainable LSTM-based architecture. Their model achieved an average accuracy of 96.3% in the same dataset as [34].

In Zhu *et al.* [60], a sequential approach was followed. First, a ResNet-50 extracts spatial feature maps. Then, the feature maps of all frames are fed into a LSTM that temporally aggregates them. Finally, the aggregated features are used to predict one of 8 classes, achieving an accuracy of 98.8%. The models were pre-trained with ImageNet and fine-tuned with 3 thousand videos with 60 frames each. Each frame had its intensities normalised between [-1, 1]. The augmentation techniques employed included rotations, translations, flips, and random cropping.

2.3 Quality Assessment of Ultrasound Images

The quality of the acquired ultrasound images directly affects their analysis, measurements done and consequently the diagnosis made [32]. Therefore, the task of assessing image quality is an important step in the automation of ultrasound analysis, and various approaches can be found in the literature. Automated assessment of the quality of an echo image may provide a (discrete or continuous) quality score or categorise that image as being of poor or good quality, which represent a regression or classification problem, respectively. The different approaches can be further divided into model-based methods and DL-based methods. Since model-based methods are view-specific (they require a specific model or template for each view) and fail when applied to low contrast images [47], this section focuses on DL-based models.

Labs *et al.* [61] trained a multi-stream regression model with CNN and LSTM layers to assess the quality of apical four chamber view images for four proposed attributes (*i.e.* foreshortening, gain/contrast, time-gain compensation, and interventricular orientation), assigning an independent score for each. Four CNNs arranged in parallel were trained simultaneously on 20,780 images, with each stream containing slightly different layers, created specifically for each attribute. This regression model achieved an average accuracy of 86% in the test set. They later applied the same approach to a dataset of 33,784 frames of both apical four chambers and parasternal long axis, with similar attributes [62].

Various approaches were presented by Abdi *et al.* First, they proposed a two-layer regression CNN architecture to assess the quality of end-systolic, apical four-chamber frames [27]. The model was trained end-to-end with 2,344 images, with integer scores ranging from 0 to 5. Later, these authors improved their model by increasing the amount of training data and by applying a method of

hyperparameter optimisation called Particle Swarm Optimisation [28]. However, these methods did not fully exploit the available information. In [29], they extend their work to include other views, as well as process whole videos instead of frames only. The proposed architecture of the multi-stream regression network consisted of five models, for five views, with the same weights in the first layers, while the last layers were view-specific. Like in [61], each stream consisted of CNN and LSTM modules. This method was trained on 4,675 cine loops of 20 frames each and achieved a prediction accuracy of 86%.

Luong *et al.* [63] studied the images of echocardiograms in hospitalised, mechanically ventilated patients. Their method of quality assessment was similar to Abdi's in the sense that the model produced a single score per image (in this case from 0 to 1). Like other video-based models presented above, the DL framework was a composition of a CNN and a LSTM, with the former employing a DenseNet architecture (Figure 2.8). The individual frame scores of each video were averaged to obtain an aggregate prediction for the whole video. The dataset consisted of 14,086 echo video clips, normalised to [0,1], and the overall accuracy was 87.0%.

In the field of fetal ultrasound, Dong *et al.* [45] have proposed a general quality control framework for the apical four-chamber view. Their proposed framework consists of three networks with the following tasks: identification of the four-chamber view images in the raw data; determination of the images' gain and zoom; and detection of the required anatomical structures in the view in question. For the selected images in the first step, a DenseNet-161 was set up to classify the images into three classes (low, proper, and high) in terms of gain and zoom in a multitask learning procedure. This framework attained 99.5% and 98.8% of accuracy for each of the task, respectively). Finally, a detection network called Aggregated Residual Visual Block Net was proposed to recognise the anatomical references that must be present, calculating a bounding box to encompass each structure. The output of the three networks provided the overall quantitative score of each image.

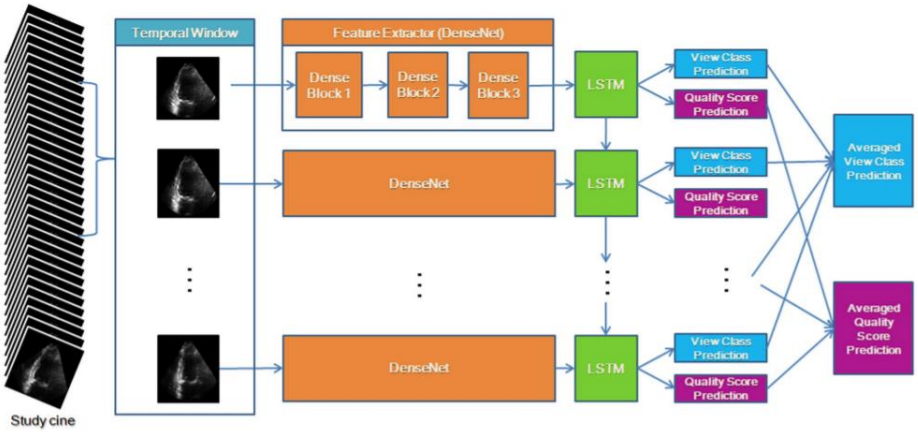


Figure 2.8 - Schematic of the network proposed in [63].

2.4 Evaluation Metrics

Multiple metrics exist to evaluate the performance of a model. In classification problems, all of them are based on the confusion matrix. This matrix is computed by comparing the model's class prediction with the ground truth annotation, counting the number of correct and incorrect predictions by class. An example matrix for two classes is presented in Figure 2.9 [64]. It should be pointed that the matrix can be computed for multi-class problems, however the values of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are counted in a per-class manner. These four values are the basis for the calculation of the metrics described below.

The selection of the metric to evaluate different models is of extreme importance and depends largely on particularities of the problem at hand. Maier-Hein *et al.* [65] presented a framework to guide the decision-making regarding metrics in biomedical image analysis problems. Following their guideline, for the present thesis, the recommendation is to use three types of metrics: a multi-class metric, which considers the performance of the model for all classes in one single value; a per-class metric to assess the performance for each class; and a multi-threshold metric, which works on a dynamic confusion matrix by varying the threshold that determines if a given class is either positive or negative.

For the first type, the gold standard is Matthews Correlation Coefficient (MCC). MCC measures the correlation between the true class and the predicted one and it is a balanced measure even in the presence of an imbalanced distribution of classes. In the binary case, the MCC is defined as in equation (1), and it ranges from -1 to +1, where 1 represents a perfect prediction, 0 a random classification, and -1 a perfect negative correlation. It has also the property of symmetry, resulting in the same score even if positive and negative classes are inverted. In the multiclass case, the MCC is defined as in equation (2) and the minimal value is no longer -1, as it can vary depending on the class distribution [64, 65].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (1)$$

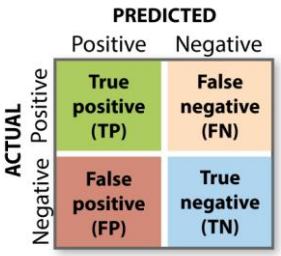


Figure 2.9 - Structure of a confusion matrix [64].

$t_k = number\ of\ times\ class\ K\ ocurred$
$p_k = number\ of\ times\ class\ K\ was\ predicted$
$c = number\ of\ samples\ correctly\ predicted$
$s = number\ of\ samples$

$$MCC = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}} \quad (2)$$

F1 is one of the most important per-class metrics. It combines two opposing metrics - precision and recall. Precision represents the fraction of TP in all positive predictions, while recall represents the fraction of TP in all positive cases. Importantly, there is a trade-off between these metrics, and often improving one means a reduction on the other, or vice versa. Hence the F1 score, the harmonic mean of precision and recall, combining them into a single value while penalising extreme values of either metric. This metric is computed using three of the four values in the confusion matrix and is defined as in equation (3). Its value can range from 0 (worst case) to 1 (perfect prediction). To compute the per-class F1 in a multi-class problem, each class is assessed in a one-versus-the-rest manner [64, 65].

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (3)$$

Multi-threshold metrics calculate metric scores based on multiple thresholds. In this regard, the most used metric is the Area Under the Curve (AUC), which represents the measure of separability of the receiver operating characteristic (ROC) curve *i.e.* the capability of the model to distinguish the classes. This curve is computed by calculating both recall and specificity (fraction of TN in all negative cases) for multiple threshold values, each representing a point of the curve (Figure 2.10). Like the previously presented metric, when in multi-class problems, the computation is adapted to a one-versus-the-rest manner [64, 65].

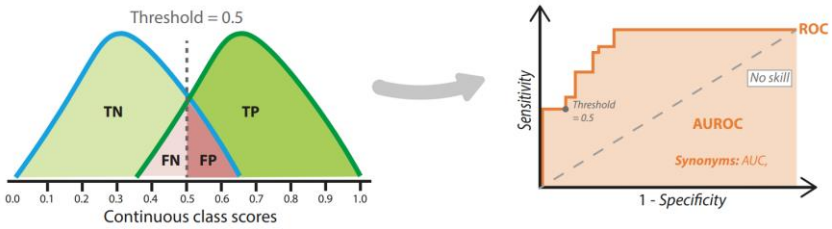


Figure 2.10 - Representation of the AUC computation [64].

3. Automatic Classification of FoCUS Views

For any automated processing of FoCUS videos, the first step is the identification of the imaged cardiac view. So far, most state-of-the-art algorithms for this task were trained on conventional echocardiogram images, which have higher image quality and far less instability than the ones acquired in FoCUS settings. In this chapter, the development of a deep learning framework to automatically classify the cardiac views in FoCUS videos is presented. First, the dataset used to train the models is introduced. Then, a video-based model for view classification is proposed, exploiting a multi-frame input to combat frame-level variations. Finally, different methods to leverage spatio-temporal information of FoCUS video clips in the proposed model are investigated and compared.

3.1 Dataset

3.1.1 General Description

The FoCUS dataset used in this work was collected from the *Hospital de Clínicas de Porto Alegre* (HCPA, Brazil) with the ethical approval of the Ethics Committee for Research (ECR) in Life and Health Sciences of the University of Minho (CEICVS 039/202) and the ECR of the HCPA (5.334.879). The collected data consists of saved clips from examinations performed by residents in FoCUS training from 2020 to 2022. Each study has multiple videos in MP4 format with various durations and frame rates, each corresponding to one of the cardiac views described in Section 1.2.2. An exam does not necessarily consist of one video per view, but may include, for example, multiple videos from one view and none from another. These clinical exams encompass different pathological conditions and were performed for different clinical reasons and with different ultrasound machines, so they are representative of the clinical practice. All identifiable information in the videos was anonymised.

3.1.2 Data Preparation

As described in the literature, to obtain DL models with higher generalisation capabilities, it is advantageous to isolate the scan sector of the image, erasing all auxiliary text and information present

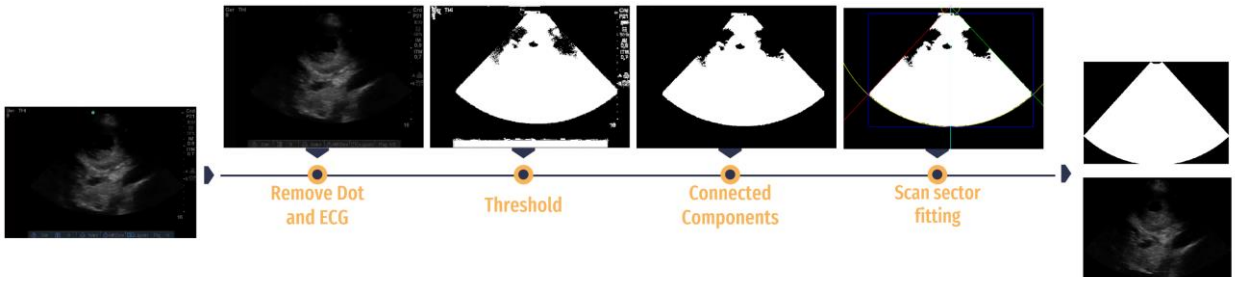


Figure 3.1 - Routine steps to isolate the scan sector.

around it. To accomplish this, a preprocessing routine was developed using MATLAB (MathWorks Inc, USA), involving several steps (Figure 3.1).

As mentioned earlier, the ultrasound probe has a mark, visible on the screen, that indicates the orientation of the probe relative to the patient's head. This mark appears as a coloured dot, often in green and most commonly on the image's left side (as per the acquisition guidelines). The first step was to detect this dot by converting the image to the HSV colour space and filter any pixel having the possible colours for the dot: green, orange, or blue. This was achieved by defining a set of threshold limits (per HSV channel) for each colour separately. Any identified area is then erased. Note that, if one of these colours is detected on the image's right side, there is the possibility that the video was acquired in an inverted orientation with respect to the acquisition guidelines. In this scenario, a warning is given to trigger a visual inspection of the video and, if confirmed, the respective label ('Inverted') is linked to the video.

In some videos, the ECG line was projected at the bottom of the image, and it must also be removed. For this purpose, like the method above, the ECG line is considered present if there is a large object with its characteristic colour in the screen's lower section. In this case, instead of deleting the thresholded pixels, the highest point of the ECG mark (the point with the lower Y-coordinate) determines where to crop the image (keeping the upper portion only), removing the ECG region from the video.

The next step creates a first draft of the mask. First, a threshold is applied to select all pixels with an intensity above 2, and then all frames are summed into one image. This results in a binary image where pixels with a value of 0 correspond to those that always had values below 2 throughout the entire video. The connected components are calculated, and all elements with an area above 20 000 are kept.

After this operation, the scan is usually isolated, but the mask may not perfectly fit the shape of the scan sector. To correct this, the first points of the binary mask from each side are detected and two lines are fitted to these points (one per side). Since the sector is symmetrical, the line that forms the wider angle with the vertical axis is assumed to be the correct one, and it is mirrored to form the other. The lateral sides of the sector are now defined, leaving the upper and lower parts which follow a circular shape. A circumference is defined by its centre and its radius. The centre here is defined by the

intersection of the lateral lines. To calculate the radii, first a bounding box of the scan sector mask initially detected through thresholding is defined and its intersection with the lateral lines results in an upper and a lower point. The distance from these points to the centre point defines both radii. The sector is now fully defined, and the final mask is obtained. The mask is applied to all frames of a video, successfully isolating the scan sector.

3.1.3 Annotation

The annotation of the view type of each video was made by an internal medicine physician proficient in point-of-care ultrasound. Each video was classified as one the seven views presented in Section 1.2.2. Dubious videos, acquired 'between' standard views or with extreme low quality, were flagged and revised later, aiming to reduce inevitable human errors. Part of the dataset was also given a grade regarding the overall quality of the video in a range from 1 to 5, where 5 represents excellent quality, 3 represents a sub-optimal video (either in terms of image gain/depth or by the absence of some anatomical features) but still interpretable, and 1 a very poor quality video where any assessment is untrustworthy.

