Automated Assessment of Coronary Artery Stenosis in X-ray Angiography using Deep Neural Networks

Dinis Lourenço Tavares Rodrigues

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Prof. Arlindo Manuel Limede de Oliveira
Prof. Mário Alexandre Teles de Figueiredo

Examination Committee

Chairperson: Prof. José Eduardo Charters Ribeiro da Cunha Sanguino
Supervisor: Prof. Arlindo Manuel Limede de Oliveira
Member of the Committee: Prof. Alexandre José Malheiro Bernardino
Prof. Miguel Nobre Menezes

February 2020

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

Ao Professor Arlindo Manuel Limede de Oliveira por me ter dado esta oportunidade única de trabalhar num tema tão extraordinário e por todo o apoio ao longo desta dissertação. Ao Professor Mário Alexandre Teles de Figueiredo por toda a orientação e conselhos. Ao Professor Miguel Nobre Menezes por toda disponibilidade e trabalho conjunto no tratamento de dados. Por fim, ao meu pai António Manuel Tavares Rodrigues e à minha mãe Otília Maria Moura Lourenço Rodrigues, por toda a vossa dedicação em mim.

Resumo

Existem vários métodos para avaliação quantitativa da gravidade de uma estenose da artéria coronária, bem como diferentes medidas, levando a uma gestão distinta dos procedimentos de tratamento. É de extrema importância identificar e classificar adequadamente todas as estenoses de um indivíduo. Uma implementação de uma estrutura de três etapas utilizando aprendizagem profunda foi projetada para automatizar a detecção e avaliação da gravidade da estenose. Este estudo apresenta um novo conjunto de dados clinicamente obtidos de sequências de angiografia coronária invasiva de raios-X (ACI) devidamente desidentificadas de 480 pacientes do Hospital de Santa Maria. Para cada sequência, imagens com contraste radio-opaco foram anotadas, definindo a visibilidade total da estenose. Caixas delimitadoras da estenose foram anotadas por um médico especialista em imagens de referência seguido por técnicas de processamento de imagem para propagação das caixas em cada imagem. Dinâmica de transferência de conhecimento de redes neuronais profundas são exploradas para aprendizagem supervisionada em cada etapa. Com aplicação de redes neuronais convolucionais para a seleção do ângulo correspondente da Artéria Coronária Esquerda / Direita (ACE / ACD) atingindo uma exatidão de 0,97. Detectores de disparo único são utilizados para a detecção de estenose atingindo 0,83/0,81 mAR para ACE/ACD respectivamente. Uma nova abordagem de reforço de região de interesse com CNN para regressão da gravidade da estenose do RCA foi explorada. O nosso método demonstra a importância de transferência de aprendizagem na avaliação da gravidade da estenose com dados limitados, alcançando bons desempenhos. Para o melhor do conhecimento do autor, esta é a primeira vez que o iFR foi usado como uma métrica para tarefas de avaliação automática da gravidade da estenose, usando técnicas de aprendizagem profunda.

Palavras-chave: Doença de Artéria Coronária (DAC), Rede Neural Convolucional (RNC), Angiografia Coronária Invasiva (ACI), Deteção Automática de Estenose, Classificação de Imagem, Transferência de Aprendizagem

Abstract

Several methods for quantitative severity assessment of coronary artery stenosis exist as well as different measures, leading to distinct management of treatment procedures. It is of upmost importance to properly identify and classify all possible stenosis on an individual. A deep-learning three-step framework implementation was designed to automate the detection and assessment of stenosis severity. This study showcases a new clinically obtained dataset of properly de-identified X-ray invasive coronary angiography (ICA) sequences of 438 patients from *Hospital de Santa Maria*. For each sequence, radio-opaque contrast filled frames were annotated, defining full stenosis visibility with stenosis bounding boxes being annotated by an expert physician on reference frames followed by image processing techniques for propagation at each frame. Transfer learning dynamics of deep neural networks are exploited for supervised learning at each step, employing CNN's for angle view selection of the Left/Right Coronary Artery (LCA/RCA) achieving 0.97 Accuracy, single-shot detectors for stenosis detection achieving 0.83/0.81 mAR for LCA/RCA respectively and a new region of interest boost approach with CNN's for stenosis severity regression of the RCA was explored. Our method showcases the importance of transfer learning in stenosis severity assessment with limited data, achieving considerable performances. To the best of the author's knowledge, this is the first time that iFR was used as a metric for stenosis severity assessment tasks using deep learning techniques.

Keywords: Coronary Artery Disease (CAD), Convolutional Neural Network (CNN), Invasive Coronary Angiography (ICA), Stenosis Detection, Image Classification, Transfer Learning

Contents

List of Tables

List of Figures

Acronyms

CAD Coronary Artery Disease.

CHD Coronary Heart Disease.

CNN Convolutional Neural Network.

COCO Common Object in Context.

DICOM Digital Imaging and Communications in Medicine.

FFR Fractional Flow Reserve.

FN False Negative.

FP False Positive.

GUI Graphical User Interface.

ICA Invasive Coronary Angiography.

iFR instantaneous Wave-free Ratio.

IHD Ischemic Heart Disease.

IoU Intersection-over-Union.

LAD Left Anterior Descending.

LCA Left Coronary Artery.

LCx Left Circumflex.

mAP mean-Average Precision.

mAR mean-Average Recall.

MLP Multi Layer Perceptron.

NCD Non-Communicable diseases.

NMS Non-Maximum Suppression.

QCA Quantitative Coronary Angiography.

RCA Right Coronary Artery.

RMA Right Marginal Artery.

RPD Right Posterior Descending.

SGD Stochastic Gradient Descent.

TN True Negative.

TP True Positive.

Chapter 1

Introduction

1.1 Motivation

There's been a quest to understand the proper functioning and morphology of the human heart. Diagnosis quality on Coronary Artery Disease (CAD) has been shifting across decades, and it's clearly shown that further research in this subject is increasing.

In the latest published study from the Global Burden of Diseases, Injuries, and Risk Factors (GBD), cardiovascular diseases, including coronary artery disease, is the leading non-communicable disease (NCD) in regards to global mortality accountability. Responsible for an estimated 17.8 million deaths in 2017, showing an increase of 21.1% in occurrences, from 2003 to 2017, with groups of low and middle income being the most affected [1].

Although several resources have been invested in prevention, proper available CAD assessment and treatment procedures still aren't reachable to the most general public due to the high costs involved.

The interest in producing reliable artificial intelligent applications and automated procedures has been increasing in this new technology era. Clear evidence that proper automated and reliable ways of quantifying stenosis lesions are being applied is lacking. Motivated by the opportunity to contribute to the medical research field and the general public health care, with the application of state-of-the-art technologies, an implementation using a three-step framework is proposed. The proposed solution uses a deep convolutional neural network to automatically identify invasive coronary angiography (ICA) angle views, and to detect and estimate stenosis severity. The approach can be used to aid expert physicians in stenosis severity assessment procedures reducing clinical malpractices [2] and improving patient welfare during procedures, by reducing the time complexity.

1.2 Topic Overview

Coronary artery disease (CAD), also known as coronary heart disease (CHD) or ischemic heart disease (IHD), [3] is characterized by plaque buildup, a waxy substance, inside the coronary arteries. This buildup can partially or totally block blood flow in the coronary arteries, leading to improper delivery of oxygen-rich blood to the heart, weakening the heart muscle, and possibly leading to heart failure [4]. Current standard diagnosis methods rely on an expert physician to assess the issue, off or on-site, using non-invasive or invasive procedures.

1.3 Objectives

This study uses a new curated medical dataset of X-ray invasive coronary angiography with optimal interval frames annotated with stenosis bounding boxes and labeled with a quantitative iFR severity assessment measure, providing a path for novel implementations of automatic stenosis assessment. Additionally, a three-step framework based on deep neural networks is presented for coronary angle view selection, stenosis detection and stenosis quantitative severity assessment.

1.4 Thesis Outline

The present document is divided into eight chapters:

- **Chapter 2 State-of-the-art**: An overview of previous work carried in CAD assessment.
- **Chapter 3 Coronary Artery Disease**: Describes the overall notations, definitions, and issues regarding CAD assessment.
- **Chapter 4 Medical Data**: Presents the processing steps of the applied medical dataset in this study.
- **Chapter 5 Implementation**: Describes a three-step framework for automatic stenosis assessment using deep neural networks.
- **Chapter 6 Results**: Showcases the performance metrics of each step of the framework and presents comparisons with related work by other authors.
- **Chapter 7 Conclusions**: An overview through the completed study and contributions.
- **Chapter 8 Future Work**: A brief discussion on how to improve the current medical dataset and possible implementations based on segmentation and attention mechanisms, with the objective of improving stenosis detection and severity assessment.

Chapter 2

State-of-the-Art

Subjective visual estimation of CAD severity by visually assessing coronary angiograms was standard practice for several years until it was shown to be inadequate due to high degrees of intra-observer and inter-observer variability [5]. Significant efforts have been made to develop new methods and algorithms to objectively quantify the severity of CAD through quantitative coronary angiography (QCA) [6]. QCA starts with an invasive coronary angiography procedure (ICA), where the introduction of a contrast medium through heart catheterization is performed. Different processes exist, but the core procedure remains the same. A reference two-dimensional ICA still image is chosen, normally in the end-diastole, where the coronary arteries are more expanded. The image is forwarded to a countour-detection algorithm. Together with several other parameters, QCA can quantify the severity of the stenosis by assigning a percentage metric where 0% indicates no blood flow and 100% indicates perfect flow [7].

In 2014 the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS) published their guidelines where a careful evaluation on this issue with all evidence available at the time was made. The guidelines aim to assist health professionals in selecting the best management strategies for an individual patient with a given condition. QCA (previously recommended) in these guidelines was no longer favored for stenosis assessment through ICA, but rather a new upcoming procedure called fractional flow reserve (FFR)[8]. Invasive measurement through FFR is performed by placing clinically suited wires (connected to an external system) with a pressure sensor in the coronary artery. With the pressure sensor, pressure values are recorded during several heartbeats. The result is a pressure ratio between the distal and proximal location of the stenosis were a value below 0.71 indicates abnormal blood flow with a need for further treatment procedure [9].