3.1.4 Division into Sets

A total of 713 exams were gathered, which correspond to 4,029 videos. The exams were randomly split into 6 groups of equal proportion, maintaining sample independence by ensuring that all videos of each exam fall within the same group. One set of 119 exams was used as a held-out test set, and the other 5 groups were used for training and validation in a 5-fold cross validation method. In sum, 3,414 videos were included in the training/validation set and 615 in the test set. The relative distribution of each view is shown in Figure 3.2, where a smaller prevalence of the short-axis views is noticeable. On a later stage, a new set of videos were made available (928 videos from 170 exams), which were used as a secondary test set.

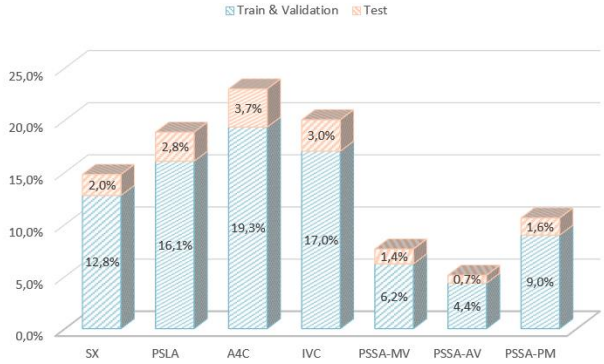


Figure 3.2 - Relative distribution of videos for every view in training/validation and test sets.

3.2 Video-Based Classification using Multi-Frame CNN

3.2.1 Methods

This section presents a method to automatically classify which cardiac view a FoCUS video belongs to, whose approach is based on four components: (1) a preprocessing routine that standardises each video (Section 3.2.1.1); (2) a CNN that takes multiple frames of the video as input, processing them as a whole (Section 3.2.1.2); (3) a training scheme with a more flexible loss function and random augmentations that integrate domain knowledge (Section 3.2.1.3); and (4) an inference strategy that allows aggregation of multiple clips of the original video to achieve a video-level prediction (Section 3.2.1.4).

3.2.1.1 Data Preprocessing

First, each video is downsampled to f frames per second (FPS), where f is set to 10. Then, a clip is created by randomly selecting 32 consecutive frames. If there are not enough frames, empty frames are added. The resulting clip (and its scan sector mask) are padded along the shorter axis to achieve a square aspect ratio and resized to 224x224 using bilinear interpolation. Finally, the clip is converted to grayscale and the pixel intensities divided by 255 (to normalise to the range of [0,1]).

3.2.1.2 Network Architecture

The proposed architecture is depicted in Figure 3.3. In this design, each frame of the input video clip is passed through a shared spatial feature extractor based on the ResNet-18 [42]. The resulting feature maps are subsequently combined through a global average pooling layer, and then recombined into 128 final features for class prediction. This spatial feature extractor also introduces a component presented by Park *et al.* [66], the blur block. Taking advantage of the spatial consistency of each image by aggregating adjacent feature map points, they showed the blur block stabilises the resulting feature maps and improves robustness. To achieve this, first a combination of the activations tanh and ReLU transform each point into a probability, and then an average pooling layer with a kernel size of 2 and a stride of 1 is applied to perform the blur operation (Figure 3.3C). The tanh activation was applied in a temperature-scaled manner, *i.e.* following equation (4) where τ represents the temperature hyperparameter (empirically set to 10). Although the authors in [66] suggest adding this block after every downsampling layer, we propose to add it after every residual block. The proposed model has a total of 11,246,784 parameters.

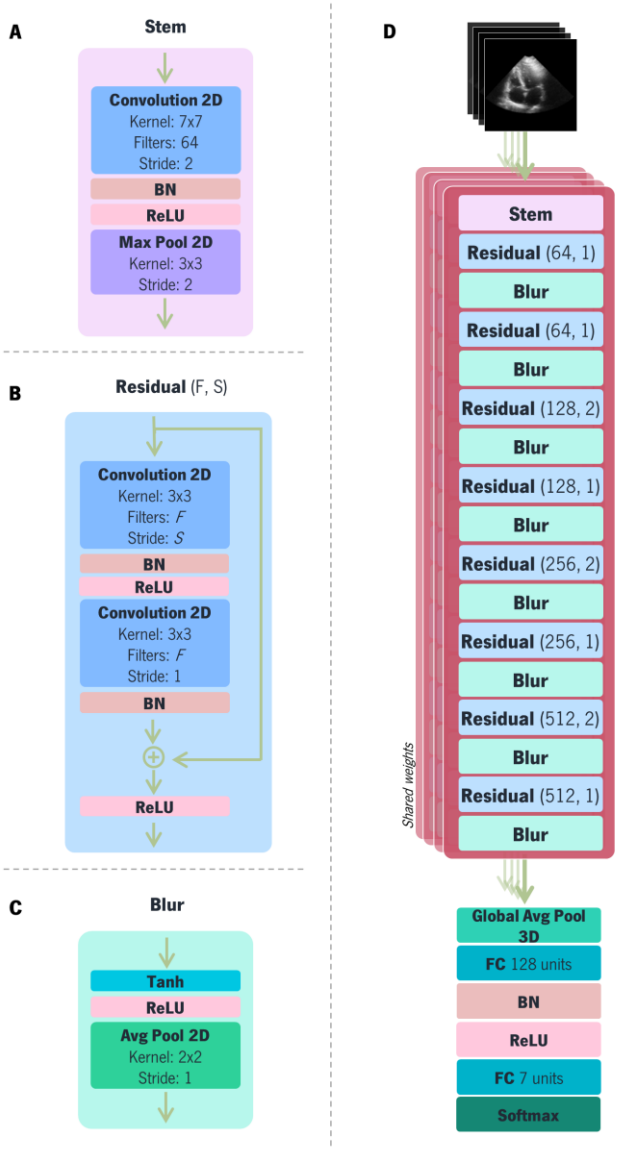


Figure 3.3 - Illustration of the proposed network.

BN – Batch Normalisation; FC – Fully Connected.

$$tanh_{\tau}(z) = \tau \times tanh\,(z/\tau) \tag{4}$$

3.2.1.3 Model Training

The model was trained for 75 epochs with a batch size of 8. The Adam optimiser [67] was used with an initial learning rate set to $1{\times}10^{-3}$, updated using a cosine decay schedule [68]. Weights in the convolutional and fully connected layers were initialised with the normal distribution proposed in [69]. The PolyLoss [70] was employed as loss function. In [70], by leveraging the description of the cross-entropy loss as a Taylor expansion, the authors have shown that increasing the first polynomial coefficient of the cross-entropy loss systematically increases a network's performance. This function is defined in

equation (5), where P_t represents the prediction's probability of the target class label, and ε is a tunable parameter (empirically set to 1.5). To combat overfit, L2 regularisation on the network's weights was applied with a weight of 5×10^{-4}.

$$PolyLoss = -\log(P_t) + \varepsilon(1 - P_t) \quad (5)$$

During training, one proposes to apply, on-the-fly, three types of data augmentation techniques (intensity-, spatial- and temporal-based), increasing data variability, regularising the network, and further preventing overfitting. Specifically, intensity-based transformations were applied with a 15% probability and consisted of additive brightness (from -25 to 25, in a scale of [0, 255]), contrast (up to 25%) and gamma correction (in the range of [0.7, 1.5]). The spatial transformations, which were applied with a 50% probability, included scaling (up to 15%), rotations (in the range of [-10°, 10°]) and translations (up to 5% the image's width/height). These ranges and transformations were selected to simulate different settings of the US machine and placements of the transducer. To maintain the scan sector fixed in the centre and correctly oriented (and thus truthful to the US imagery), the sector scan mask is applied after the transformations. These were implemented using the Solt package [71]. Finally, a third type of augmentation is proposed, which works on the temporal dimension and aims to simulate patients with different heart rates. To keep it realistic, the range of variation was chosen considering the typical physiological and pathological values of an adult's cardiac rhythm: a normal rate of 60 bpm, and a value of 40 and 100 bpm for bradycardia and tachycardia, respectively. Assuming a normal rate for the input video clip, this corresponds to sampling it with a frame rate of ωf, with ω being a random value in the range of [0.6, 1.6].

To set up the learning environment, the Keras framework was used with Tensorflow as backend. Experiments were carried out on a workstation with an Intel Core i9-10980XE CPU, a NVIDIA RTX 6000 with 24 GB of VRAM, and 128 GB of RAM.

3.2.1.4 Inference routine

During inference, instead of providing a prediction for the video based on a single video segment, we propose to aggregate predictions of multiple clips spread across the video into a video-level classification. To do so, after downsampling the input video to a frame rate of 10 Hz, clips of 32 consecutive frames are extracted at every half a second (a step of 5 frames). Only complete clips (with the desired length) are considered. In cases where the original video does not meet this minimal length, only one clip is created, and its length is corrected by adding empty frames at the end. The video-level

classification is computed by averaging all clips' prediction vectors and identifying the class with maximum value.

3.2.2 Experiments, Results and Discussion

The performance of the proposed network, evaluated on MCC and per-class F1-score (plus macro- and micro-averages), for the validation set and the two independent test sets is summarised in Table 3.1. The performance metrics of the test sets were calculated in two ways: as the mean of the metrics calculated for each model of the five folds ('Single'); and after model ensembling, *i.e.* after averaging the predictions of the five models ('Ensemble'). The comparison of the validation and test results shows that the model can generalise well. The ensembling technique shows a good improvement over the single model prediction, which demonstrates that the combination of several estimates leads to a better and more reliable final prediction. The ROC of the individual classes for the ensemble results of test set 1 can be found in Figure 3.4. These curves also show good performance in all classes, as they are very close to the upper left corner, which is reflected in a very high value of the corresponding AUC. Both curves and metrics shown in Table 3.1 reveal a lower ability to classify the PSSA views. This sub-optimal performance is due to their visual similarities and their lower prevalence in the dataset. This is even more noticeable for the PSSA-MV and PSSA-PM views, as they represent spatially close anatomical regions and are often mixed up even by experts (potentially even leading to wrong labels). On the other hand, the IVC view achieves a slightly better performance because it is so distinct from the other views.

Ablation Study

A series of experiments were conducted to investigate the value of the algorithmic decisions made. All networks were trained under the same conditions. The 5-fold cross-validation results can be found in Table 3.2. This study shows that the proposed pipeline produces a significant improvement in all metrics over its frame-level counterpart (last row). Note that the frame-level model processes only one

Table 3.1 - Performance of the proposed architecture

		MCC	**F1**							
			SX	PSLA	A4C	IVC	PSSA-MV	PSSA-AV	PSSA-PM	Macro	Micro
Validation		0.9276	0.9715	0.9654	0.9636	0.9898	0.7795	0.8464	0.8607	0.9110	0.9400
Test 1	Single	0.9244	0.9623	0.9679	0.9664	0.9825	0.7760	0.9049	0.8429	0.9147	0.9372
	Ensemble	0.9531	0.9756	0.9870	0.9767	0.9873	0.8542	0.9474	0.9028	0.9473	0.9610
Test 2	Single	0.9279	0.9649	0.9682	0.9751	0.9947	0.7762	0.8928	0.8190	0.9130	0.9401
	Ensemble	0.9403	0.9753	0.9831	0.9818	1.0000	0.7895	0.9296	0.8247	0.9263	0.9504

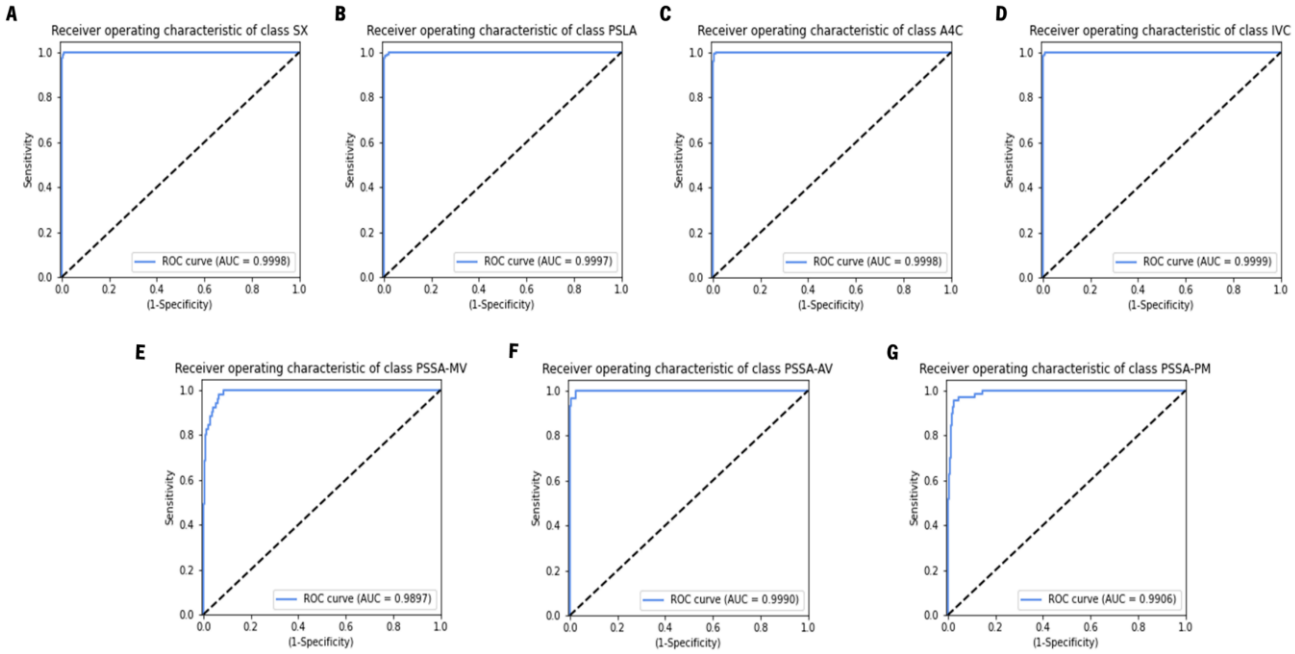


Figure 3.4 - ROC and respective AUC of each class.