In 2018 a new version of ESC/EACTS guidelines was published, introducing an equally recommended procedure for quantitative stenosis assessment [10]. The instantaneous wave free ratio (iFR) is a procedure that, contrary to FFR, needs no adenosine administration, a vasodilator that regulates the heart rhythm. The principle remains the same: by placing clinically suited wires in the coronary arteries, it results in a ratio between the stenosis's proximal and distal pressures, where a value below 0.90 indicates abnormal blood flow and a need for further treatment procedure. Instead of recording several pressure values like in FFR, it can provide an accurate quantitative measure by analyzing the heart in a specific cardiac cycle, in the diastolic period, which is characterized by being very stable with minimum microvascular resistance [11].

With more access to computing power, it opened a path that's been leading to increased neural networks and convolutional neural network (CNN) applications. Early developments in deep learning techniques applied in cardiology imaging began by focusing on the structural segmentation of the ventricles and stenosis centerline extraction from MRI and CT scans with the objective of aiding expert physicians in visual assessment [12–16].

To deal with the lack of public ICA datasets Antczak and Liberadzki [17] generated and trained a custom CNN on thousands of artificial 32 by 32 pixel patches mimicking the presence of stenosis. Using a sliding window with the patches dimensions on the original frame with the trained CNN, detection performance increased, but real test images were very few and were also scaled down.

To automate the process from start to finish in stenosis assessment Au et al. [18] showcased a pipeline composed by three uniquely designed CNN's with the intention of detecting, segmenting and classifying stenosis severity through QCA annotations in ICA reference images of the left coronary artery (LCA). Their study included 1024 study participants using only RCA viewing angles and reference frames. A detector variant of the single-shot detector YOLO [19] was developed with the objective of determining fixed dimensional regions of 192 by 192 pixels on which a stenosis was present. With the proposed region another custom segmentation deep learning architecture was built, based on U-Net [20], to automatically segment every pixel where the stenosis was present. Afterwards another but yet small custom CNN with only five convolutional layers was built to classify the segmented frame.

Cong et al. [21] also developed a three-stage end-to-end workflow for stenosis characterization. With a dataset wich included RCA and LCA sequences from 194 patients, the process of viewing angle selection is initialized with transfer learning and fine-tuning of the InceptionV3 [22]. The reference frame selection in each ICA sequence, which lies under the optimal frame interval, when the radio opaque contrast is fully introduced and all coronary arteries and stenosis are best seen, it's usually selected by an expert physician. Features extracted from the last convolution layer of the InceptionV3, are used to train a bi-directional LSTM, taking advantage of the temporal dimension to extract the exact frame of the sequence corresponding to the reference frame. With the extracted frame another InceptionV3 is fine-tuned in a classification manner for the stenosis assessment under QCA labels. The detection of the stenosis is then performed as a weakly supervised method by employing class activation maps using Grad-CAM [23] to identify the most important regions based on the weights contribution for the respective frame classification result. These detections are then evaluated against expert physician manual annotations of 35 by 35 pixel bounding boxes.

Focusing in the detection task of the stenosis Wu et al. [24] developed a novel single-shot architecture using the VGG16, a feature extractor from which feature maps from low and high level convolutional layers are extracted. Those are then passed into a classification and regression sub network, to estimate bounding box coordinates and the respective confidence scores. Their dataset constituted ICA sequences from a different range of viewing angles of 134 participants where only bounding boxes for verified stenosis above the QCA threshold were annotated and compared against estimations.

Chapter 3

Coronary Artery Disease

3.1 Characterization

The heart muscle, like any other organ or tissue in the human body, needs oxygen-rich blood to survive and endure. Blood is supplied to the heart by its own vascular system, named coronary circulation. The aorta (the main blood supplier to the body) branches into two main coronary arteries. These coronary arteries then also branch into smaller arteries, which supply oxygen-rich blood to the entire heart muscle [25].

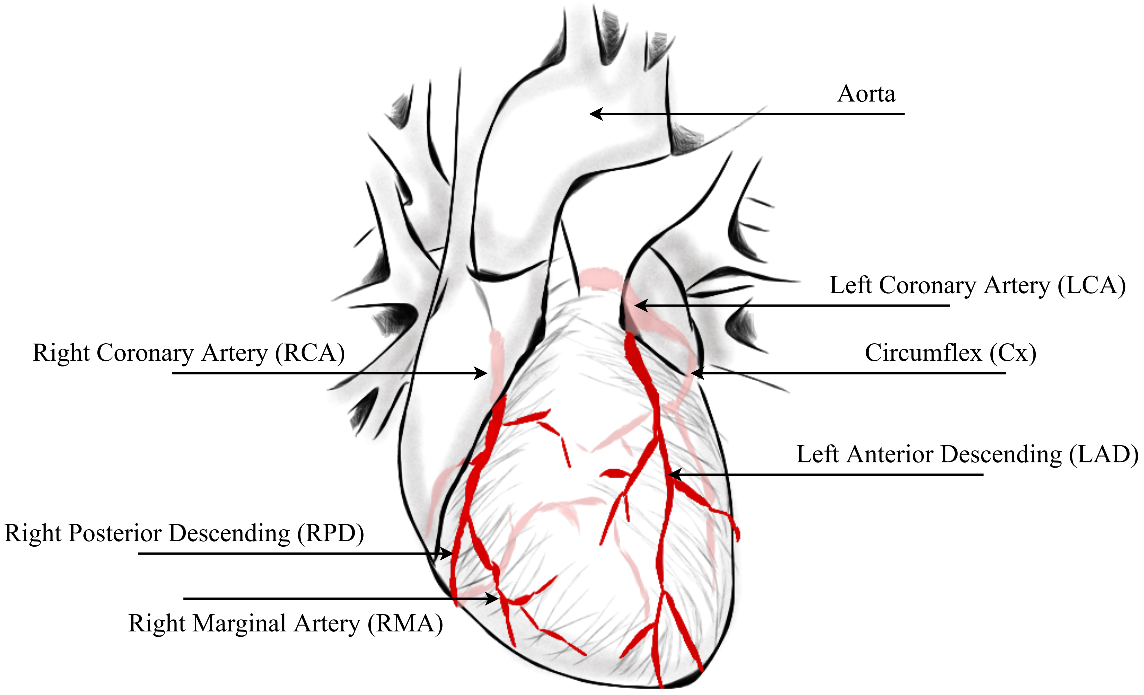

Aorta
Left Coronary Artery (LCA)
Right Coronary Artery (RCA)
Circumflex (Cx)
Left Anterior Descending (LAD)
Right Posterior Descending (RPD)
Right Marginal Artery (RMA)

Figure 3.1: Coronary arteries illustration.

The left main coronary artery (LCA) supplies blood to the left side of the heart muscle (the left ventricle and left atrium). It then branches into the left anterior descending (LAD) artery and to the left circumflex (LCx) artery encircling the heart. The right coronary artery (RCA) supplies blood to the right ventricle, the right atrium, and the SA (sinoatrial) and AV (atrioventricular) nodes, which regulate the heart rhythm. It branches into the right posterior descending artery (RPD) and the right marginal artery (RMA). Together with the left anterior descending artery, the right coronary artery helps to supply blood to the middle or septum of the heart [26].

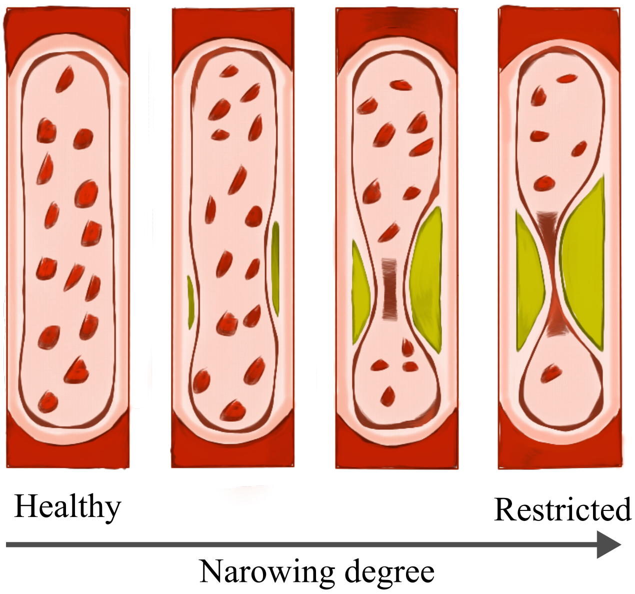


Figure 3.2: Fatty deposits restricting blood flow, narrowing the coronary artery.

CAD is characterized by the narrowing (stenosis) or blockage of the coronary arteries due to the buildup of cholesterol and fatty deposits on the inner walls of the arteries. The stenosis may restrict blood flow to the heart muscle by clogging the artery thus causing abnormal artery function [27]. These can show up in multiple scenarios, i.e multiple coronary arteries and respective branches.

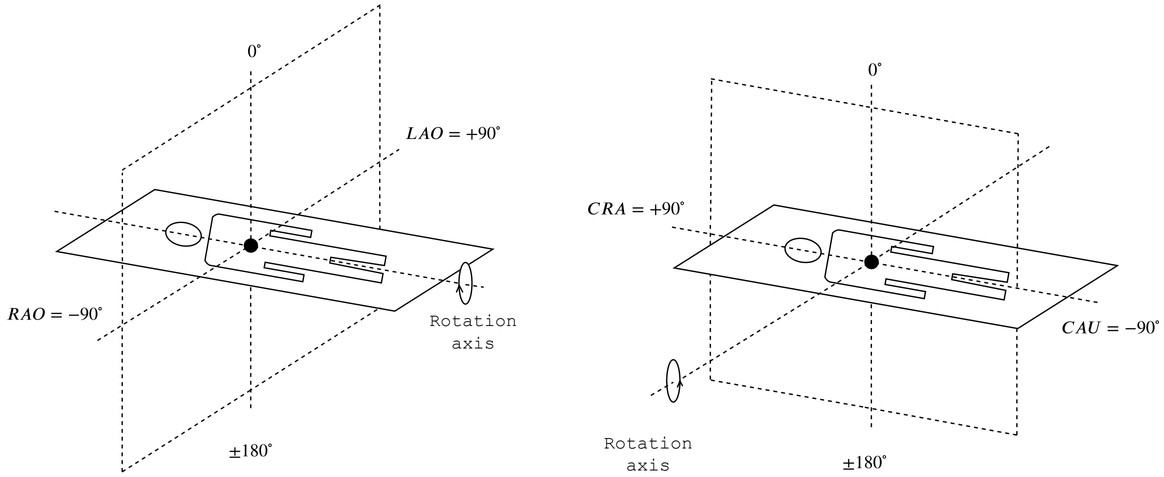
Untreated cases may lead to stable and unstable angina (chest pain), myocardial infarction (heart attack), and sudden cardiac death (absence of blood flow)[28]. To prevent further CAD advances, correct stenosis diagnosis is important in order to help in the definition further steps of treatment.

3.2 Stenosis Assessment

3.2.1 Coronary Angiography

The current standard in coronary artery stenosis evaluation is invasive coronary angiography (ICA) [11]. In this procedure, the patient lies on an X-ray table and a radio-opaque contrast agent is administered intravenously through a catheter directly into the coronary arteries by an expert physician. With the assistance of a C-arm X-ray unit that moves rotationally in two perpendicular planes (see Figure 3.3), two-dimensional images are then taken, to build a temporal sequence of frames recording the entire procedure and capturing the radio-opaque contrast flow.

The different retrieved image sequences are then individually separated and visually assessed by



(a) Right Anterior Oblique (RAO), right side of the patient, to Left Anterior Oblique (LAO), left side of the patient

(b) Cranial (CRA), upper side of the patient, to Caudal (CAU), lower side of the patient

Figure 3.3: Rotational perpendicular planes of C-arm X-ray unit.

the expert for a first evaluation, by observing the radio-opaque contrast flow, which translates to the blood flow in the coronary arteries, looking for any abnormal behavior. For confirmation, a quantitative stenosis severity procedure is followed.

Coronary System	Angle (degrees)		Observed Arteries
	LAO	CRA	
LCA	-20	20	LAD and LCx
	0	40	LAD and LCx
	0	-30	LAD
	50	-30	LAD and LCx
	50	30	LAD
RCA	30	0	RCA
	-30	0	RCA
	0	30	RPDA and RMA

Table 3.1: Common viewing angles in ICA.

Multiple occurrences of stenosis can show in coronary arteries and depending on the patient's morphology, the expert physician may move the C-arm X-ray unit in different angles in order to reach and visualize certain coronary arteries where the most common angles can be summed in Table 3.1 [29, 30].

3.2.2 Quantitative Coronary Assessment

QCA is a computer-assisted procedure widely considered as a gold standard for measuring stenosis percentage. It is typically performed either 'online' immediately after coronary angiography for clinical decision making or 'offline' by angiographic core laboratory experts (for clinical trials). It involves visually annotating diseased coronary arterial segments and the area surrounding each stenosis to determine the percent diameter stenosis. Standard workflow for QCA consists of a multi-step analytical pipeline. Single-frame images that best demonstrate the stenosis are selected by the analyst. Images are then manually annotated by segmenting a portion of the lesion (i.e. stenosis) of interest, consisting of the lesion as well as the surrounding healthy vessel. Finally, annotated lesions are analyzed for percent diameter stenosis relative to the reference vessel diameter of the lesion of interest [31].

3.2.3 Fractional Flow Reserve

Invasive measurement through Fractional Flow Reserve (FFR) is performed by placing clinically suited wires (connected to an external system) with a pressure sensor in the coronary artery. Through the time of several heartbeats pressure values are recorded with the pressure sensor. The result is a pressure ratio between the distal and proximal location of the stenosis where a value below 0.71 indicates abnormal blood flow with a need for further treatment procedure [32].

3.2.4 Instantaneous Wave-Free Ratio (iFR)

Instantaneous wave-free ratio (iFR) assessment is based on the physical law described by the Hagen-Poiseuille equation, giving the pressure drop of Newtonian fluids with constant density in laminar flow, running in a constant cross-section long cylindrical pipe. This law can be seen as a deviation from Ohms Law ($U = RI$) with

$$P = Q \times R \tag{3.1}$$

where P, Q, and R is the respective pressure, flow, and resistance. By discarding many variations, the problem can be simplified to a very simple model (see Figure 3.4) in order to obtain the pressure drop caused by the stenosis

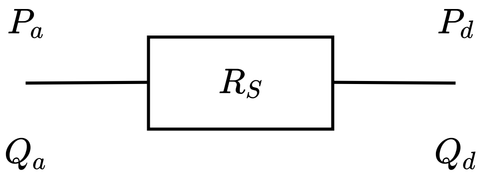


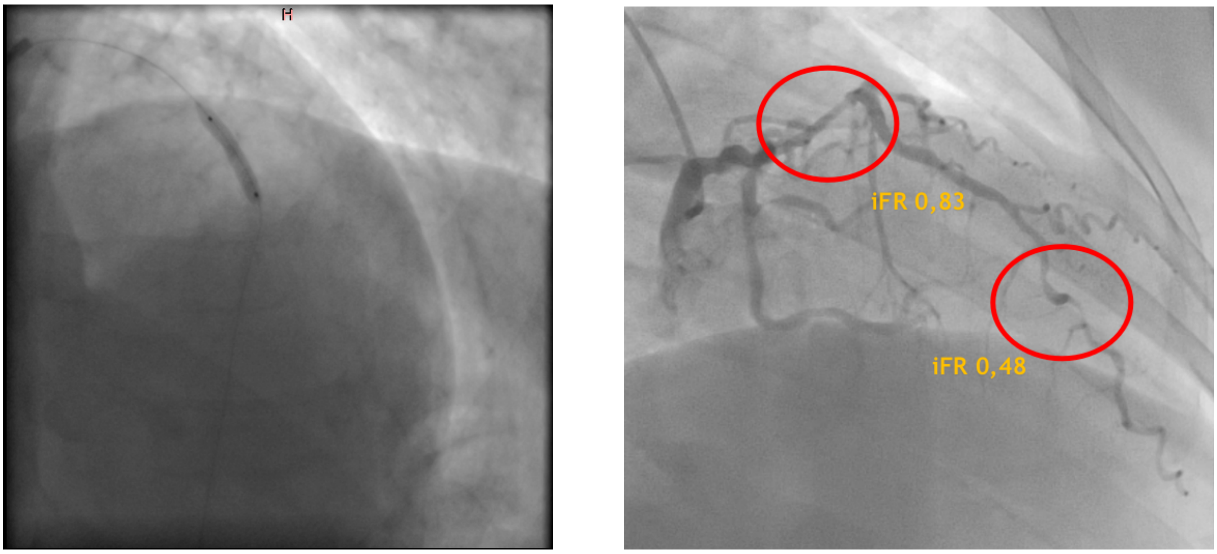
Figure 3.4: Simplified iFR measurement model.

where the total pressure drop across the stenosis is given by

$$\begin{gathered} P_a - P_d = (Q_a - Q_d) \times R_S \\ \text{Pressure change} = \text{Flow change} \times \text{Constant Resistance} \\ \Delta P \approx \Delta Q \times R_s \end{gathered} \tag{3.2}$$

with P_a being the proximal pressure (beginning), P_d the distal pressure (end) and R_S the constant resistance given by the stenosis. [11, 33, 34].

When the stenosis limits the blood flow, the pressures P_d and P_a differ over the wave-free period. iFR is performed through an invasive procedure, by placing medical suited wires with a pressure sensor from the proximal to the distal coronary stenosis (see Figure 3.5). The pressure values are then computed by proper medical systems and average the ratio P_a/P_d from usually five cardiac cycles in the diastolic period, with a minimum of one. When we are in the presence of normal blood flow, iFR has a ratio of 1.0 Any value below 0.90 will indicate abnormal flow and consequently further treatment procedure indication [35].



(a) Clinical suited wire placement

(b) iFR assessed quantitative value

Figure 3.5: iFR proceadure

From the coronary angiography to the quantitative iFR measurement, the procedure takes an average time of one hour, with the coronary angiography taking only about ten minutes to visualize every coronary artery.

3.2.5 Issues

The objective of the iFR measure is to quantify the severity of the stenosis. The expert then determines whether the patient needs to proceed to further treatment procedures or not. The coronary angiography takes only a fraction of the total assessment procedure time. The expert physician needs to evaluate each individually coronary artery and correspondingly determine each iFR value. So in sum: (1) the patient endures the coronary angiography, (2) the iFR measurement, and (3) if necessary the treatment procedure time.

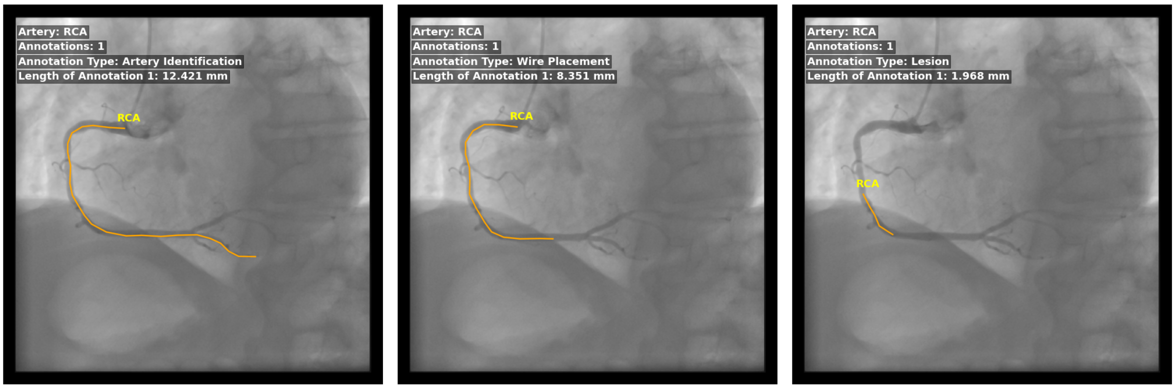
Chapter 4

Medical Data

4.1 Overview

The available data for this work was firstly curated by cardiologist Miguel Nobre Menezes from *Unidade de Cardiologia de Intervenção Joaquim Oliveira, Serviço de Cardiologia* from *Hospital de Santa Maria, Centro Hospitalar Lisboa Norte*. It is composed of 9378 clinically obtained invasive coronary angiography single and multiple ICA image sequences of 438 subjects, ranging from the year 2015 until 2019. The data was properly de-identified to preserve participant privacy, and each subject was over the age of eighteen.