Table 3.2 - Ablation study on the proposed methods

	Techniques applied					**MCC**	**F1**	
	Clip-based Input	Temporal Augmentation	PolyLoss	Original Blur	Blur every residual		Macro	Micro
Proposed	**x**	**x**	**x**		**x**	**0.9276**	0.9110	**0.9400**
	x	x			x	0.9265	0.9087	0.9391
	x	x		x		0.9234	0.9076	0.9364
	x	x	x			0.9248	**0.9114**	0.9376
	x	x				0.9192	0.9040	0.9329
	x					0.9184	0.9042	0.9323
						0.9156	0.8949	0.9300

frame at a time (instead of 32 as in the proposed clip-based input) and averages the predictions over all frames during inference to match the averaging of clips' predictions. The results also prove that temporal augmentation increases the model's performance, most probably because adding realistic variability to the training data increases the model's ability to generalise. In pair with the proposed clip-based input, the use of the PolyLoss, in opposition to the commonly used cross-entropy loss, led to the larger relative improvement in MCC and micro-averaged F1-score (and the second largest in macro-averaged F1-score). As for the blur block, its ability to reduce feature map variances and improve classification accuracy has been demonstrated. However, our proposal to implement it at every residual connection was shown to be superior to the original implementation. A synergetic performance is further observed when combined with the PolyLoss.

Inference routine

Another set of experiments was conducted to prove the merits of the proposed inference routine. Table 3.3 shows the 5-fold cross-validation results. Overall, the average prediction over multiple clips proved to be beneficial when compared to a one-clip inference technique, with the latter showing the worst performance among the tested variants. Similarly, processing the whole video at once (taking advantage of the input flexibility in CNNs) also under-performed when compared to the proposed method. However, it is substantially better than the one-clip prediction, which suggests that the assessment of the whole video, by multiple clips or as one, is crucial. As each video comprises one view only through its total duration, it could be expected that only a clip would suffice. But the conditions of the acquisition of FoCUS videos cannot be disregarded. The intrinsic urgency, and sometimes inexperience of the user, result in some instability on the videos, hence the necessity of a full analysis of the video to elude occasional moments of worst quality of the acquisition.

When considering the inference over multiple clips, other options were tested. First, the step at which clips are extracted, and consequentially the number of clips extracted, was manipulated. Then, the frame rate to which each video is downsampled before the clips are extracted was changed. Note that, in the latter, as the number of selected frames is fixed, altering the frame rate affects primarily the clip's time length but also the number of extracted clips (since only complete clips are considered for extraction). For all these variants, the results are very similar to the ones achieved by the proposed routine, which demonstrates the model's robustness to the setting of these parameters. Overall, the use of the complete video over multiple clips proved to be the key component of the proposed method. This is due to the fact that averaging multiple predictions can help reduce the model's uncertainty and ultimately improve its final prediction. The bigger difference within these variants was observed when a smaller frame rate is

Table 3.3 - Comparison of inference routines and influence of pre-processing settings

	MCC	**F1**								
		SX	PSLA	A4C	IVC	PSSA-MV	PSSA-AV	PSSA-PM	Macro	Micro
Proposed Routine	0.9276	0.9715	0.9654	0.9636	0.9898	0.7795	0.8464	0.8607	0.9110	0.9400
One clip only	0.9220	0.9734	0.9630	0.9598	0.9891	0.7650	0.8357	0.8434	0.9042	0.9353
Whole video at a time	0.9269	0.9705	0.9646	0.9642	0.9891	0.7835	0.8424	0.8575	0.9103	0.9394
Bigger Step (clip every second)	0.9279	0.9735	0.9646	0.9630	0.9905	0.7771	0.8497	0.8615	0.9114	0.9402
Smaller Step (clip at every 0.1 seconds)	0.9276	0.9745	0.9661	0.9636	0.9898	0.7771	0.8439	0.8579	0.9104	0.9400
Smaller Frame Rate (6 frames per second)	0.9258	0.9697	0.9653	0.9636	0.9927	0.7717	0.8430	0.8503	0.9081	0.9385
Bigger Frame Rate (16 frames per second)	0.9276	0.9706	0.9654	0.9629	0.9898	0.7835	0.8547	0.8563	0.9119	0.9400

selected. Here, the depth stayed as 32 frames and the step as half a second. Because the videos of the dataset are no longer than 6 seconds, this results in a maximum of two clips being extracted (opposed to the usual six), which takes less advantage of the multiple clip inference technique, hence the worst results.

Quality Influence

Leveraging of the fact that part of the videos available in the dataset were given a 5-grade overall quality score, a study was conducted to assess how the quality of the video influences the proposed model's performance. The confusion matrix resultant of the prediction of all quality-annotated samples is present in Figure 3.5A. The confusion matrices for the videos with extreme grades, *i.e.* 5 for excellent quality and 1 or 2 for very poor quality (these two grades were aggregated as there were not enough videos of grade 1 for a proper analysis), are present in Figure 3.5B and 3.5C, respectively. The corresponding metrics, as well as the ones for grades 3 and 4, are summarised in Table 3.4. As expected, as the quality of the video declines, the model shows more difficulty in classifying the correct view. This discrepancy in results further proves the importance of quality assessment of FoCUS videos, as the success of the model is clearly influenced by the video quality, even in an initial processing task such as identifying the acquired view.

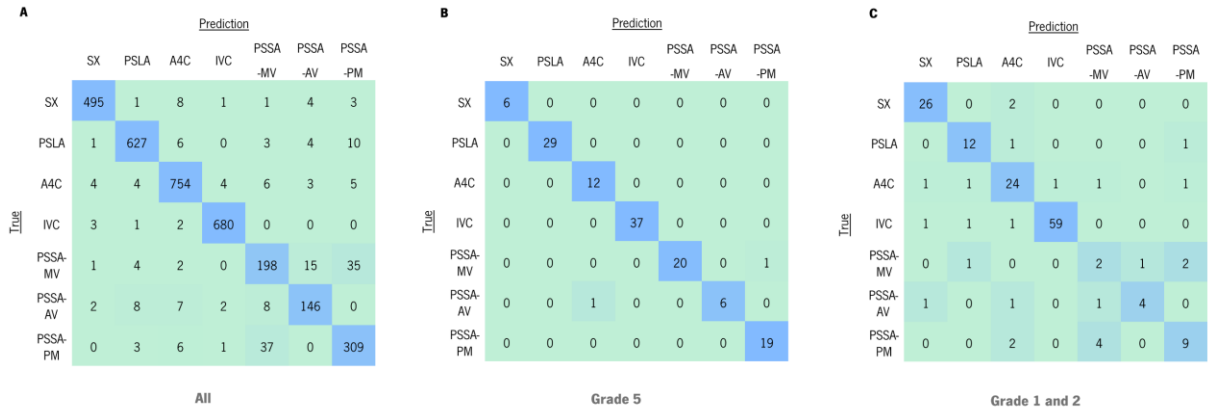


Figure 3.5 - Confusion matrices on the quality annotated set: A) all videos, B) videos with grade 5 and C) videos with grade 1 or 2.

Table 3.4 - Results on the quality annotated set per grade assigned

	MCC	F1								
		SX	PSLA	A4C	IVC	PSSA-MV	PSSA-AV	PSSA-PM	Macro	Micro
All	0.9299	0.9715	0.9654	0.9636	0.9898	0.7795	0.8464	0.8607	0.9110	0.9400
Grade 5	**0.9822**	**1.0000**	**1.0000**	0.9600	**1.0000**	**0.9756**	**0.9231**	**0.9744**	**0.9761**	**0.9847**
Grade 4	0.9557	0.9920	0.9911	**0.9863**	0.9956	0.8148	0.8148	0.9365	0.9330	0.9620
Grade 3	0.9289	0.9565	0.9571	0.9643	0.9877	0.7317	0.8571	0.8889	0.9062	0.9390
Grade 1 and 2	0.8188	0.9123	0.8276	0.8000	0.9672	0.2857	0.6667	0.6429	0.7289	0.8447

3.3 Video-Based Classification using Spatio-temporal Features

As previously stated, FoCUS videos are acquired in conditions that provoke a lot of instability. As such, not every frame contains useful information, which might be ignored when one considers the global average pooling layer used in the previously proposed model. Hence, this section describes a comparative study between different architectures to leverage of the temporal information present in the multiple frames that constitute the video.

3.3.1 Methods

3.3.1.1 Overview

For this study, different network architectures were developed and tested. All architectures stem from the previously presented framework, incorporating the same spatial feature extractor (red block in Figure 3.3), while maintaining the same data preprocessing, model training and inference routines described in Section 3.2.1. The analysed methods of aggregating the feature maps of each frame can be organised into three main groups, as represented in Figure 3.6. The weighting group (Section 3.3.1.2) explores strategies to give more importance to some frames than others, performing a weighted average

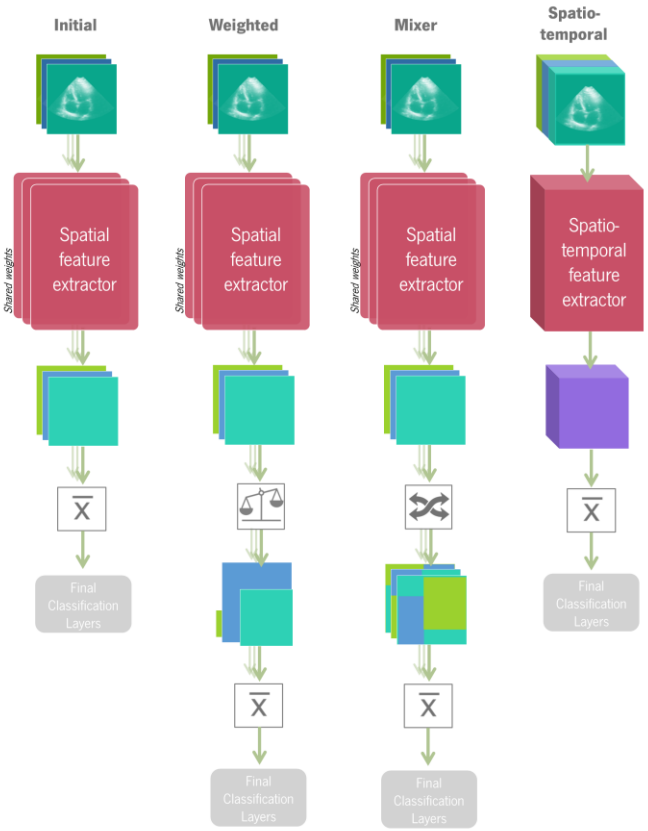


Figure 3.6 - Overview of the analysed methods to compute spatio-temporal features.

of the feature maps instead of the regular average considered in the initial proposal. Mixers (Section 3.3.1.3) encompass schemes than blend the feature maps of the frames, combining them into new ones before averaging them for the final prediction. The final group, the spatio-temporal (Section 3.3.1.4), takes the initial scheme and adapts it to compute, from the first layer, features that leverage both spatial and temporal information.

3.3.1.2 Weighting Networks

The implemented weighting networks are represented in Figure 3.7. The first step in all of them is spatially averaging the feature maps of each frame, and then each one has a different strategy to compute the weights. These weights are passed through a softmax activation to scale them to a range of [0,1], and are then multiplied to each corresponding set of features and summed, performing in this way the weighted average. The number of parameters of each one of these networks is indicated in Table 3.5.

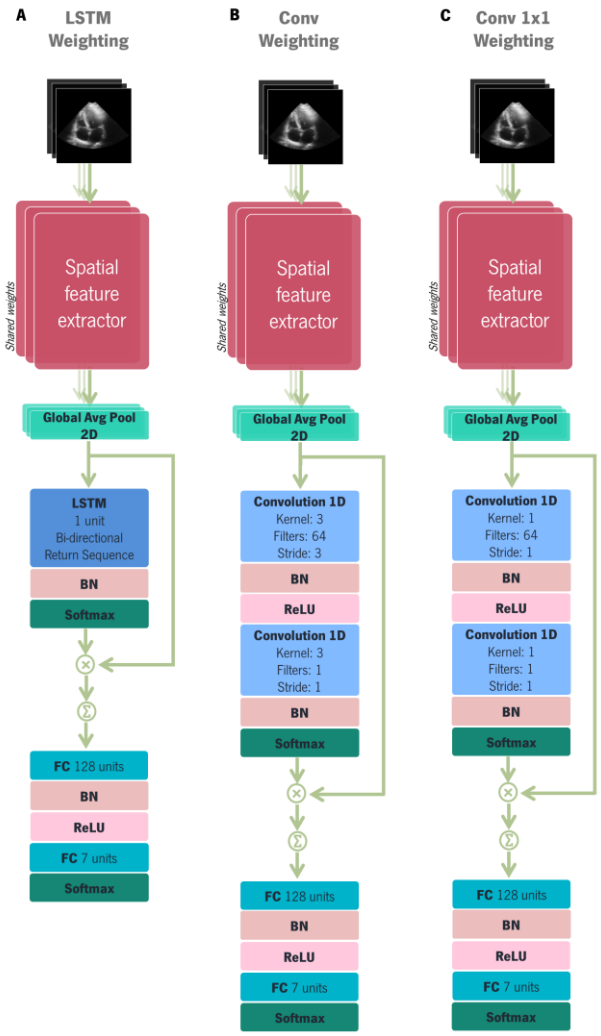


Figure 3.7 - Illustration of the weighting networks implemented.

BN – Batch Normalisation; FC – Fully Connected.

Table 3.5 - Number of parameters in each implemented weighting network

Network	Number of Parameters
LSTM Weighting	11,250,896
Conv Weighting	11,345,540
Conv 1x1 Weighting	11,279,876

The presented architectures are inspired by the work of Wang *et al.* [72], namely by their RNN and temporal convolutional methods for feature aggregation. In the former, each weight is computed using a LSTM layer, which considers the sequential relation of each frame to the other ones in the clip (Figure 3.7A). The second method relies in applying convolutions over the temporal dimension, where the kernel size controls the receptive temporal field, *i.e.* the number of frames considered when computing each weight. Following their proposal, we implemented a kernel size of 3, analysing each frame and its adjacent (Figure 3.7B). To compute the weights disregarding the correlation with other frames, a kernel size of 1 was also considered (Figure 3.7C).

3.3.1.3 Mixer Networks

In this group of networks, each feature map is spatially averaged and then they are combined based on their temporal information, following different methods. First, structures similar to the ones presented in the previous section were applied, namely a LSTM (Figure 3.8D), a convolution-based architecture with kernel size of 3 (Figure 3.8B) and another with size 1 (Figure 3.8A).