For each subject, in addition to the ICA image sequences themselves, the *Gender*, *Age*, *Date*, *iFR Value*, *FFR Value* and *Coronary Artery Stenosis Location* was also included at the patient level. The *Gender*, *Age* and *Date* are considered only as auxiliary metadata and do not serve as features for this study.


Artery: RCA
Annotations: 1
Annotation Type: Artery Identification
Length of Annotation 1: 12.421 mm
RCA
Artery: RCA
Annotations: 1
Annotation Type: Wire Placement
Length of Annotation 1: 8.351 mm
RCA
Artery: RCA
Annotations: 1
Annotation Type: Lesion
Length of Annotation 1: 1.968 mm
RCA

(a) Artery Annotation (b) Wire Placement Annotation (c) Stenosis Annotation

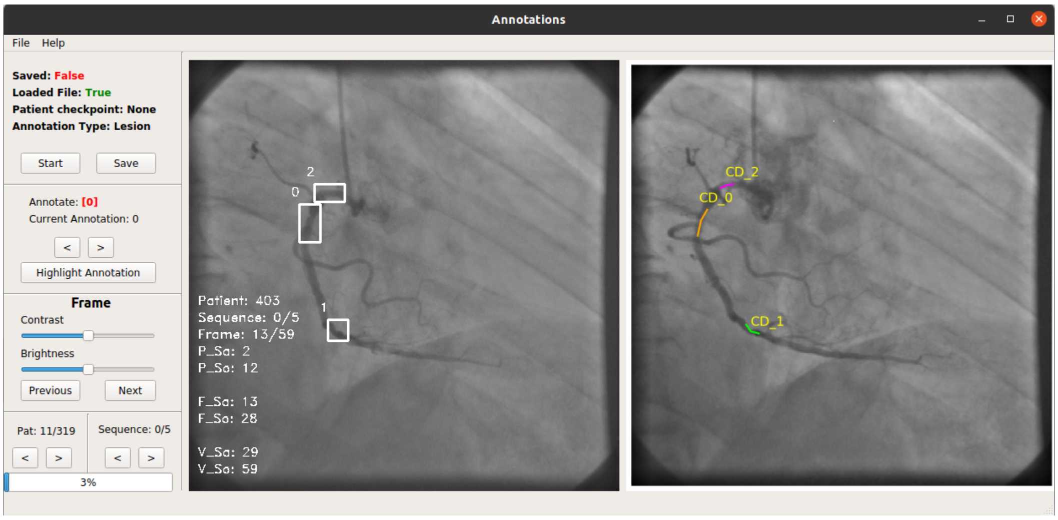
Figure 4.1: Annotations in RCA viewing angle by frame procedure indicating the full coronary artery, wire placement from which the iFR was obtained and the most visually contributing stenosis.

For the stenosis severity assessment, the iFR procedure was preformed for the majority of the patients. As for the FFR, the procedure was in some cases performed together with iFR, but overall there are insufficient assessments, resulting in a non-present value for the respective patients.

The *Coronary Artery Stenosis Location* information indicates which artery had the most contributing stenosis for which the iFR and FFR values were obtained.

For each patient with a valid iFR assessment value and stenosis location, annotations for the optimal sequences, i.e., stenosis is best seen under the radio-opaque contrast, were included in a non-destructive way using *Osirix* [36], an image processing software (see Figure 4.1).

Several stenosis occurrences may happen at any given sequence and patient, whether on the same coronary artery or multiple coronary arteries. In some viewing angles, different coronary arteries can also be seen simultaneously and the stenosis themselves.


Saved: False
Loaded File: True
Patient checkpoint: None
Annotation Type: Lesion

Annotate: [0]
Current Annotation: 0

Pat: 11/319
Sequence: 0/5
3%

Patient: 403
Sequence: 0/5
Frame: 13/59
P_Sa: 2
P_So: 12

F_Sa: 13
F_So: 28

V_Sa: 29
V_So: 59

Figure 4.2: Custom annotation graphical user interface program.

For the optimal sequences, annotations were done such that for each coronary artery containing a stenosis: (1) a unique frame was annotated showcasing the artery; (2) a unique frame was annotated showcasing how the wire of the iFR procedure was placed; (3) a unique frame was annotated showcasing all the stenosis of the corresponding artery, corresponding to the best contrast viewing frame and it is considered the reference frame of the sequence. A total of 1593 sequences accounting for 438 patients were annotated using this procedure.

4.2 Data Treatment and Annotation Process

The provided ICA image sequences were in the well known and documented DICOM format protocol.

	DICOM Metadata	
Identifier	Detail	Desired
Image Type	Image identification characteristics	No Protocols/Reports
Rows	Number of pixel rows in the frame	512
Columns	Number of pixel columns in the frame	512
Frame Time	Time interval in milliseconds between frames	33 and 66
Pixel Spacing	Physical distance (mm) between adjacent rows and columns centers	1
Photometric Interpretation	Intended interpretation of the pixel data	Monochromatic
Positioner Primary Angle	Position of the C-arm X-Ray unit from the RAO to LAO	Any
Positioner Secondary Angle	Position of the C-arm X-Ray unit from CRA to CAU	Any

Table 4.1: Relevant DICOM metadata of ICA sequences.

The dataset contained variations in dimensions, frame rates, and unique cases that would affect the result. Many of the DICOM files were patient study reports, protocols, and single-frame shots, which are not directly usable. Following the criteria in Table 4.1a first step in standardizing the dataset only through the DICOM metadata was performed. Only sequences with frame dimensions of 512 by 512 pixel were allowed with pixel values ranging in the $[0, 255]$ monochromatic scale (1 channel). Rare cases of 1024 by 1024 pixel were present having pixel values outside the allowed range. As a dimensional down-scaling and a transformation in pixel value would be necessary, those were discarded.

From the single frame annotations in the optimal sequences (see Figure 4.1) an annotation framework was developed to enrich the dataset with metadata. Starting with the optimal frame annotated by the expert physician, they are automatically transformed into bounding boxes and propagated through the entire sequence. Then it is possible to adjust the bounding boxes, define the sequence optimal intervals (when the contrast is fully introduced), group-specific angle views of the sequences, and discard unwanted sequences with image artifacts. A custom graphical user interface (GUI) (see Figure 4.2) was specifically developed to suit the requirements necessary for this task.

4.2.1 Bounding box propagation

Before propagating the annotations, they are first transformed into bounding boxes, which is done by taking the start/end coordinates and the original annotation's maximum width/height coordinates. With this information, any rectangular shape can be made and is considered a bounding box. Note that these bounding boxes are generated only in the reference frame of the sequence.