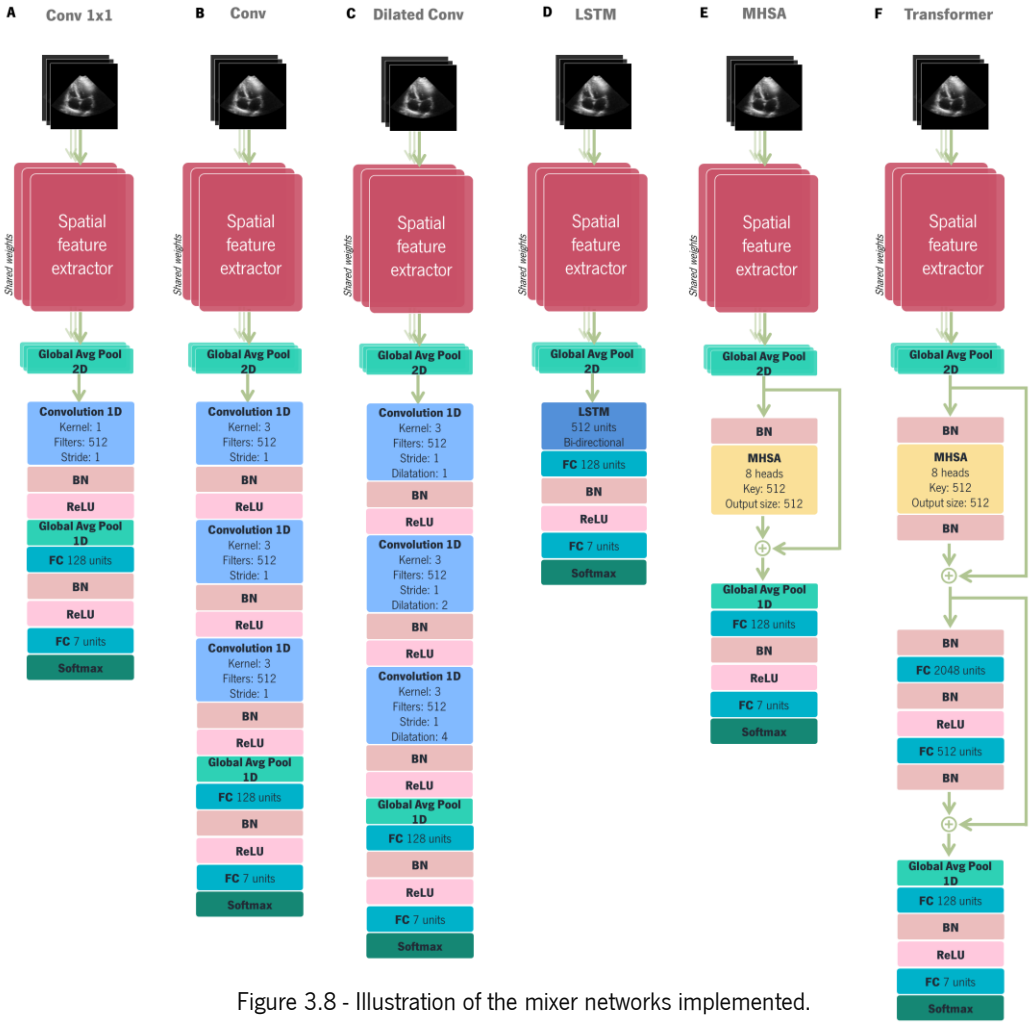


Figure 3.8 - Illustration of the mixer networks implemented.
BN – Batch Normalisation; FC – Fully Connected; MHSA – Multi Head Self Attention.

Inspired by the approach of Temporal Convolutional Networks [73], a convolution scheme was implemented with dilations (Figure 3.8C). Dilations consist in the expansion of the kernel by inserting gaps between its consecutive elements, skipping pixels. This technique allows to increase the layer's receptive field, *i.e.* the number of neighbouring frames considered while mixing the features.

Recently, attention layers such as Multi-Head Self-Attention (MHSA) grew in popularity [74]. MHSA is a module that runs through an attention mechanism several times in parallel. In light of that, two methods were implemented: one simply based on this layer (Figure 3.8E) and the other based on the encoder of the Vision Transformer [75], an attention-based network that handles long-range dependencies (Figure 3.8F).

The total number of parameters of each mixer network is summarised in Table 3.6.

3.3.1.4 Spatio-temporal Network

The final hypothesis is to extract spatio-temporal features from the clips directly. To this end, the clip is processed as a 3D block. To do so, the architecture proposed in Section 3.2 is adapted by transforming its spatial convolutions into 3D convolutions (Figure 3.9). In this type of convolution, the filter is not only applied through the two dimensions of each frame, but also across the time dimension (*i.e.* the frames). Note that all other hyperparameters (like the number of filters per convolutional layer or the total number of convolutional layers) are kept. This model has a total of 33,224,000 parameters.

Table 3.6 - Number of parameters in each implemented mixer network

Network	Number of Parameters
Conv 1x1	11,510,976
Conv	13,612,224
Dilated Conv	13,612,224
LSTM	15,445,184
MHSA	12,299,456
Transformer	14,410,944

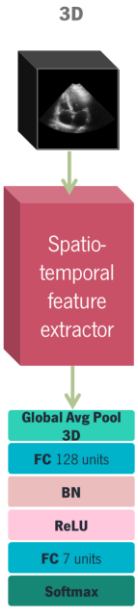


Figure 3.9 - Illustration of the spatio-temporal network implemented. BN – Batch Normalisation; FC – Fully Connected.

3.3.2 Results and Discussion

The results of the networks studied on the validation set are summarised in Table 3.7. The weighting group performed well in general, showing a performance close to the initial proposal and the Conv Weighting network even slightly outperforming it. Given the lower performance observed for the Conv 1x1 variant, one may conclude that it is not a good strategy to determine the importance of each frame only by considering its own feature maps. On the contrary, in the LSTM weighting method, where all frames contribute to the calculation of the weights, no good results were obtained either. This could be because the LSTM has a complicated internal system that hampers its training, failing to properly converge under the training conditions used. Thus, calculating the relative importance of a frame by calculating its weight considering the neighbouring frames was the best strategy of the group. The Mixer group's results show that these techniques for transforming the spatial feature maps are not helpful for the task at hand. When analysing the metrics, one notices that the larger the receptive field defining the recombined feature maps, the worse the performance (apart from the LSTM). Since the temporal features are computed from the spatial feature maps, the applied transformations are very limited, and the models cannot produce good representations. Moreover, all these networks are more complex (with increased number of trainable weights and more nonlinearities), so they might benefit from more regularisation. In particular, the MHSA and Transformer variants may require tuning of their specific hyperparameters. The spatio-temporal network, despite having roughly double the parameters of the others, showed a considerable improvement with respect to the initial model, proving the added value of extracting features with both spatial and temporal information from the outset.

Table 3.7 - Performance of the implemented methods on the validation set

	MCC	**F1**								
		SX	PSLA	A4C	IVC	PSSA-MV	PSSA-AV	PSSA-PM	Macro	Micro
Initial	0.9276	0.9715	0.9654	0.9636	0.9898	0.7795	0.8464	0.8607	0.9110	0.9400
LSTM Weighting	0.9220	0.9628	0.9651	0.9618	0.9905	0.7534	0.8696	0.8390	0.9060	0.9353
Conv Weighting	0.9283	0.9659	0.9624	0.9641	0.9949	0.7782	0.8703	0.8571	0.9133	0.9405
Conv 1x1 Weighting	0.9230	0.9667	0.9617	0.9635	0.9891	0.7583	0.8630	0.8455	0.9068	0.9361
Conv 1x1	0.9230	0.9590	0.9663	0.9628	0.9884	0.7723	0.8488	0.8476	0.9064	0.9361
Conv	0.9213	0.9696	0.9591	0.9737	0.9861	0.7390	0.8324	0.8491	0.9013	0.9347
Dilated Conv	0.9139	0.9531	0.9563	0.9623	0.9854	0.7438	0.8481	0.8319	0.8973	0.9285
LSTM	0.9216	0.9609	0.9684	0.9687	0.9913	0.7302	0.8418	0.8451	0.9009	0.9350
MHSA	0.9128	0.9618	0.9640	0.9596	0.9834	0.7064	0.8521	0.8235	0.8930	0.9277
Transformer	0.9089	0.9570	0.9647	0.9661	0.9883	0.6858	0.8299	0.8079	0.8857	0.9244
3D	0.9315	0.9617	0.9700	0.9623	0.9934	0.8008	0.8889	0.8584	0.9194	0.9432

In Table 3.8, the performance of the spatio-temporal network on the two test sets are presented. The ROC curves and the corresponding AUC of the ensembled test set 1 can be found in Figure 3.10. When compared to the results of the multi-frame approach (Table 3.1 and Figure 3.4), the improvement in the first test set is striking. However, this is not verified for the second test set, which seems to suggest one has reached a plateau on the performance one can obtain with this type of network in this dataset.

From the 2D approach to the multi-frame one and finally the 3D network, results kept improving as more temporal information was considered, proving its value for the present task. Ultimately, a MCC of 0.9569 was achieved in the test set, an excellent result compared to the state-of-the-art results on view identification and considering the challenging particularities of FoCUS videos (compared to a conventional echocardiogram). Further improvements would probably only come from alterations to the dataset, such as reanalysing the videos of the dataset and correcting possible mislabelled videos, or gathering more data for the underrepresented views (*e.g.*, all three PSSA). One could argue that eliminating extremely bad quality videos could also help, however this dataset is representative of day-to-day FoCUS acquisitions and as such these should be included to guarantee an adequate generalisation to a real-world scenario.

Table 3.8 - Results of the spatio-temporal network on the test sets

		MCC	**F1**								
			SX	PSLA	A4C	IVC	PSSA-MV	PSSA-AV	PSSA-PM	Macro	Micro
Test 1	Single	0.9377	0.9730	0.9684	0.9824	0.9918	0.8155	0.8967	0.8555	0.9262	0.9483
	Ensemble	0.9569	0.9814	0.9870	0.9932	0.9917	0.8515	0.9667	0.8794	0.9501	0.9642
Test 2	Single	0.9250	0.9523	0.9705	0.9702	0.9926	0.7871	0.9030	0.8104	0.9123	0.9377
	Ensemble	0.9352	0.9600	0.9773	0.9765	0.9947	0.8079	0.9189	0.8290	0.9235	0.9461

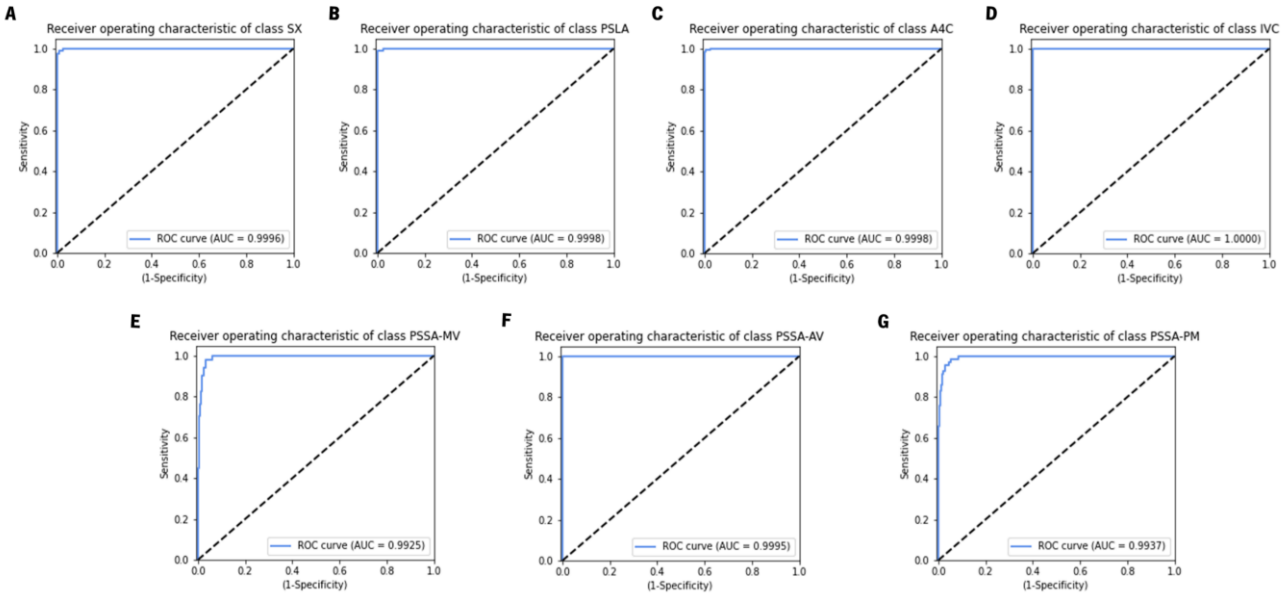


Figure 3.10 - ROC and respective AUC of each class for test set 1 with the spatio-temporal network.

4. Quality Assessment of FoCUS Videos

Since the clinician's ability to interpret an echo study highly depends on image quality, suboptimal images can adversely alter patient care. Automatic quality assessment of FoCUS videos is therefore highly desirable to ensure that captured videos are suitable for health assessment, or even to exclude uninterpretable ones from further automatic analyses. In this chapter, view-specific quality assessment has been investigated, taking advantage of the model developed in the previous chapter. First, the dataset used to train the models is introduced, as well as the quality attributes annotated for each view. Then, the methodologies developed for this study are presented. Finally, the focus is on the results achieved and their discussion.

4.1 Dataset

The FoCUS dataset used in this study has the same origin as the one used in the previous chapter. As such, it falls under the same description and preparation routine as depicted in Sections 3.1.1 and 3.1.2, respectively. Since the goal is to create view-specific models, the dataset was then separated per view. However, given the effort involved in annotating a relevant quantity of videos per view, the work here presented focuses on the SX, A4C and IVC views only. In Section 4.1.1, the annotation of the quality attributes of each video of these views is described. Then, Section 4.1.2 presents the division of each view dataset into sets.

4.1.1 Annotation

Two experts were responsible for establishing the relevant attributes of quality to be annotated: a cardiologist who specialises in cardiac imaging, and an internal medicine doctor proficient in point-of-care ultrasound. The difference in background is representative of the reality of the many users of FoCUS. As quality assessment is a subjective process, several meetings took place between the development team and the two experts to reach a consensus on which features are relevant in each cardiac view. The different backgrounds of the experts allowed for a good definition of what to expect from a good FoCUS acquisition, as one is used to high quality and good image interpretability, and the other understands the

urgency of the exam. The two established the list of possible feedback that should be given on each video, which led to the guidelines that were then followed during the annotation of the videos.

The three views have four common quality attributes assessed: (1) the image gain, which can be classified as 'Insufficient', 'Appropriate', or 'Excessive'; (2) the image depth, which follows the same possibilities; (3) the overall quality of the video, selecting a score of '1', '2' or '3'; and (4) the video orientation, where one assesses if the video was acquired in an inverted orientation with respect to the acquisition guidelines (in which case it is labelled as 'Inverted'). The other annotations are view-specific as they target the assessment of particular anatomical structures or features of a given view. Most of these attributes constitute a binary label and, in this case, the positive class is assigned to the videos where the anatomical reference or feature assessed is not present.