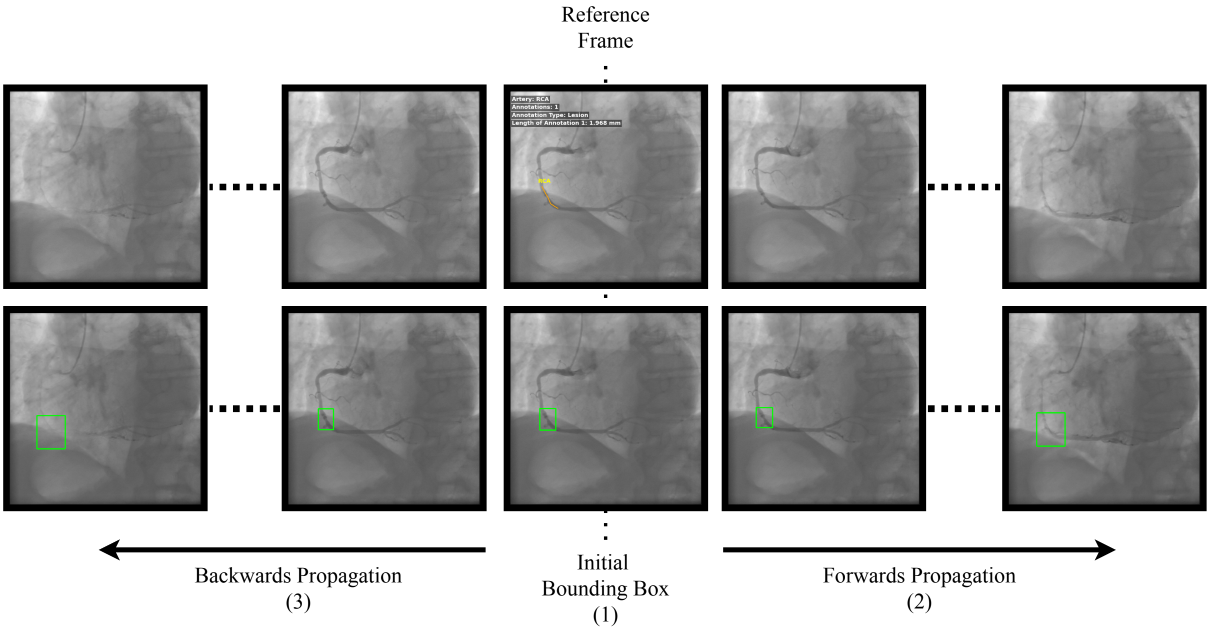


Figure 4.3: Bounding box annotations propagation

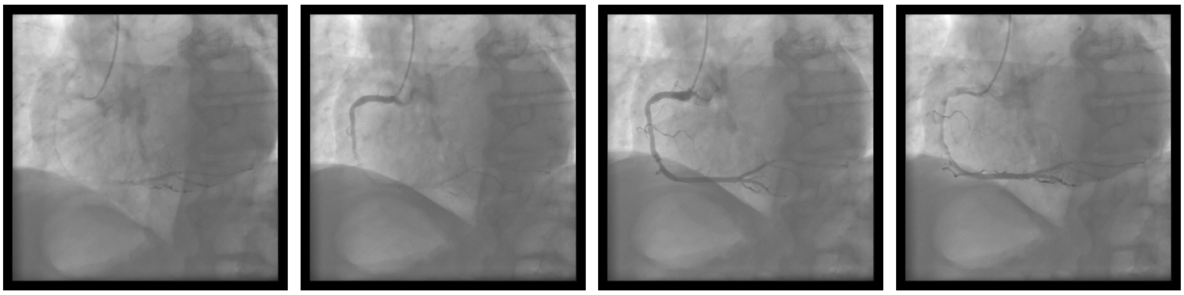
To alleviate the labour of having to annotate all the bounding boxes in every single frame manually, the object tracking algorithm *Discriminative Correlation Filter Tracker with Channel and Spatial Reliability* [37] was implemented using the OpenCV image processing library. By defining our initial template of the object, i.e. our initial bounding box, a set of correlation filters (Histogram of Gradients and Colour filters) is initially trained and used to estimate and update the new object position at the next frame. The correlation filters response is computed and weighted by their channel reliability. All the hyperparameters and implementation details were followed according to the original paper.

The propagation is made through a three-step process (see Figure 4.3): (1) the initial bounding box is obtained by transforming the initial annotation of the reference frame to a bounding box; (2) using the tracking algorithm the initial bounding box is propagated to the forward part of the sequence; (3) The same reference bounding box is propagated to the backwards part of the sequence. Using this procedure the reference bounding box is propagated to all frames of the sequence.

Even though the tracking algorithm worked as intended, in some frames the bounding box was not properly placed due to the occlusion of the stenosis, rapid shift of the lesion, frame rate and situational sudden movements. These misplaced bounding boxes need to be manually corrected such that they have a perfect fit in the region of interest.

4.2.2 Optimal interval definition

Not every frame of the sequence is suitable and can't be considered an optimal frame. Only a specific interval of the sequence, when the radio-opaque contrast is fully visible and introduced, filling all of the coronary arteries with the stenosis best seen, can be considered optimal.



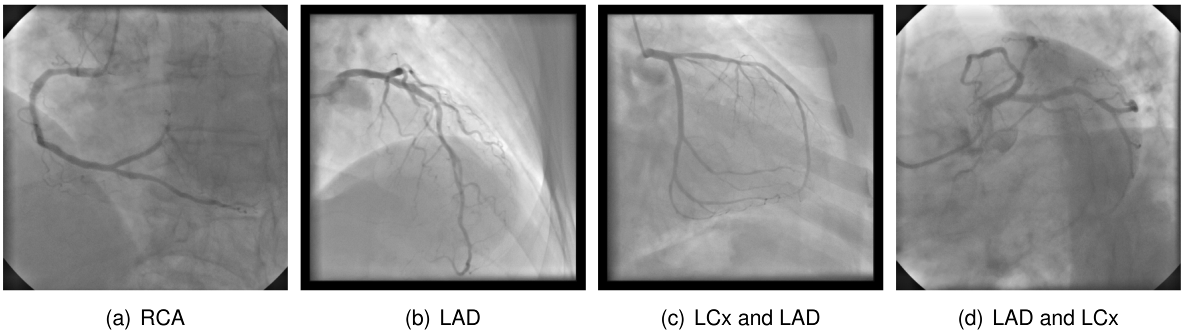
(a) No contrast is present (b) Contrast being introduced (c) Contrast fully introduced (d) Contrast is vanishing

Figure 4.4: Different sequence intervals of the radio opaque contrast placement

Four different intervals (see Figure 4.4) need to be manually labelled as: (a) no radio-opaque contrast is present, and no coronary arteries and stenosis are visible; (b) the radio-opaque contrast is being introduced but has not yet filled all the coronary arteries and stenosis; (c) the radio-opaque contrast has been fully introduced, and all the coronary arteries and stenosis are optimally viewed (d) the radio-opaque contrast is vanishing from the coronary arteries and is not an optimal viewing case.

4.2.3 Angle view selection

From the present metadata in each DICOM sequence file, it is possible to access the X-ray C-arm unit's angles in which the sequence was taken. A correlation between the machine's angles and the coronary arteries' common viewing angles (see Table 3.1) is feasible. However, in many cases, they do not directly correspond to each other since several angle variations occur during the procedure, mainly due to distinct patient morphology.


(a) RCA

(b) LAD

(c) LCx and LAD

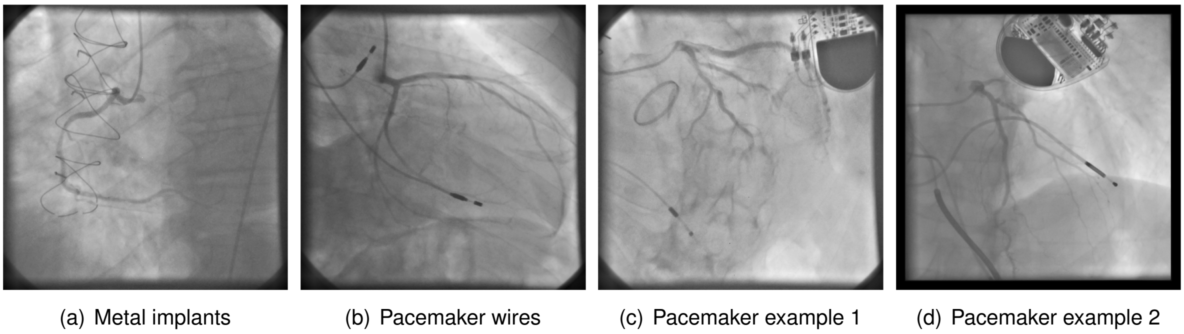
(d) LAD and LCx

Figure 4.5: Most common viewing angles from left to right, with the corresponding coronary arteries

To address this, sequences are firstly grouped by their respective angles (obtained by the DICOM metadata) and then manually filtered to the most common viewing angle names (see Figure 4.5). For simplicity reasons, sequences containing the only RCA and LAD are denoted RCA view and LCA view.

4.2.4 Metal implants

A minority of patients had pacemakers and/or metal implants. Being radio-opaque, they completely/partially obstructed the coronary arteries in some sequences.


(a) Metal implants

(b) Pacemaker wires

(c) Pacemaker example 1

(d) Pacemaker example 2

Figure 4.6: Examples of image artefacts present in the sequences

As these sequences (see Figure 4.6) only represent a small portion of the dataset, they were discarded as they would affect the optimal dataset quality.

4.2.5 Processed dataset

All of the sequences with implants were initially discarded and did not go through the bounding box propagation procedure nor frame interval selection. Additionally, some sequences also required further curation from the expert physician, as the annotations could not be programmatically retrieved.

The 1593 annotated sequences from 438 patients were processed with the described steps, resulting in 1294 optimal sequences (see Table 4.2) with 20819 optimal frames, where the bounding box was placed. From the optimal sequences, the most common angles (see Figure 4.5) were grouped.

			Frames						
Sequence Detail	Patients	Sequences	No Contrast	Introducing	Optimal	Vanishing	Stenosis Annotations	iFR below	iFR above
Total Sequences	438	1593	0	0	0	0	4234	554	1005
Discard	72	115	0	0	0	0	338	40	73
With Implants	82	184	0	0	0	0	472	81	96
Optimal	392	1294	11582	1294	20819	39266	3424	433	836
Optimal RCA	91	235	2249	235	3983	6323	309	25	210
Optimal LCA	126	155	1323	155	2474	5077	225	70	85
Optimal LCx/LAD	111	118	1155	118	1912	3869	159	53	65
Optimal LAD/LCx	90	92	865	92	1590	2616	105	43	49
No Lesion RCA	48	54	465	54	748	1450	0	0	0
No Lesion LCA	17	18	153	16	190	538	0	0	0

Table 4.2: Processed dataset, with all optimal sequences, frame intervals, stenosis annotations, and iFR values count.

From the raw dataset, additional sequences without any stenosis were also annotated with their optimal intervals to enhance the processed dataset.

Due to the limited sequences of *Optimal LCx/LAD* and *Optimal LAD/LCx*, only *Optimal RCA* and *Optimal LCA* viewing angle sequences and the respective *No Lesion* sequences contributed to this study.

Chapter 5

Implementation

5.1 Concepts

5.1.1 Neural Networks

The human brain has been a source of inspiration in the world of computer science and vision problems. The brain has a large number of neurons ($\sim 10^{11}$), which, although being slow (~ 1ms) comparatively to today's electronic devices, are able to seemingly achieve very complex and difficult tasks with ease in real-time.

One of the first attempts to model the human brain was by McCulloch and Pitts [38] in the early 1940s. By studying the neuron's biological anatomy, he tried to understand how the brain could perceive highly complex patterns by using many neurons connected.

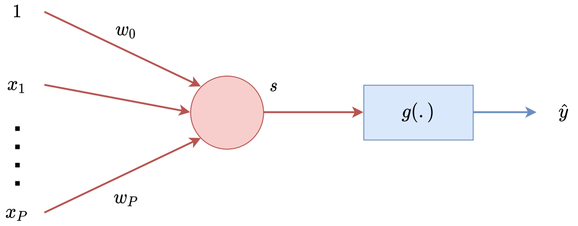


Figure 5.1: Simplified model of the neuron

The proposed artificial neuron model (MP neuron) (see Figure 5.1) has a linear part

$$s = \begin{bmatrix} 1 & \mathbf{x}^T \end{bmatrix} \mathbf{w} = \tilde{\mathbf{x}}^T \mathbf{w} \tag{5.1}$$

which represents the weighted sum of the inputs, with $\mathbf{w} = [w_0, \ldots, w_P]^T$ being the weight vector. It is followed by non-linearity

$$\hat{y} = g(s) = \begin{cases} 1 & \text{if} \quad s \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

where $g : \mathbb{R} \to \mathbb{R}$ is known as the activation function. In the 1950's Rosenblatt [39] proposed an improved version, the perceptron together with an iterative algorithm to train the weights of the artificial neuron: (1) start with a training set $\mathcal{T} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$ with $x^k \in \mathbb{R}^P$ and $y^k \in 0, 1$; (2) randomly initialize the weights $w_i(0), i = 0 \ldots, P$; (3) present a new training sample $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ to the model and compute the output $\hat{\mathbf{y}}(t) = g(\tilde{\mathbf{x}}^T(t)\mathbf{w}(t-1))$; (4) update the weights according to

$$w_i(t) = w_i(t-1) + \eta x_i(t)\epsilon(t), \quad \epsilon(t) = \mathbf{y}(t) - \hat{\mathbf{y}}(t) \tag{5.3}$$

where $\mathbf{y}$ is the desired outcome of the input $\mathbf{x}(t)$ and η is the learning rate that models the rate of convergence. Then repeat steps 3 and 4 until convergence.

The algorithm was proven to solve any binary problem, as long the training set could be separated by a hyperplane in feature space. Otherwise, and as well, with regression problems that are not binary, the algorithm would not converge.

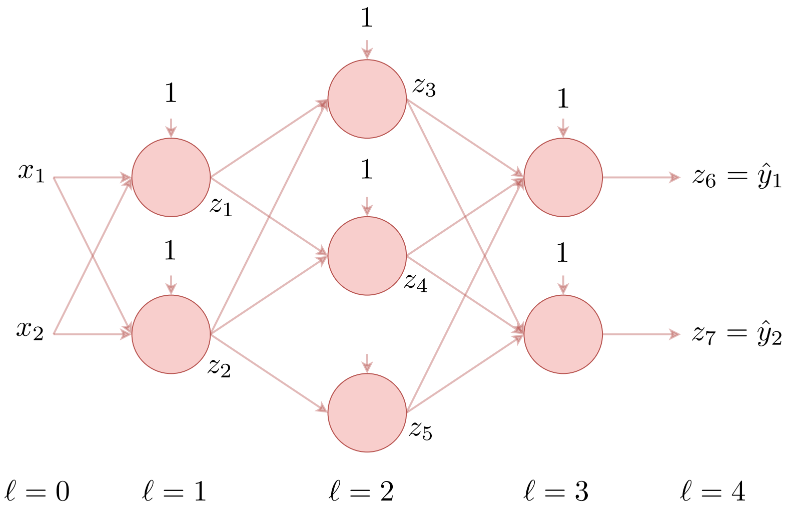


Figure 5.2: Simplified multilayer perceptron architecture

To solve the perceptron's limitations, multilayer perceptron (MLP) architectures (see Figure 5.2) with continuous and differentiable activation functions were developed.

Each unit s_i of layer L_ℓ is connected to a unit s_j of the next layer $L_{\ell+1}$ through a weight w_{ij} such that for each unit $j \in L_{\ell+1}$

$$\begin{gathered} s_j = w_{0j} + \textstyle\sum_{i \in L_\ell} w_{ij} z_i \\ z_j = g(s_j) \end{gathered} \tag{5.4}$$

The network can thus be seen as a non-linear function

$$\hat{y} = f(\mathbf{x}, \mathbf{w}) \tag{5.5}$$

that maps the input space to the output space controlled by a set of weights $\mathbf{w}$. The optimal weights are then estimated using a training set $\mathcal{T}$, through loss optimization

$$\mathcal{L} = \frac{1}{N} \sum_{k=1}^{N} \mathcal{L}_k(\mathbf{y}^k, \hat{\mathbf{y}}^k) \tag{5.6}$$

where $\hat{\mathbf{y}}^k$ is the network output for a given input $\mathbf{x}^k$ and the loss function $\mathcal{L}$ quantifies the deviation between the output and the ground truth $\mathbf{y}^k$.

The minimization of $\mathcal{L}$ is commonly achieved with the gradient descent algorithm

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t), \quad \Delta w_{ij}(t) = -\eta \left. \frac{\partial \mathcal{L}}{\partial w_{ij}} \right|_{w(t)} \tag{5.7}$$

If in batch mode, the gradient vector $\Delta \mathbf{w}$ at time step t includes all the training patterns of $\mathcal{T}$. If only a portion of the training patterns is used, it is called mini-batch mode. If only one pattern is used at a time, it is called online mode, a situation also known as stochastic gradient descent (SGD).

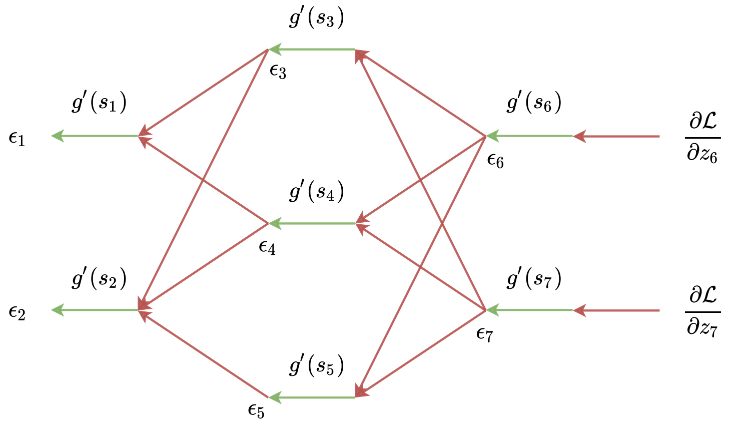


Figure 5.3: Backpropagation network

To obtain each weight, w_{ij}, the derivative of the loss function $\mathcal{L}$ with respect to each weight is necessary [40, 41]. This derivative is calculated for all network weights using the chain rule of differentiation (see Figure 5.3).

5.1.2 Convolutional Neural Networks

Convolutional neural networks (CNN) have been the standard for image analysis in many fields, such as image classification, recognition, and object detection. In contrast to MLP, which was inspired by the general connections between neurons of the brain, CNN's were inspired by how the visual cortex and its receptive field respond to specific image patterns.

An image is a matrix with dimensions *Width* x *Height* x *Channels* with values in a given range, usually from 0 to 255. CNN's are then able to capture those features by convolving them with a set of learned kernels w (filters) given by

$$v_{\ell n}^{xy} = \sum_{m} \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} w_{\ell mn}^{pq} v_{(\ell-1)m}^{(x+p)(y+q)} \tag{5.8}$$

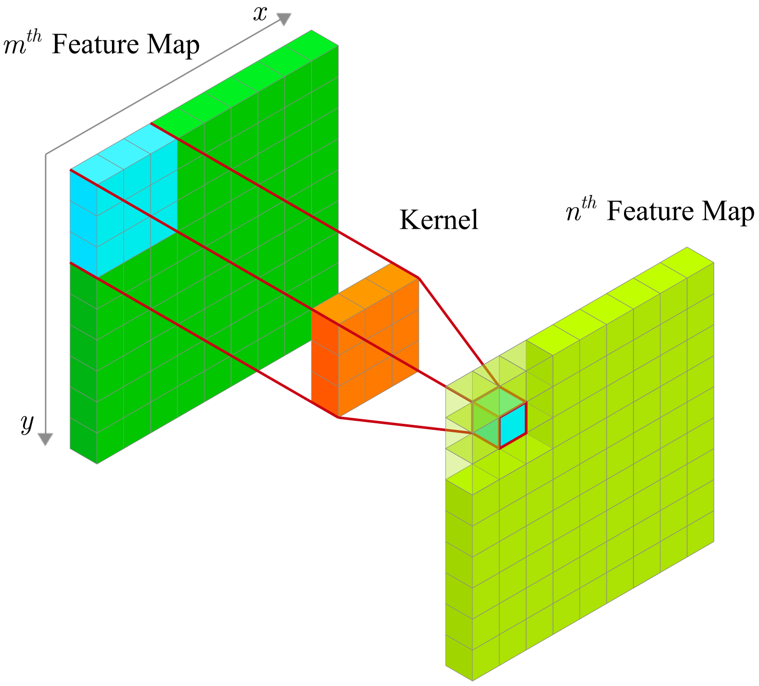


Figure 5.4: Convolution operation illustration

where P and Q are the dimensions of the kernel, and $w^{pq}_{\ell mn}$ is the value of the kernel at position (p, q) connected to the m^{th} feature map. These are then often followed by sub-sampling. The result is a lower-dimensional and easier to process feature map, which still bears a strong relation to the original image features (see Figure 5.4).

The initial convolutional layers of a network, which have the most resolution, are responsible for capturing low-level features like edges and color. By stacking even more convolutions (Deep Convolutional Neural Networks), followed by sub-sampling, the architecture begins to capture even more high-level features like curves, texture, and patterns [42]. It is then possible to connect the flattened feature map to a fully connected classification/regression layer and start the training process for the specific task at hand.

5.1.3 Transfer Learning

One problem with neural networks is that they will overfit the training set when there isn't enough data to train them due to the network's large number of weight parameters.