In the SX view, there are four specific attributes that were considered. Following the guidelines for this view, the heart apex should be visible. When it is not, the video is labelled as 'Apex not present'. On the other hand, the AV outflow tract should not be visible, so if it is present in the video, the label of 'AV tract' is assigned. The goal in this view is to clearly visualise all four chambers of the heart, as well the heart crux (interventricular and interatrial septum, MV and tricuspid valve). However, inappropriate intersection of the imaging plane with the cardiac anatomy can result in the sub-optimal visualisation of these structures or they may even be missing, which leads to the video being labelled as 'Misaligned chambers'. Furthermore, ideally, the interatrial septum should be very close to perpendicular to the ultrasound beam, and the interventricular septum should make an angle wider than 45° with the vertical axis. When this is not witnessed in the video, it is marked as 'Incorrect Anatomical Orientation'.

The A4C view also displays, ideally, the four chambers and the heart crux. However, here the goal is to picture the septum vertically. Hence, when this aspect is not present, the video is labelled as 'Incorrect Anatomical Orientation'. Each video is also assigned the label 'Misaligned chambers' when the incorrect positioning of the imaging plane prevents the correct and simultaneous visualisation of the four chambers (without foreshortening and without imaging the aortic tract). In this view particularly, the right ventricle is frequently displayed with poor to no definition of its outer wall, which was labelled as 'Sub-optimal RV'. Misplacement of the probe can also originate what was labelled as 'Incorrect windowing position', which visually translates into the top center of the field of view (FOV) not being aligned with the LV apex (often positioned towards the right side of the heart). Additionally, the label 'Near-field artifact' was attributed to videos that displayed an ultrasound artifact characterised as clutter beneath the near field. Although the cause of this artifact is attributed to the transducer itself and not the technique of

acquisition, as it can be an obstacle to a correct interpretation of the cardiac function, its presence was deemed important to assess.

As described in Section 1.2.2, in an ideal acquisition of the IVC view, the suprahepatic vein and the RA confluence are visualised. As such, the absence of these structures was assessed and labelled as 'Suprahepatic not present' and 'Confluence not present', accordingly. A common mistake is to fail the correct centering of the sector scan's FOV in the IVC when attempting to image all other relevant anatomical references. This attribute was assessed in a multi-class manner, with each video classified as either 'Confluence centered', 'Correctly centered' or 'Liver centered'.

It should be noted that, despite the established criteria, quality assessment is a subjective process laced with decision variability, even when an image is reassessed by the same observer. Because of this, a certain level of label noise is expected to be present in this dataset.

4.1.2 Division into Sets

The SX dataset was composed of 819 videos from 592 exams. For the A4C view, 1,168 videos from 663 different exams were gathered. Finally, the IVC dataset summed up to 1,073 videos of 847 exams. These datasets are of relatively small size and revealed severe imbalance in some of the tasks. Because of this, the cross-validation method became unfeasible as it could not be ensured that every class was adequately represented in all sets. Instead, while maintaining sample independence, each dataset was randomly divided into 6 groups of equal proportion and four fixed groups were used for training, and the remaining two are used for validation and test, respectively. The relative distribution of the labels of each task in the three sets of the SX, A4C and IVC datasets is shown in Figures 4.1, 4.2 and 4.3, respectively. The 'Image Orientation' task is not represented in these figures because it was very

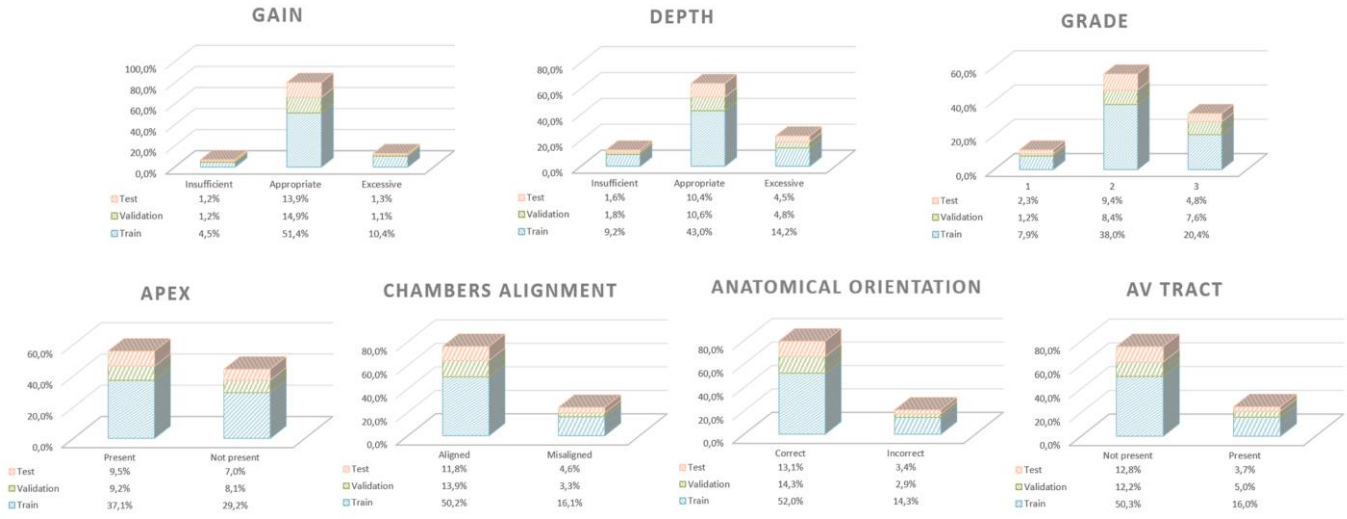


Figure 4.1 - Relative distribution of videos per class in training, validation, and test sets for each task of the SX dataset.

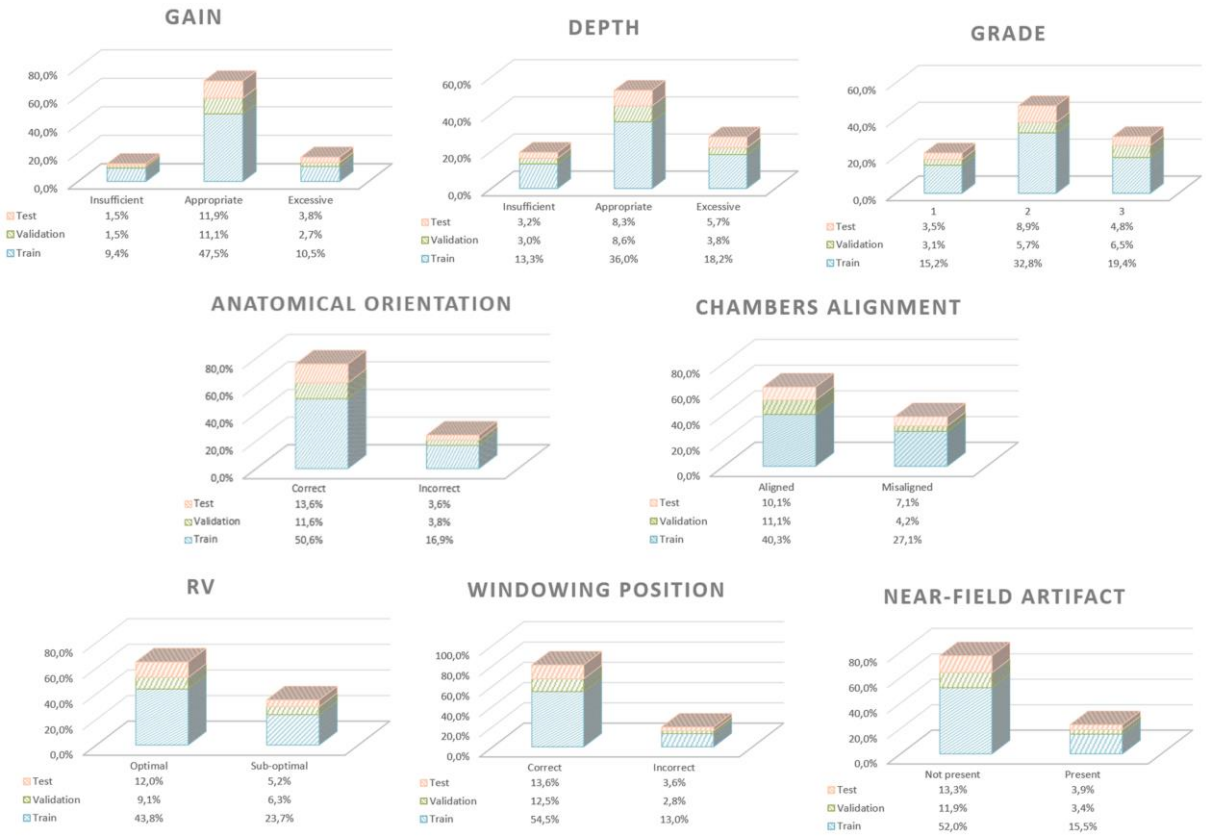


Figure 4.2 - Relative distribution of videos per class in training, validation, and test sets for each task of the A4C dataset.

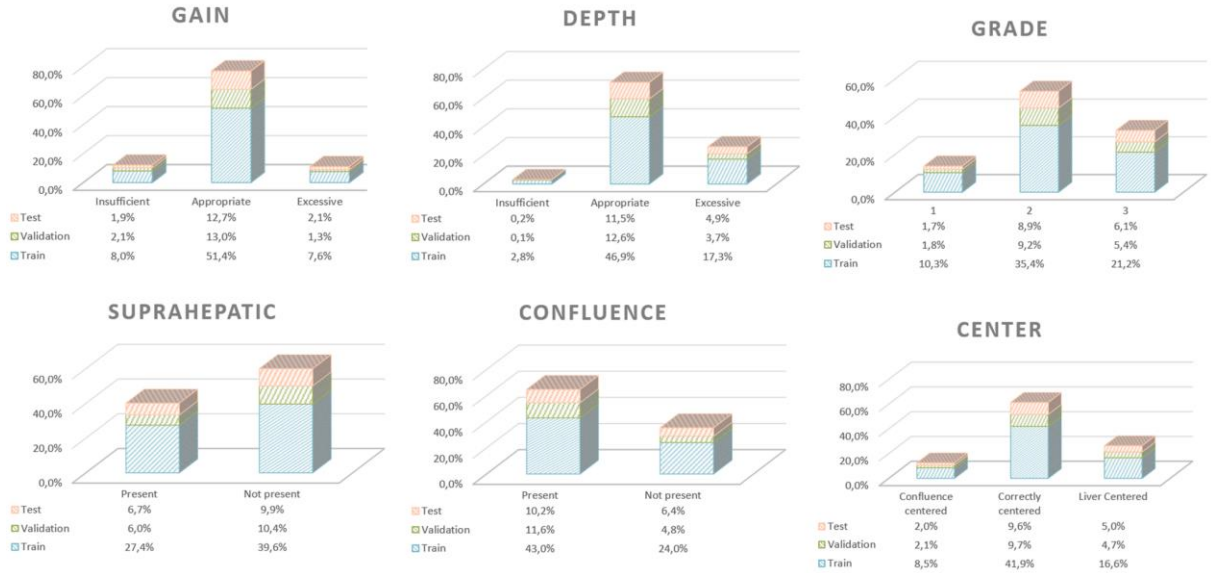


Figure 4.3 - Relative distribution of videos per class in training, validation, and test sets for each task of the IVC dataset.

underrepresented in the three views. Hence, the few examples were corrected with a horizontal flip, and then during training this label was simulated on the fly.

4.2 METHODS

This task uses the model developed in the previous task as starting point, as it shown good ability to process these images. However, since the 3D model requires a large amount of data, which is not

available for this task, the model used as backbone is the multi-frame CNN presented in Section 3.2. This limited amount of data also makes the model prone to overfitting. To prevent it, dropout layers were added before each FC layer with a dropout rate of 15%.

To classify every identified attribute per view, three strategies were investigated. On the one hand, one considered the training of an independent model per attribute, as depicted in Figure 4.4A. On the other hand, a multi-task strategy was studied, where view-specific networks are implemented predicting the multiple outputs simultaneously. Whitin this multi-task proposal, two variants were explored: one where the recombination of the spatial features is specific for each task (Multi-task 1, Figure 4.4B); and one where common features are computed for all tasks, differing only on the last fully connected layer (Multi-task 2, Figure 4.4C). The number of parameters of each model is shown in Table 4.1 (together with the number of independent models considered in each strategy and view).

4.2.1 Implementation Details

The routines of data preprocessing, model training and inference described in Section 3.2.1 are maintained, except for the initial learning rate that is now set to 3×10^{-4} and the augmentations' ranges which were reduced to not affect the attributed labels. The lower initial learning rate was set to avoid training instability, which was observed for certain tasks and, particularly, for the multi-task scenario.

As previously mentioned, due to extremely scarce videos with the 'Inverted' label, the examples for training and evaluation of this task were simulated on the fly. In the case of the independent model of

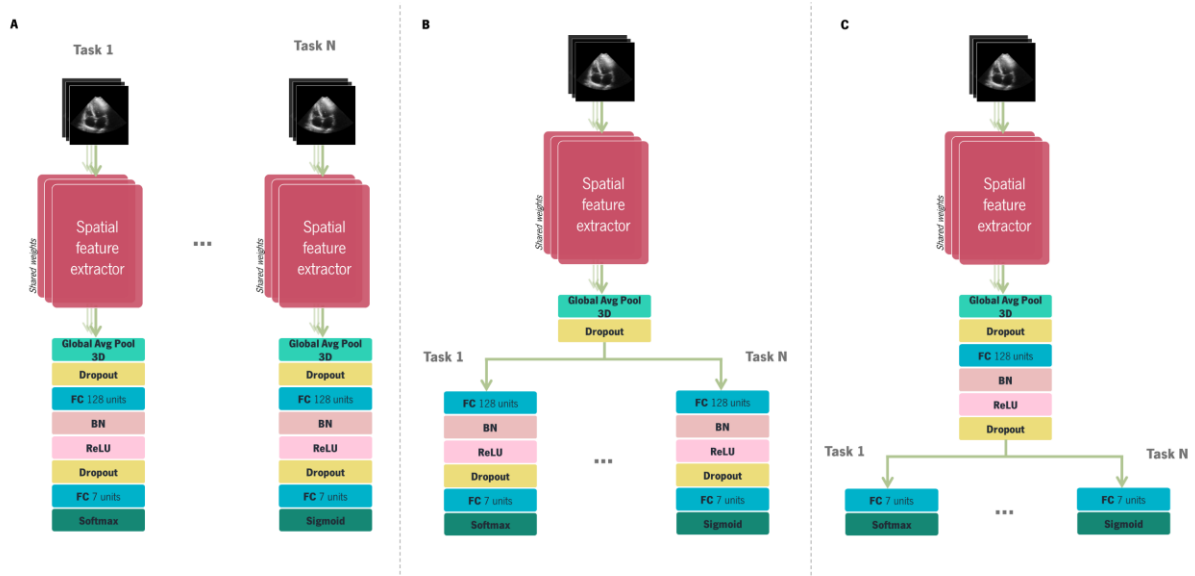


Figure 4.4 - Illustration of the view-specific networks implemented: A) Independent, B) Multi-task 1 and C) Multi-task 2. BN – Batch Normalisation; FC – Fully Connected.