Transfer learning is a technique to overcome the small dataset limitation, where the network is trained in two steps. First, there's a pre-training stage where the network is trained under a broad and wide domain dataset with various classes/labels/categories, like ImageNet [43], which has over 20 thousand categories and over 14 million images. The fine-tuning stage is where the network is trained once more, but this time with the smaller dataset. The pre-trained network targets the specific target task of interest, which may have fewer category examples than the first step's broader dataset. This step is commonly performed by replacing the last layer of the network (classification layer) and replacing it with a fully connected randomized weighted layer with the number of categories require. The pre-training

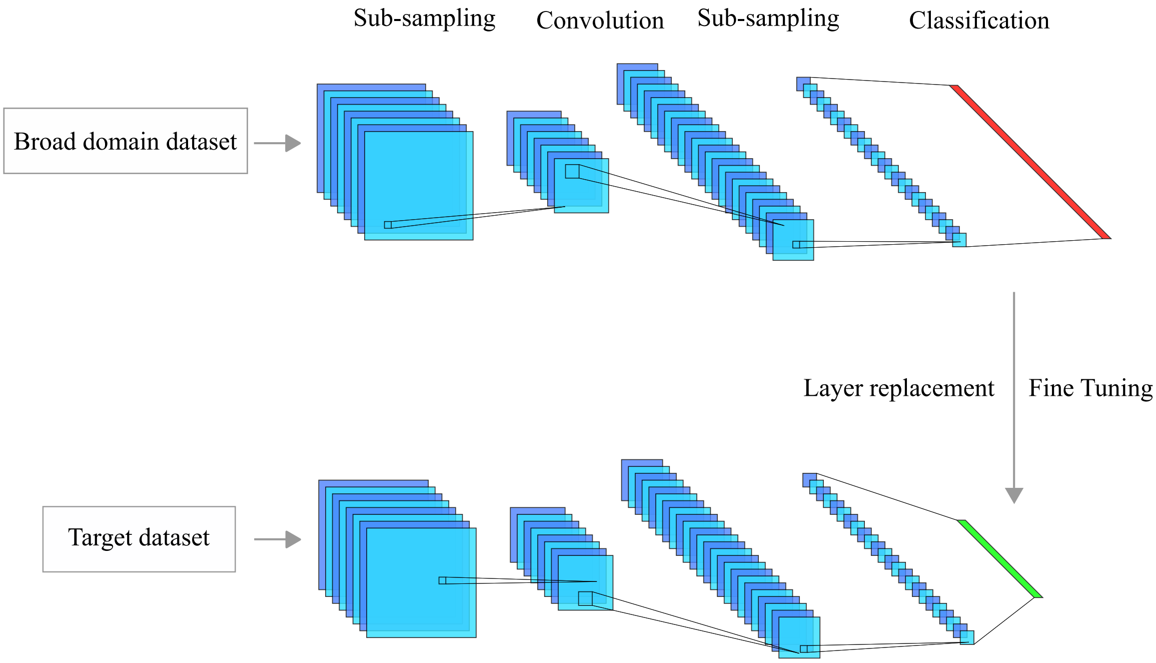


Figure 5.5: Transfer Learning and fine tuning example by replacing the last classification layer

step helps the network learn general features that can then be reapplied on the specific target task (see Figure 5.5).

One of the benefits of transfer learning is that researchers often publish their state-of-the-art architectures with the pre-trained weights available. It's then possible to take advantage of the huge computing time required in pre-training and give more focus to the fine-tuning stage where changes in hyperparameters may be necessary. Although training a network from scratch would be considered the standard procedure in a training pipeline, it has been shown that in small datasets and medical tasks specifically, training from scratch may have a similar performance to transfer learning methods [44].

5.1.4 Cross-Validation

The standard approach when training and evaluating a model is to randomly split our training set $\mathcal{T}$ into three subsets: (1) one for training the model; (2) one for validation during training, and (3) one to test on unseen data when the training process is finished. But when working with small datasets and limited data, the split operation is highly susceptible to selection bias. For example, if our test subset only contains easy examples, the final evaluation will deliver high results, but if the test subset only contained hard examples, the opposite result would occur.

In k-fold cross-validation, every observation is ensured to appear in both training and test subsets. The training set $\mathcal{T}$ is randomly split into k exclusive folds $\{\mathcal{T}_1, \mathcal{T}_2, \dots \mathcal{T}_k\}$ with equal size, where the model is trained and tested k times. In each iteration $t \in \{1, 2, \dots, k\}$ the model is trained on $\mathcal{T} \setminus \mathcal{T}_t$ and evaluated on $\mathcal{T}_t$. In stratified cross-validation, the folds are stratified, ensuring approximately the same proportions of classes/labels as in the original dataset $\mathcal{T}$

Performance of cross-validation is then obtained by computing the mean and variance from the

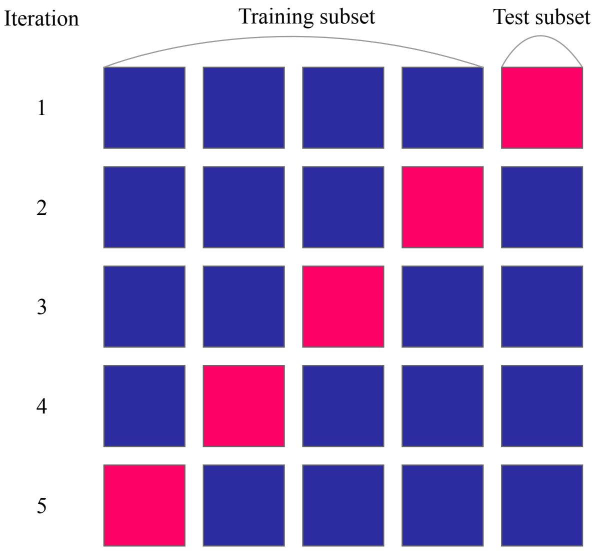


Figure 5.6: 5-Fold Cross-validation procedure example.

evaluation on the test subset at each iteration (see Figure 5.6). A good performance in cross-validation shows a high value in the mean and a low variance, indicating that the model could generalize well to all unseen data [45].

5.2 Angle View Classification

The initial phases of detecting and assessing a stenosis require that a coronary viewing angle must first be filtered and selected. Since every sequence was previously labeled with the respective viewing angle, this task is address as a classification problem, as one frame can only belong to a specific viewing angle. The ResNet-50 was chosen as the architecture for this task.

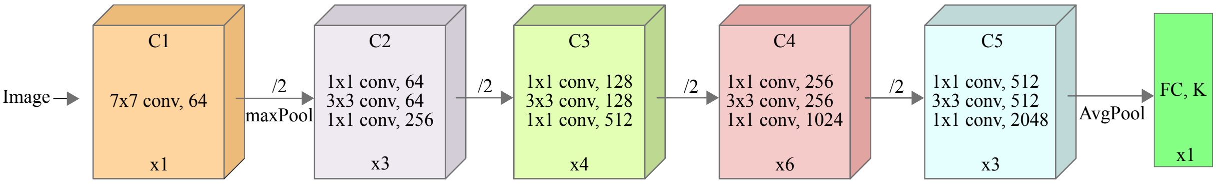


Figure 5.7: ResNet-50 simplified architecture. Bottom of each block denotes the repeated set of layers (Kernel size, operation, number of channels) in each convolution block (C). After each block ($C1$ to $C5$) spatial dimension is reduced. The final layer is a fully connected layer FC with K units. Shortcut connections occur every two layers but are omitted for readability.

The ResNet is characterized by shortcut connections that skip one or more layers, a technique that improves the flow of relevant feature information into deeper layers [46].

The shortcut connections allow the summation of the previous layer outputs to the outputs of the stacked layers, contributing to a better feature propagation across the deep network and can be described by

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x} \tag{5.9}$$

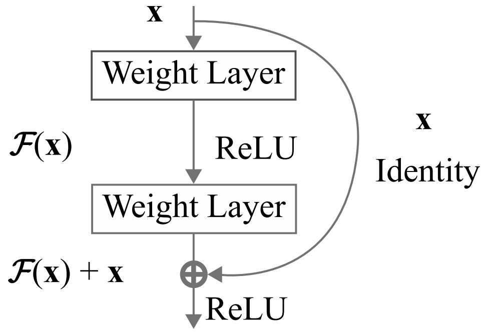

x
Weight Layer
$\mathcal{F}(\mathbf{x})$
ReLU
Weight Layer
x
Identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
ReLU

Figure 5.8: Residual block with shortcut connection used in the ResNet layers.

where $\mathcal{F}(\mathbf{x}, \{W_i\})$ is the residual mapping to be learned, eg. in Figure 5.8, $\mathcal{F} = W_2\text{ReLU}(W_1\mathbf{x})$. After each residual block, sub-sampling is applied directly from the convolutional layers using a stride of dimension 2. Since the dimensions of $\mathcal{F}$ and $\mathbf{x}$ must be equal, a linear projection W_s is made to match the correct dimensions between such that

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + W_s\mathbf{x} \tag{5.10}$$

only when there are spatial dimension changes and where W_s is the identity matrix. For each layer of the ResNet network, except the fully connected layer, the ReLU activation function [47] is used

$$\text{ReLU}(x) = x^+ = \max(0, x) \tag{5.11}$$

with batch normalization layers [48], This standardizes a layer's inputs for each mini-batch by maintaining the output close to zero with a standard deviation close to one.