Table 4.1 - Number of parameters of each view-specific model

	SX	A4C	IVC
Independent	11,246,784 × 8	11,246,784 × 9	11,246,784 × 7
Multi-task 1	11,710,016 × 1	11,764,288 × 1	11,644,096 × 1
Multi-task 2	11,247,680 × 1	11,247,808 × 1	11,247,808 × 1

this task, the images were transformed with a horizontal flip (and labelled accordingly) with a probability of 50%, guaranteeing a balanced representation of the class. During inference, the images were simulated at the same rate. However, in the multi-task scenario, adjustments were made. The training was performed while simulating 10% of images inverted. This decision was made because the presence of inverted images, if in high prevalence, could cause noise and hamper the training of the other tasks. With a lower but sufficient rate, the network is able to deal with the class imbalance and learn relevant information for the 'Image Orientation' task, without compromising the learning process for the other tasks. Logically, the inference was also adapted, being performed in two stages: first, one assesses the performance of the other tasks by running the inference on the original images of the validation/test set; and second, one assessed the 'Image Orientation' task by considering 50% of the validation/test images inverted (to allow a direct comparison with the results observed for the 'Independent' model).

Given the limitations on data quantity, one also investigated the use of a pre-training strategy like transfer learning. As such, the three networks presented in Figure 4.4 were trained with two different schemes for weights initialisation and training. On the one hand, all weights were initialised with the normal distribution proposed in [69] and trained for 75 epochs. On the other hand, the weights of the spatial feature extractor obtained in Section 3.2 were instead used to initialise the networks. In the latter, after initialising the network, the transferred weights were kept frozen (*i.e.* were not updated) for 25 epochs (optimising solely the last layers of the network), and updated normally throughout the remaining 50 epochs (keeping the total number of training epochs equal to the random initialisation scenario).

4.3 Results and Discussion

Since cross-validation was not possible to perform, and only one validation group was set, each model was trained three times to account for the random initialisation and the training paradigm, and to emphasise the robustness of the results. The network's hyper-parameters such as dropout rate, L2 regularisation weight, learning rate, optimiser and batch normalisation momentum were studied and optimised to the values previously presented on the validation set (data not shown). The value of using the PolyLoss and the Blur block was also confirmed on the validation set.

The results for each task trained independently on both validation and test sets can be found in Appendix A. These results are summarised in Table 4.2, represented by the average results for each view. Note that the value of F1 represents the positive class result for binary tasks and the macro F1 for multi-class tasks. Also, the results are shown, once again, by averaging the metrics obtained by the three repetitions or by the ensembling technique. Once again, model ensembling has proven to be beneficial, with the metrics of all tasks surpassing those obtained by a single model. Indeed, the latter is associated with higher uncertainty (which is perceivable by the variability seen among the three repetitions), particularly in these conditions of limited data, which is mitigated by the ensemble. It is also noticeable that the results on the validation and test sets are very similar for most of the tasks, which shows a good generalisation of the models. In some of the SX tasks, there is a slight decrease in the metric scores. However, this variability can be attributed to the different distribution of classes on the validation and test sets and even the low representation of these classes.

The following analyses consider results obtained for the test set with the ensemble technique. Of note, this decision was made to keep the assessment unbiased, since at no point was the test data used or analysed for the design of the networks or the tuning of the models' training conditions.

The confusion matrices of each task of the SX, A4C and IVC views are shown in Figures 4.5, 4.6 and 4.7 respectively. Across all 24 tasks, an average F1 value of 0.7243 was achieved, indicating good performance on these tasks. However, it should not be disregarded that this value is inflated by the results of the relatively easy 'Image Orientation' task.

When analysing these results, it is noticeable that the performance is better in tasks that analyse the image as a whole, *i.e.* the attribute being assessed alters the image globally (or nearly) by modifying how the anatomy is perceived. Examples include tasks like 'Gain', 'Depth', 'Image Orientation', 'Anatomical Orientation', 'RV', 'Window Position' or 'Center'. There are although some exceptions, in which tasks targeting global-like features do not show the same success. Across all views, the 'Grade' task falls under this exception. Despite representing a global characteristic of the image, it is a task with a subjective nature, as several factors must be considered in the annotation process, with their relative importance

Table 4.2 - Average per-view performance of the independent networks

		SX		**A4C**		**IVC**	
		MCC	F1	MCC	F1	MCC	F1
Validation	Single	0.5703	0.6999	0.5981	0.7222	0.5653	0.7268
	Ensemble	0.5838	0.7053	0.6053	0.7259	0.5723	0.7168
Test	Single	0.5225	0.6702	0.6098	0.7373	0.6038	0.7231
	Ensemble	0.5528	0.6887	0.6273	0.7489	0.6270	0.7354

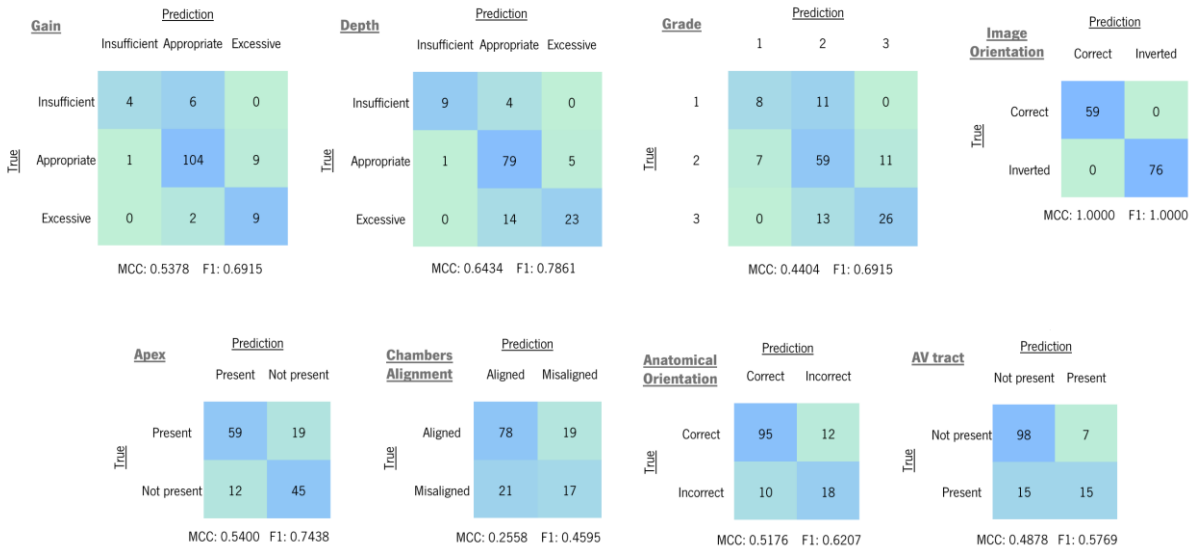


Figure 4.5 - Confusion matrices of the independent SX models ensembled on test set.

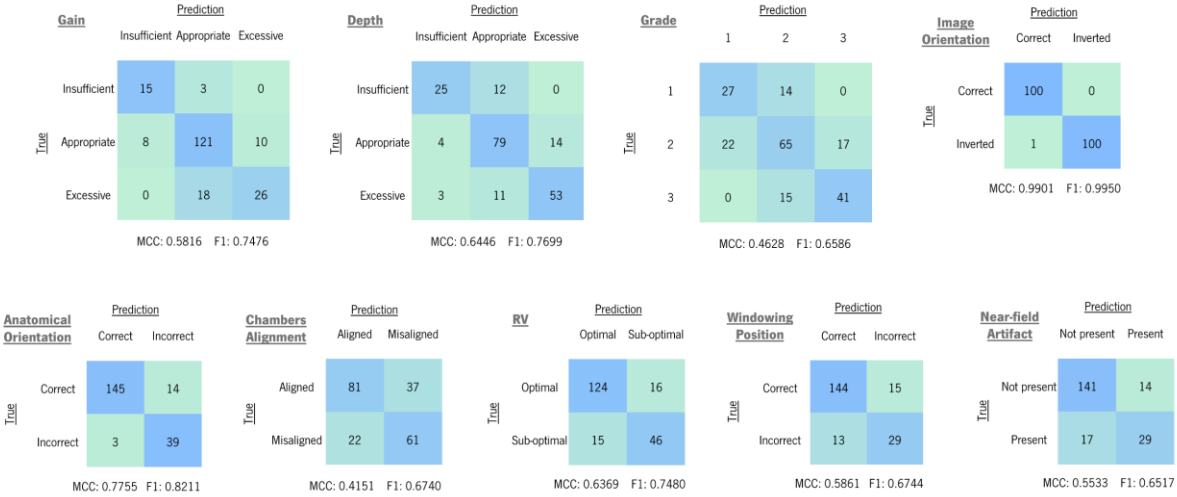


Figure 4.6 - Confusion matrices of the independent A4C models ensembled on test set.

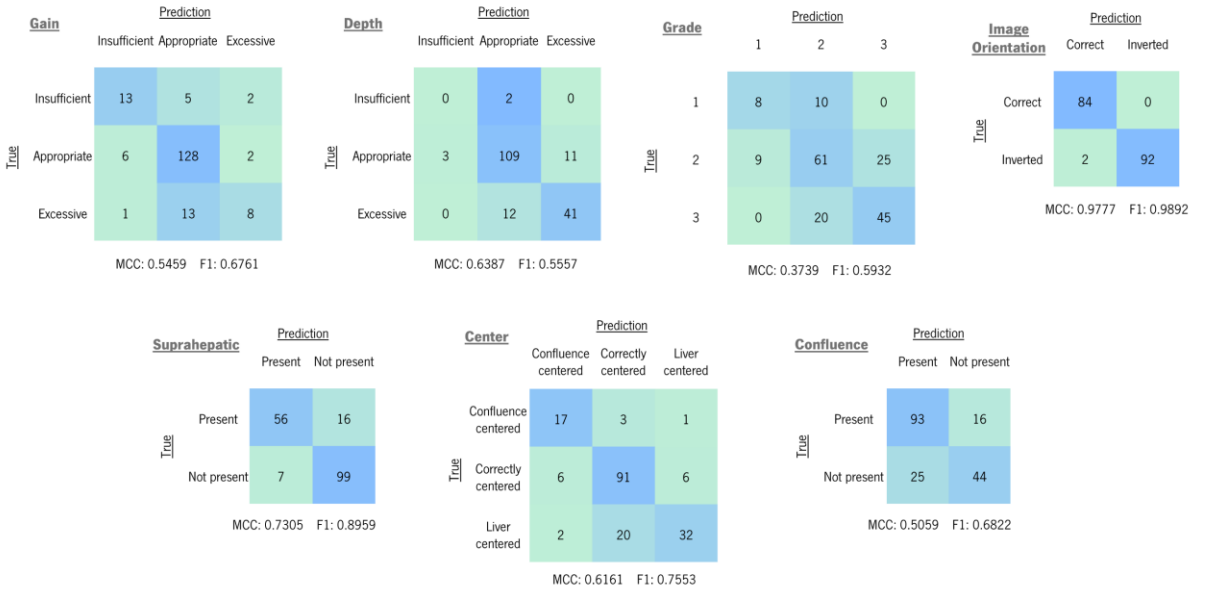


Figure 4.7 - Confusion matrices of the independent IVC models ensembled on test set.

established by the observer when assessing the video itself and some of which (like the image contrast) being dependent on the observer's personal preference. Indeed, not only do the many factors cause some intraclass variability, but more noise is expected in these labels, which translates into poorer performance. The 'Anatomical Orientation' task for the SX view also shows poorer performance, once again related with the subjectivity in labelling. In this view, the main reference point is the inclination of the interventricular septum, and it is subjective at which limit one considers it too vertical. Although a similar comment could be made about the A4C view, the latter is far more standardised in clinical practice and is thus easier to label (as one intuitively perceives when it is not acquired correctly). In the IVC view, the 'Depth' task also presents a sub-optimal result, which was expected given the underrepresentation of the 'Insufficient' class in the dataset (and particularly in the test set). The task 'Chamber Alignment' also does not achieve the same results as the other tasks in both views. In this case, the low metric values are a product of the high intraclass variability, as the misalignment can be caused by different acquisition errors (foreshortening, five chambers present, wrong intersection, etc.) that can have very distinct appearances. This variability makes the task more challenging, hence the worst performance. It is worth noting that the 'Image Orientation' task for the IVC view performs well, but slightly below the results obtained in the other views. This is because this task is more complex in this view, as the absence of some anatomical references like the suprahepatic vein and the right atrium confluence makes the image symmetrical (showing only the cava crossing the liver horizontally).

Following the same logic, the attributes representing the presence or absence of localised structures, such as 'Apex', 'AV tract', 'Near-field Artifact' and 'Confluence', show lower performance metrics. An exception to this behaviour is the 'Suprahepatic' task. This task performed very well despite being a small structure, which is explained by the contrast it creates when present. The IVC is a view that is typically hyperechoic, and the suprahepatic vessel stands out as a hypoechoic region, easing its assessment.

The results of the other implemented approaches, namely the two methods of multi-tasking and transfer learning, are present in full in Appendix B and summarised in Table 4.3. Although the multi-task strategy seemed helpful for the IVC tasks, it caused a significant decrease in performance for the other views. This result is explained by the larger dependency between tasks in the IVC view. For example, as the suprahepatic vein and the right atrium confluence are localised structures, when present they are a key indicator of whether the image is inversed or not, and the same principle applies to the 'Center' task. This does not apply to the other views, as their tasks are mainly independent (except for the A4C tasks of 'Anatomical Orientation' and 'Windowing Position'), which explains the poorer performance of a multi-

Table 4.3 - Average per-view performance of the implemented networks

		SX		A4C		IVC	
		MCC	F1	MCC	F1	MCC	F1
Random Initialisation	Independent	0.5528	0.6887	0.6273	0.7489	0.6270	0.7354
	Multi-task 1	0.5096	0.6666	0.5920	0.7248	0.6415	0.7458
	Multi-task 2	0.5061	0.6651	0.5930	0.7239	0.6444	0.7482
Transfer Learning	Independent	0.5031	0.6630	0.5970	0.7247	0.5996	0.7272
	Multi-task 1	0.5028	0.6609	0.5814	0.7164	0.6286	0.7398
	Multi-task 2	0.5224	0.6711	0.5968	0.7229	0.6400	0.7495

task strategy. It is also noticeable that between the two implemented architectures of multi-tasking, there is little difference in the results. Yet, a closer look to the results per task shows that the 'Grade' task always shows a better, or at least equal, performance for the multi-task schemes when compared to the independent models. As previously stated, this attribute is assessed based on multiple factors, and most of these factors are contemplated by the other tasks. As such, the training of a single model targeting the classification of the multitude of tasks clearly benefits the prediction of the overall video quality.

The transfer learning technique, in general, did not add any value, with the only exception being the Multi-task 2 network for the SX and A4C views. This is likely because the features learnt for the view identification (used here as initialization) are not relevant for the view-specific tasks of quality assessment. Additionally, the networks are relatively small and the number of videos probably enough for the spatial feature extractor to learn good representations (given its shared nature across input frames), which decreases the necessity for a better weight initialization.

To conclude, the implemented strategies failed to improve the performance of the quality assessment tasks. Since the multi-task scheme showed improvements in some tasks, perhaps it would be interesting in the future to investigate the creation of subsets of the tasks. Instead of all tasks in one model, tasks related to each other would be trained in one model, sharing features, while the others on independent models. Inferring all tasks in one model is probably too complex, as the same/very similar set of features have to represent very distinct details concerning multiple tasks. In terms of transfer learning, the technique should not be disregarded, however, the task from which the weights were transferred was not ideal. Finding a better, more related pre-training task to quality assessment could be key. Moreover, the results showed worst performance on the tasks that represented the presence of localised structures. To improve this, it should be considered some network architectural changes to allow focusing on specific regions of the video (spatial attention, squeeze and excitation modules [76], etc.). In addition, changing the task from classification to an object detection problem, as previously proposed in

the literature [45], could also lead to a better performance in these cases. Overall, the main obstacle was the simultaneous presence of label noise and an unbalanced distribution of classes in most tasks. A closer look to the distribution of classes in Figures 4.1, 4.2, and 4.3 and the respective results per task in the Appendices shows an inverse correlation between the imbalance ratio and the performance of each task. This shows the large impact the disproportion of classes has in the results. The only tasks where this relation does not hold is when a higher level of label noise is present (*e.g.*, 'Grade'). More specific and rigid annotation guidelines could be made for these attributes to mitigate this issue, although some subjectivity in quality assessment is always foreseeable. Note that, although some loss functions have been proposed to mitigate the effect of noisy labels, these techniques often fail to simultaneously deal with class imbalance (exacerbating its effects) [77]. As such, new strategies that can manage the presence of these two problems are imperative to improve the results on the present dataset.

5. CONCLUSION

In this thesis, the focus was to develop a deep learning pipeline to perform automatic quality assessment of FoCUS videos. To achieve this, a two-stage approach was proposed. First, an algorithm for automatic classification of FoCUS views was developed. This allowed the separation of the videos by views, which are then passed through view-specific models that assess multiple quality attributes.

For the view identification task, a 3D CNN based on the ResNet-18 was created along with a training strategy that leverages domain knowledge into the augmentation scheme, and a multi-clip based inference routine. The experimental results showed that the computation of features considering spatial and temporal information from the start improves the algorithm's performance. In conclusion, the developed framework presents itself as a first-rate solution to the task at hand. Out of seven classes, four of them achieve F1 scores higher than 0.98 on the test set. The other three (all from the parasternal short-axis window) underperformed in comparison, mainly because they have many visual similarities as they are spatially close. Strategies to tackle this limitation and improve their classification should be studied. Another challenge present in the classification of these three views is their underrepresentation in the dataset. Thus, more data of these classes would also be beneficial.

A framework based on the one developed for view identification was implemented to perform quality assessment in three views (SX, A4C and IVC). As annotated data is scarce, in this case, the multi-frame approach was used, as it also had achieved good performance and it did not require as much data as the 3D version. A total of twenty-four attributes were assessed across the three views and showed a good performance with an average F1-score of 0.7243. Notwithstanding, the development of this framework faced two main challenges. First, due to the subjectivity that quality assessment implies, noisy labelling was present in certain tasks. Additionally, along with the relatively small size of the dataset for each view, class imbalance was highly present among the tasks. Hence, it is imperative to find solutions to deal with both problematics in an effective way for these results to be improved. Finally, to apply the developed methods in clinical practice, more annotated data should the gathered and the work should be extended for the remaining four views.

BIBLIOGRAPHY

[1] B. A. Stanton and B. M. Koeppen, *Berne & Levy Physiology*, 7th ed. Elsevier, 2018.

[2] G. J. Tortora and B. Derrickson, *Principles of Anatomy and Physiology*, 12th ed. Wiley, 2009.

[3] S. S. Mader, *Understanding Human Anatomy and Physiology*, 5th ed. McGraw-Hill Publishing, 2004.

[4] K. M. van de Graaff, *Human Anatomy*, 6th ed. McGraw-Hill Publishing, 2002.

[5] R. E. Klabundle, *Cardiovascular Physiology Concepts*, 2nd ed. Wolters Kluwer, 2012.

[6] A. Vander, J. Sherman, and D. Luciano, *Human Physiology: The Mechanisms of Body Function*, 8th ed. McGraw-Hill Publishing, 2001.

[7] 'Inferior Vena Cava'. https://radiologykey.com/inferior-vena-cava-4/ (accessed Dec. 12, 2021).

[8] S. K. Zhou *et al.*, 'A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises', *Proceedings of the IEEE*, vol. 109, no. 5, pp. 820–838, Aug. 2020, doi: 10.1109/JPROC.2021.3054390.

[9] K. L. Moore, A. F. Dalley, and A. M. R. Agur, *Anatomia Orientada para a Clínica*, 7th ed. Guanabara Koogan, 2014.

[10] K. A. Lara Hernandez, T. Rienmüller, D. Baumgartner, and C. Baumgartner, 'Deep learning in spatiotemporal cardiac imaging: A review of methodologies and clinical usability', *Computers in Biology and Medicine*, vol. 130, Mar. 2021, doi: 10.1016/j.compbiomed.2020.104200.

[11] R. Ribes, A. Luna, P. Kuschnir, J. C. Vilanova, and J. M. Jimenez-Hoyuela, *Learning Cardiac Imaging*, 1st ed. Springer, 2010.

[12] J.-L. Zamorano, J. Bax, F. Rademakers, and J. Knuuti, *The ESC Textbook of Cardiovascular Imaging*. Springer, 2010. doi: 10.1007/978-1-84882-421-8.

[13] K. T. Spencer, B. J. Kimura, C. E. Korcarz, P. A. Pellikka, P. S. Rahko, and R. J. Siegel, 'Focused Cardiac Ultrasound: Recommendations from the American Society of Echocardiography', *Journal of the American Society of Echocardiography*, vol. 26, no. 6, pp. 567–581, Jun. 2013, doi: 10.1016/j.echo.2013.04.001.

[14] Z. Feng, J. A. Sivak, and A. K. Krishnamurthy, 'Two-stream attention spatio-temporal network for classification of echocardiography videos', *International Symposium on Biomedical Imaging*, pp. 1461–1465, Apr. 2021, doi: 10.1109/ISBI48211.2021.9433773.

[15] A. N. Neskovic *et al.*, 'Focus cardiac ultrasound core curriculum and core syllabus of the European Association of Cardiovascular Imaging', *European Heart Journal - Cardiovascular Imaging*, vol. 19, no. 5, pp. 475–481, May 2018, doi: 10.1093/ehjci/jey006.

[16] M. H. Jafari *et al.*, 'Automatic biplane left ventricular ejection fraction estimation with mobile point-of-care ultrasound using multi-task learning and adversarial training', *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, pp. 1027–1037, 2019, doi: 10.1007/s11548-019-01954-w.

[17] P. Andrus and A. Dean, 'Focused cardiac ultrasound', *Global Heart*, vol. 8, no. 4, pp. 299–303, 2013, doi: 10.1016/j.gheart.2013.12.003.

[18] '7 Benefits of Portable Ultrasound Machines', Aug. 22, 2019. https://www.nationalultrasound.com/benefits-portable-ultrasound-machines/ (accessed Dec. 20, 2021).

[19] A. J. Labovitz *et al.*, 'Focused Cardiac Ultrasound in the Emergent Setting: A Consensus Statement of the American Society of Echocardiography and American College of Emergency Physicians', *Journal of the American Society of Echocardiography*, vol. 23, no. 12, pp. 1225–1230, Dec. 2010, doi: 10.1016/j.echo.2010.10.005.

[20] D. Farsi *et al.*, 'Focused cardiac ultrasound (FOCUS) by emergency medicine residents in patients with suspected cardiovascular diseases', *Journal of Ultrasound*, vol. 20, no. 2, pp. 133–138, Jun. 2017, doi: 10.1007/s40477-017-0246-5.

[21] N. J. Soni, R. Arntfield, and P. Kory, *Point-of-Care Ultrasound*, 2nd ed. Elsevier, 2020.

[22] A. EL-Khuffash, *Neonatologist Performed Echocardiography Teaching Manual*. 2019.

[23] A. A. Mohamed, A. A. Arifi, and A. Omran, 'The basics of echocardiography', *Journal of the Saudi Heart Association*, vol. 22, no. 2, pp. 71–76, Mar. 2010, doi: 10.1016/j.jsha.2010.02.011.

[24] A. C. Y. To and L. Rodriguez, 'Fundamentals of Doppler Echocardiography'. https://thoracickey.com/fundamentals-of-doppler-echocardiography/ (accessed Dec. 12, 2021).

[25] T. R. Cawthorn *et al.*, 'Development and Evaluation of Methodologies for Teaching Focused Cardiac Ultrasound Skills to Medical Students', *Journal of the American Society of Echocardiography*, vol. 27, no. 3, pp. 302–309, Mar. 2014, doi: 10.1016/j.echo.2013.12.006.

[26] G. Lopez-Garrido, 'Self-Determination Theory and Motivation', Jan. 04, 2021. https://www.simplypsychology.org/self-determination-theory.html (accessed Dec. 18, 2021).

[27] A. H. Abdi *et al.*, 'Automatic quality assessment of apical four-chamber echocardiograms using deep convolutional neural networks', *SPIE Medical Imaging 2017: Image Processing*, vol. 10133, Feb. 2017, doi: 10.1117/12.2254585.

[28] A. H. Abdi *et al.*, 'Automatic Quality Assessment of Echocardiograms Using Convolutional Neural Networks: Feasibility on the Apical Four-Chamber View', *IEEE Transactions on Medical Imaging*, vol. 36, no. 6, pp. 1221–1230, Jun. 2017, doi: 10.1109/TMI.2017.2690836.

[29] A. H. Abdi *et al.*, 'Quality Assessment of Echocardiographic Cine Using Recurrent Neural Networks: Feasibility on Five Standard View Planes', *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Sep. 2017, doi: 10.1007/978-3-319-66179-7_35.

[30] M. Schneider *et al.*, 'A machine learning algorithm supports ultrasound-naïve novices in the acquisition of diagnostic echocardiography loops and provides accurate estimation of LVEF', *International Journal of Cardiovascular Imaging*, vol. 37, no. 2, pp. 577–586, Feb. 2021, doi: 10.1007/s10554-020-02046-6.

[31] S. Liu *et al.*, 'Deep Learning in Medical Ultrasound Analysis: A Review', *Engineering*, vol. 5, no. 2, pp. 261–275, Apr. 2019, doi: 10.1016/j.eng.2018.11.020.

[32] Z. Akkus *et al.*, 'Artificial Intelligence (AI)-Empowered Echocardiography Interpretation: A State-of-the-Art Review', *Journal of Clinical Medicine*, vol. 10, no. 7, Apr. 2021, doi: 10.3390/jcm10071391.

[33] G. Litjens *et al.*, 'State-of-the-Art Deep Learning in Cardiovascular Image Analysis', *JACC: Cardiovascular Imaging*, vol. 12, no. 8P1, pp. 1549–1565, Aug. 2019, doi: 10.1016/j.jcmg.2019.06.009.

[34] X. Gao, W. Li, M. Loomes, and L. Wang, 'A fused deep learning architecture for viewpoint classification of echocardiography', *Information Fusion*, vol. 36, pp. 103–113, Jul. 2017, doi: 10.1016/j.inffus.2016.11.007.

[35] 'Neural Networks', Aug. 17, 2020. https://www.ibm.com/cloud/learn/neural-networks (accessed Dec. 20, 2021).

[36] 'Convolutional Neural Networks', Oct. 20, 2020. https://www.ibm.com/cloud/learn/convolutional-neural-networks (accessed Dec. 20, 2021).

[37] S. Saha, 'A Comprehensive Guide to Convolutional Neural Networks - the ELI5 way', Dec. 15, 2018. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 (accessed Dec. 20, 2021).

[38] P. Baheti, 'Activation Functions in Neural Networks [12 Types & Use Cases]', Jul. 19, 2022. https://www.v7labs.com/blog/neural-networks-activation-functions (accessed Sep. 29, 2022).

[39] A. Bindal, 'Normalization Techniques in Deep Neural Networks', Feb. 10, 2019. https://medium.com/techspace-usict/normalization-techniques-in-deep-neural-networks-9121bf100d8 (accessed Sep. 29, 2022).

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet classification with deep convolutional neural networks', *Advances in Neural Information Processing Systems*, vol. 25, no. 5, Jun. 2012, doi: 10.1145/3065386.

[41] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *International Conference on Learning Representations*, Sep. 2015, [Online]. Available: http://arxiv.org/abs/1409.1556

[42] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition', *Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[43] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, 'Densely connected convolutional networks', *Conference on Computer Vision and Pattern Recognition*, vol. 2017-January, Nov. 2017, doi: 10.1109/CVPR.2017.243.

[44] C. Szegedy *et al.*, 'Going Deeper with Convolutions', *Conference on Computer Vision and Pattern Recognition*, Sep. 2015, doi: 10.1109/CVPR.2015.7298594.

[45] J. Dong *et al.*, 'A Generic Quality Control Framework for Fetal Ultrasound Cardiac Four-Chamber Planes', *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 4, pp. 931–942, Apr. 2020, doi: 10.1109/JBHI.2019.2948316.

[46] K. He, X. Zhang, S. Ren, and J. Sun, 'Identity Mappings in Deep Residual Networks', *Computer Vision – ECCV 2016*, vol. 9908, pp. 630–645, Mar. 2016, doi: 10.1007/978-3-319-46493-0_38.

[47] G. Zamzmi, L. Y. Hsu, W. Li, V. Sachdev, and S. Antani, 'Harnessing Machine Intelligence in Automatic Echocardiogram Analysis: Current Status, Limitations, and Future Directions', *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 181–203, 2021, doi: 10.1109/RBME.2020.2988295.

[48] D. G. Gungor, B. Rao, C. Wolverton, and I. Guracar, 'View Classification and Object Detection in Cardiac Ultrasound to Localize Valves via Deep Learning', *Medical Imaging with Deep Learning*, 2020.

[49] Y. Gao, Y. Zhu, B. Liu, Y. Hu, G. Yu, and Y. Guo, 'Automated recognition of ultrasound cardiac views based on deep learning with graph constraint', *Diagnostics*, vol. 11, no. 7, Jul. 2021, doi: 10.3390/diagnostics11071177.

[50] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, 'Fast and accurate view classification of echocardiograms using deep learning', *npj Digital Medicine*, vol. 1, no. 6, Dec. 2018, doi: 10.1038/s41746-017-0013-1.

[51] M. Blaivas and L. Blaivas, 'Are All Deep Learning Architectures Alike for Point-of-Care Ultrasound?: Evidence From a Cardiac Image Classification Model Suggests Otherwise', *Journal of Ultrasound in Medicine*, vol. 39, no. 6, pp. 1187–1194, Jun. 2020, doi: 10.1002/jum.15206.

[52] A. Madani, J. R. Ong, A. Tibrewal, and M. R. K. Mofrad, 'Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease', *npj Digital Medicine*, vol. 1, no. 59, Dec. 2018, doi: 10.1038/s41746-018-0065-x.

[53] H. Vaseli *et al.*, 'Designing lightweight deep learning models for echocardiography view classification', *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 10951, Mar. 2019, doi: 10.1117/12.2512913.

[54] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen, and L. Lovstakken, 'Real-Time Standard View Classification in Transthoracic Echocardiography Using Convolutional Neural Networks', *Ultrasound in Medicine and Biology*, vol. 45, no. 2, pp. 374–384, Feb. 2019, doi: 10.1016/j.ultrasmedbio.2018.07.024.

[55] A. Chartsias *et al.*, 'Contrastive Learning for View Classification of Echocardiograms', *Simplifying Medical Ultrasound. ASMUS 2021. Lecture Notes in Computer Science*, vol. 12967, Aug. 2021, doi: 10.1007/978-3-030-87583-1_15.

[56] K. Kusunose, A. Haga, M. Inoue, D. Fukuda, H. Yamada, and M. Sata, 'Clinically feasible and accurate view classification of echocardiographic images using deep learning', *Biomolecules*, vol. 10, no. 5, May 2020, doi: 10.3390/biom10050665.

[57] J. P. Howard *et al.*, 'Improving ultrasound video classification: An evaluation of novel deep learning methods in echocardiography', *Journal of Medical Artificial Intelligence*, vol. 3, Mar. 2020, doi: 10.21037/jmai.2019.10.03.

[58] Z. Ye, Y. J. Kumar, G. O. Sing, F. Song, X. Ni, and J. Wang, 'Artificial intelligence-based echocardiogram video classification by aggregating dynamic information', *KSII Transactions on Internet and Information Systems*, vol. 15, no. 2, pp. 500–521, Feb. 2021, doi: 10.3837/tiis.2021.02.007.

[59] A. I. Shahin and S. Almotairi, 'An Accurate and Fast Cardio-Views Classification System Based on Fused Deep Features and LSTM', *IEEE Access*, vol. 8, pp. 135184–135194, 2020, doi: 10.1109/ACCESS.2020.3010326.

[60] L. Zhu, Z. Xu, and T. Fang, 'Analysis of Cardiac Ultrasound Images of Critically Ill Patients Using Deep Learning', *Journal of Healthcare Engineering*, vol. 2021, 2021, doi: 10.1155/2021/6050433.

[61] R. B. Labs, M. Zolgharni, and J. P. Loo, 'Echocardiographic Image Quality Assessment Using Deep Neural Networks', *Annual Conference on Medical Image Understanding and Analysis*, vol. 12722 LNCS, pp. 488–502, 2021, doi: 10.1007/978-3-030-80432-9_36.

[62] R. B. Labs, A. Vrettos, J. Loo, and M. Zolgharni, 'Automated Assessment of Transthoracic Echocardiogram Image Quality Using Deep Neural Networks', *Intelligent Medicine*, Aug. 2022, doi: /10.1016/j.imed.2022.08.001.

[63] C. Luong *et al.*, 'Automated estimation of echocardiogram image quality in hospitalized patients', *The International Journal of Cardiovascular Imaging*, vol. 37, no. 1, pp. 229–239, Jan. 2021, doi: 10.1007/s10554-020-01981-8.

[64] A. Reinke *et al.*, 'Common Limitations of Image Processing Metrics: A Picture Story', Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.05642

[65] L. Maier-Hein *et al.*, 'Metrics reloaded: Pitfalls and recommendations for image analysis validation', Jun. 2022, [Online]. Available: http://arxiv.org/abs/2206.01653

[66] N. Park and S. Kim, 'Blurs Behave Like Ensembles: Spatial Smoothings to Improve Accuracy, Uncertainty, and Robustness', *International Conference on Machine Learning*, May 2021, [Online]. Available: http://arxiv.org/abs/2105.12639

[67] D. P. Kingma and J. Ba, 'Adam: A Method for Stochastic Optimization', *International Conference on Learning Representations*, Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.6980

[68] I. Loshchilov and F. Hutter, 'SGDR: Stochastic Gradient Descent with Warm Restarts', *International Conference on Learning Representations*, Aug. 2016, [Online]. Available: http://arxiv.org/abs/1608.03983

[69] K. He, X. Zhang, S. Ren, and J. Sun, 'Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification', *International Conference on Computer Vision*, 2015.

[70] Z. Leng *et al.*, 'PolyLoss: A Polynomial Expansion Perspective of Classification Loss Functions', *International Conference on Learning Representations*, Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.12511

[71] A. Tiulpin, 'SOLT: Streaming over Lightweight Transformations', *GitHub repository*, 2019, [Online]. Available: https://github.com/MIPT-Oulu/solt

[72] J. Wang *et al.*, 'Automated interpretation of congenital heart disease from multi-view echocardiograms', *Medical Image Analysis*, vol. 69, Apr. 2021, doi: 10.1016/j.media.2020.101942.

[73] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, 'Temporal Convolutional Networks for Action Segmentation and Detection', *Conference on Computer Vision and Pattern Recognition*, 2017, [Online]. Available: https://github.com/colincsl/

[74] A. Vaswani *et al.*, 'Attention Is All You Need', *Conference on Neural Information Processing Systems*, Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762

[75] A. Dosovitskiy *et al.*, 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', *International Conference on Learning Representations*, Oct. 2021, [Online]. Available: http://arxiv.org/abs/2010.11929

[76] J. Hu, L. Shen, and G. Sun, 'Squeeze-and-Excitation Networks', *Conference on Computer Vision and Pattern Recognition*, Jun. 2018.

[77] H. Song, M. Kim, D. Park, Y. Shin, and J. G. Lee, 'Learning From Noisy Labels With Deep Neural Networks: A Survey', *IEEE Transactions on Neural Networks and Learning Systems*, 2022, doi: 10.1109/TNNLS.2022.3152527.

APPENDICES

Appendix A

Table A.1 - Average performance of the independent networks for each task of the SX view

		Gain		Depth		Grade		Image Orientation		Apex		Chambers Alignment		Anatomical Orientation		AV Tract	
		MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1
Validation	Single	0.4424	0.6317	0.5342	0.6990	0.3823	0.5861	1.0000	1.0000	0.6546	0.8188	0.4179	0.5368	0.5267	0.6114	0.6047	0.7158
	Ensemble	0.5179	0.6720	0.5572	0.7176	0.4168	0.5879	1.0000	1.0000	0.6624	0.8261	0.3928	0.5195	0.5470	0.6275	0.5765	0.6923
Test	Single	0.4782	0.6458	0.6006	0.7561	0.4170	0.6187	1.0000	1.0000	0.5248	0.7374	0.2108	0.4346	0.4785	0.5894	0.4702	0.5800
	Ensemble	0.5378	0.6915	0.6434	0.7861	0.4404	0.6308	1.0000	1.0000	0.5400	0.7438	0.2558	0.4595	0.5176	0.6207	0.4878	0.5769

Table A.2 - Average performance of the independent networks for each task of the A4C view

		Gain		Depth		Grade		Image Orientation		Anatomical Orientation		Chambers Alignment		Right Ventricle		Windowing Position		Near-field Artifact	
		MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1
Validation	Single	0.5279	0.6972	0.5618	0.7170	0.6219	0.7467	0.9888	0.9947	0.6667	0.7509	0.4745	0.6323	0.4487	0.6812	0.5281	0.6156	0.5646	0.6642
	Ensemble	0.5586	0.7149	0.5825	0.7272	0.6160	0.7431	0.9889	0.9947	0.6914	0.7692	0.4535	0.6190	0.4629	0.6887	0.4926	0.5846	0.6012	0.6914
Test	Single	0.5564	0.7255	0.6171	0.7519	0.4616	0.6581	0.9901	0.9950	0.7645	0.8124	0.3826	0.6570	0.6305	0.7442	0.5564	0.6507	0.5293	0.6413
	Ensemble	0.5816	0.7476	0.6446	0.7699	0.4628	0.6586	0.9901	0.9950	0.7755	0.8211	0.4151	0.6740	0.6369	0.7480	0.5861	0.6744	0.5533	0.6517

Table A.3 - Average performance of the independent networks for each task of the IVC view

		Gain		Depth		Grade		Image Orientation		Suprahepatic		Confluence		Center	
		MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1
Validation	Single	0.5992	0.7506	0.5227	0.5873	0.3809	0.5793	0.9259	0.9653	0.5015	0.8361	0.4892	0.6536	0.5379	0.7157
	Ensemble	0.5925	0.7426	0.5384	0.5171	0.3811	0.5761	0.9335	0.9688	0.5455	0.8525	0.4748	0.6446	0.5405	0.7162
Test	Single	0.5581	0.6856	0.6110	0.5458	0.3671	0.5776	0.9814	0.9911	0.6396	0.8584	0.4838	0.6723	0.5857	0.7310
	Ensemble	0.5459	0.6761	0.6387	0.5557	0.3739	0.5932	0.9777	0.9892	0.7305	0.8959	0.5059	0.6822	0.6161	0.7553

Appendix B

Table B.1 - Average performance of the implemented networks for each task of the SX view

		Gain		Depth		Grade		Image Orientation		Apex		Chambers Alignment		Anatomical Orientation		AV Tract	
		MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1
Random Initialisation	Independent	0.5378	0.6915	0.6434	0.7861	0.4404	0.6308	1.0000	1.0000	0.5400	0.7438	0.2558	0.4595	0.5176	0.6207	0.4878	0.5769
	Multi-task 1	0.4224	0.6149	0.6309	0.7742	0.4584	0.6477	0.9851	0.9934	0.5231	0.7424	0.1239	0.3797	0.4608	0.5806	0.4726	0.6000
	Multi-task 2	0.4542	0.6315	0.6223	0.7710	0.4318	0.6451	0.9561	0.9799	0.5231	0.7424	0.1462	0.3896	0.5207	0.6182	0.3945	0.5429
Transfer Learning	Independent	0.4106	0.6137	0.4366	0.6401	0.2867	0.5300	0.9850	0.9935	0.5313	0.7419	0.3841	0.5714	0.4895	0.6032	0.5013	0.6102
	Multi-task 1	0.3542	0.5797	0.4813	0.6763	0.3625	0.5729	0.9550	0.9801	0.5471	0.7538	0.2431	0.4615	0.5658	0.6400	0.5131	0.6230
	Multi-task 2	0.4171	0.5938	0.5157	0.7270	0.3672	0.5782	0.9550	0.9801	0.5953	0.7786	0.3279	0.5195	0.4668	0.5600	0.5345	0.6316

Table B.2 - Average performance of the implemented networks for each task of the A4C view

		Gain		Depth		Grade		Image Orientation		Anatomical Orientation		Chambers Alignment		Right Ventricle		Windowing Position		Near-field Artifact	
		MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1
Random Initialisation	Independent	0.5816	0.7476	0.6446	0.7699	0.4628	0.6586	0.9901	0.9950	0.7755	0.8211	0.4151	0.6740	0.6369	0.7480	0.5861	0.6744	0.5533	0.6517
	Multi-task 1	0.5080	0.7114	0.6629	0.7781	0.5318	0.7077	0.9706	0.9849	0.6823	0.7500	0.3614	0.6486	0.5676	0.7068	0.5019	0.6000	0.5413	0.6353
	Multi-task 2	0.4723	0.6777	0.6118	0.7491	0.5408	0.7127	0.8870	0.9368	0.6632	0.7347	0.3853	0.6631	0.6487	0.7597	0.5062	0.5974	0.6217	0.6835
Transfer Learning	Independent	0.3689	0.5960	0.5862	0.7254	0.5280	0.7019	1.0000	1.0000	0.7280	0.7865	0.3943	0.6552	0.5955	0.7259	0.5578	0.6410	0.6141	0.6905
	Multi-task 1	0.4168	0.6326	0.6493	0.7697	0.5372	0.7106	0.9610	0.9798	0.7057	0.7692	0.4135	0.6806	0.5591	0.7015	0.5524	0.6420	0.4378	0.5618
	Multi-task 2	0.4361	0.6516	0.6230	0.7455	0.5768	0.7364	0.9610	0.9798	0.6772	0.7473	0.4460	0.6952	0.5624	0.7050	0.5402	0.6111	0.5486	0.6341

Table B.3 - Average performance of the implemented networks for each task of the IVC view

		Gain		Depth		Grade		Image Orientation		Suprahepatic		Confluence		Center	
		MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1
Random Initialisation	Independent	0.5459	0.6761	0.6387	0.5557	0.3739	0.5932	0.9777	0.9892	0.7305	0.8959	0.5059	0.6822	0.6161	0.7553
	Multi-task 1	0.4845	0.6513	0.7236	0.5806	0.4662	0.6445	0.9888	0.9947	0.6473	0.8661	0.5267	0.7101	0.6533	0.7735
	Multi-task 2	0.5285	0.6712	0.7507	0.5885	0.5109	0.6725	0.9560	0.9783	0.6473	0.8661	0.5113	0.6963	0.6063	0.7647
Transfer Learning	Independent	0.5052	0.6841	0.5800	0.5285	0.3741	0.5805	0.9777	0.9892	0.6161	0.8436	0.5813	0.7368	0.5630	0.7274
	Multi-task 1	0.5815	0.7278	0.7048	0.5788	0.4055	0.5840	0.9777	0.9892	0.5647	0.8295	0.5590	0.7259	0.6073	0.7431
	Multi-task 2	0.5624	0.7052	0.6080	0.5395	0.4609	0.6220	0.9777	0.9892	0.5988	0.8482	0.6433	0.7794	0.6287	0.7629