In our application, the ResNet-50 was initially pre-trained on ImageNet with 224 by 224 images. For fine-tuning, the last layer was replaced by a fully connected layer with two output units, and random weights obtained through *Xavier* initialization method [49]. In this initialization the weights follow a uniform distribution within the bounds $\left[-\frac{\sqrt{6}}{\sqrt{n_{in}+n_{out}}}, \frac{\sqrt{6}}{\sqrt{n_{in}+n_{out}}}\right]$ with n_{in} and n_{out} being the in-going and out-going connections respectively, ending with softmax activation function

$$softmax(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}. \tag{5.12}$$

Softmax activation normalizes the units vector $\mathbf{z}$, with sum add up to 1, being interpreted as probabilities. The network was then trained under the categorical cross-entropy with loss defined as

$$\mathcal{L}_{angle}^{cls} = \text{CCE} = -\sum_{i=1}^{K} y_i \log(\hat{y}_i). \tag{5.13}$$

Additional training details

The ResNet-50 was trained and evaluated using 5-fold stratified cross-validation at sequence-level, for 30 epochs and with a batch size of 32. To improve convergence speed, stochastic gradient descent with the *Adam* [50], a first-order gradient-based optimization algorithm was performed, with weights w_{ij} being updated according to

$$w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\hat{m}_{w_{ij}}}{\sqrt{\hat{v}_{w_{ij}}} + \epsilon}, \tag{5.14}$$

$$\hat{m}_{w_{ij}} = \frac{m_{w_{ij}}(t+1)}{1-\beta_1}, \tag{5.15}$$

$$\hat{v}_{w_{ij}} = \frac{v_{w_{ij}}(t+1)}{1-\beta_2}, \tag{5.16}$$

$$m_{w_{ij}}(t+1) = \beta_1 m_{w_{ij}}(t) + (1-\beta_1)\frac{\partial \mathcal{L}}{\partial w_{ij}}, \tag{5.17}$$

$$v_{w_{ij}}(t+1) = \beta_2 v_{w_{ij}}(t) + (1-\beta_2)\left(\frac{\partial \mathcal{L}^{cls}_{angle}}{\partial w_{ij}}\right)^2 \tag{5.18}$$

m and v are initialized at zero and are the moving averages of the first and second moment of the gradients (mean and variance) with weight decaying factors $\beta_1 = 0.9$ and $\beta_2 = 0.99$. With $\hat{m}$ and $\hat{v}$ correcting the bias towards zero, from the initialization procedure. To prevent zero division, $\epsilon = 10^{-8}$ is set. The initial learning rate was initialized at $\eta = 10^{-5}$ and was reduced by a factor of 0.2 on *plateau*, or if the loss did not decrease after 5 epochs.

To reduce early stages of overfitting and large gradient updates to the network, due to the weight initialization of the last fully connected layer, a two-stage training workflow was assembled where: (1) for the first 15 epochs, the gradient updates on all layers of the network are frozen except for the C_5 block and fully connected layer, so the gradient updates do not become too large preventing overfitting in early steps; (2) for the next 15 epochs, the gradient updates of the entire model are restored allowing the model to converge in its entirety.

5.3 Stenosis Detection

The objective of this step is to detect and estimate the position of every visible stenosis in a given frame. Given the annotated bounding boxes for the stenosis in the optimal interval, it's possible to approximate this to an object detection/recognition problem where the stenosis is the object of interest to be detected.

Early state-of-the-art machine learning solutions for object detection were based on two-stage detector architectures. The first stage could be a faster R-CNN [51], which possesses a Region Proposal Network (RPN) that generates bounding boxes and classifies each one as having an object or not with a defined threshold. Bounding boxes classified as having an object pass to the second stage which estimate the coordinates of the bounding box by regressing its values and determines its class/category. These types of networks and its variants still remain very capable of obtaining considerable results in the COCO [52] (Common Objects in COntext) dataset, with 90 object classes, the most common and used dataset for object detection. One main issue relies on the network's complicated training process as it needs a four-step process to train the RPN and the classifier. Additionally, it is challenging to overcome the network's high inference time due to the high number of parameters.

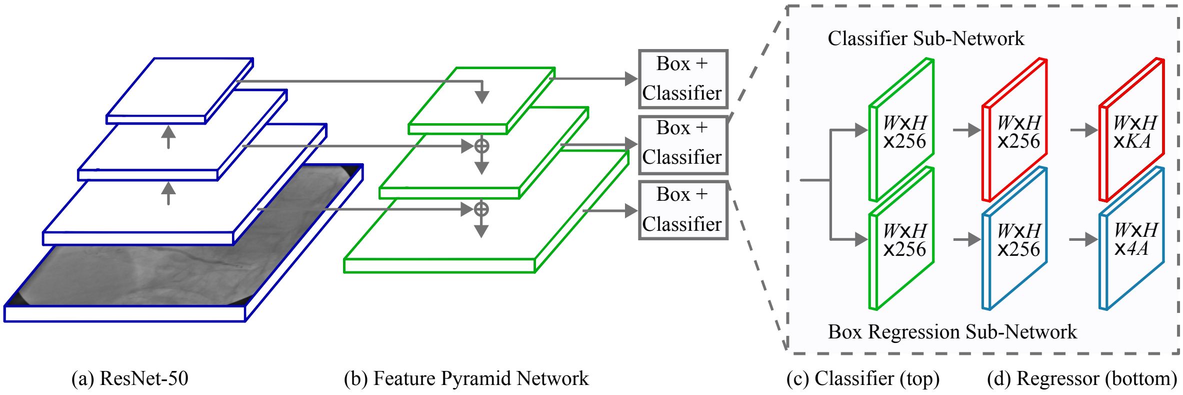


Figure 5.9: RetinaNet architecture with (a) ResNet-50 and (b) Feature Pyramid Network as feature extractor to (c) classify the lesion existence probability and (d) regress the bounding box coordinates.

We decided to adopt a state-of-the-art architecture for object detection, the RetinaNet [53], which was first pre-trained on the COCO dataset achieving a 36.9 mean average precision (mAP) score, to fine-tune it for the stenosis detection task. The RetinaNet's architecture is based on the unified single-shot detector architecture, composed of a backbone and two additional sub-networks (see Figure 5.9). The backbone is responsible for computing and extracting relevant features of the image input. The two sub-networks are responsible for correctly classifying a bounding box and regressing the estimated coordinates.

Feature Extraction

In the backbone, the ResNet-50 is first applied for deep image feature extraction.

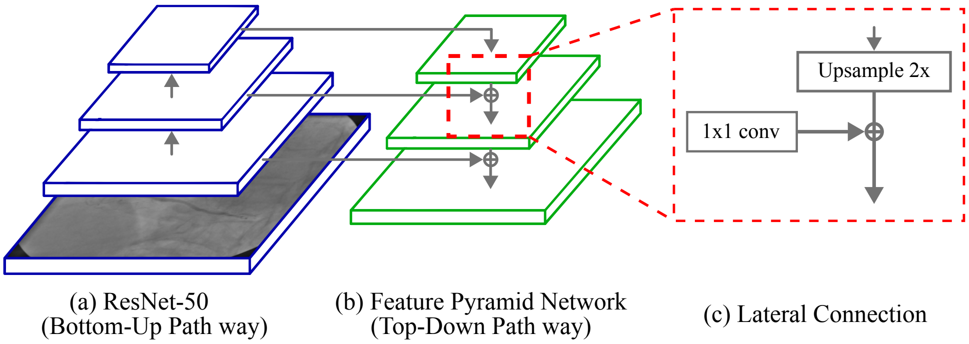

Upsample 2x
1x1 conv
(a) ResNet-50 (Bottom-Up Path way)
(b) Feature Pyramid Network (Top-Down Path way)
(c) Lateral Connection

Figure 5.10: Backbone feature extraction is composed of a (a) bottom-up pathway with the ResNet architecture and by a (b) top-down pathway which is a result of the illustrated (c) lateral connection block.

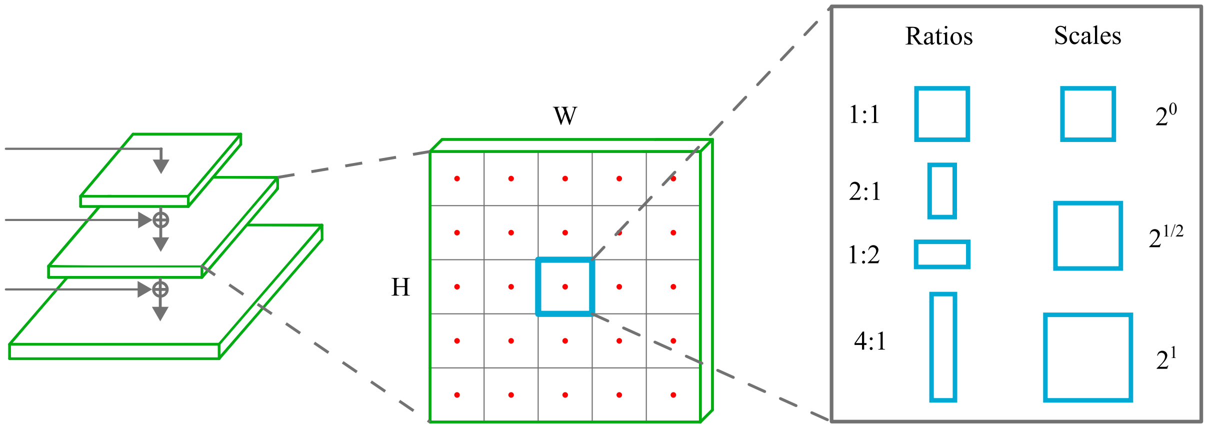
On top of the ResNet-50, Feature Pyramid Networks (FPN) [54] is also adopted as a top-down pathway to complete the RetinaNet architecture's backbone (see Figure 5.10). The bottom-up pathway is the feed-forward computation of the ResNet-50, extracting feature maps at distinct steps $\{C_1, C_2, C_3, C_4, C_5\}$

of the network. The top-down pathway then produces higher resolution features by up-sampling spatially crude but semantically stronger feature maps, which are then enhanced with features from the bottom-up pathway by means of lateral connections. Each lateral connection merges feature maps with the same spatial size from the bottom-up pathway. From the crude-resolution feature maps, they are up-sampled by a factor of two. This up-sampled feature map is merged with the corresponding bottom-up map, but since the up-sampled map differs in the number of channels, a one by one convolution is performed to match the correct channel depth.

The process is started with the one by one convolution on the C_5 block, producing the crude feature maps with $C = 256$ channels. A 3 by 3 convolution is applied to reduce the up-sampling process's effects for each merged feature map. The final set of feature maps is called $\{P_3, P_4, P_5, P_6, P_7\}$, and is computed from the last feature maps corresponding to the ResNet-50 blocks $\{C_3, C_4, C_5\}$ using the lateral connections. P_6 is obtained by applying a 3 by 3 convolution with stride 2 on the resulting feature maps of C_5 block, and P_7 is obtained by applying ReLU followed by 3 by 3 convolution with stride 2 on P_6. These additional feature maps improve larger stenosis detection. $P1$ and P_2 are not included in the feature set due to the large spatial dimension, which would affect memory and increase computation requirements.

Anchor Generation

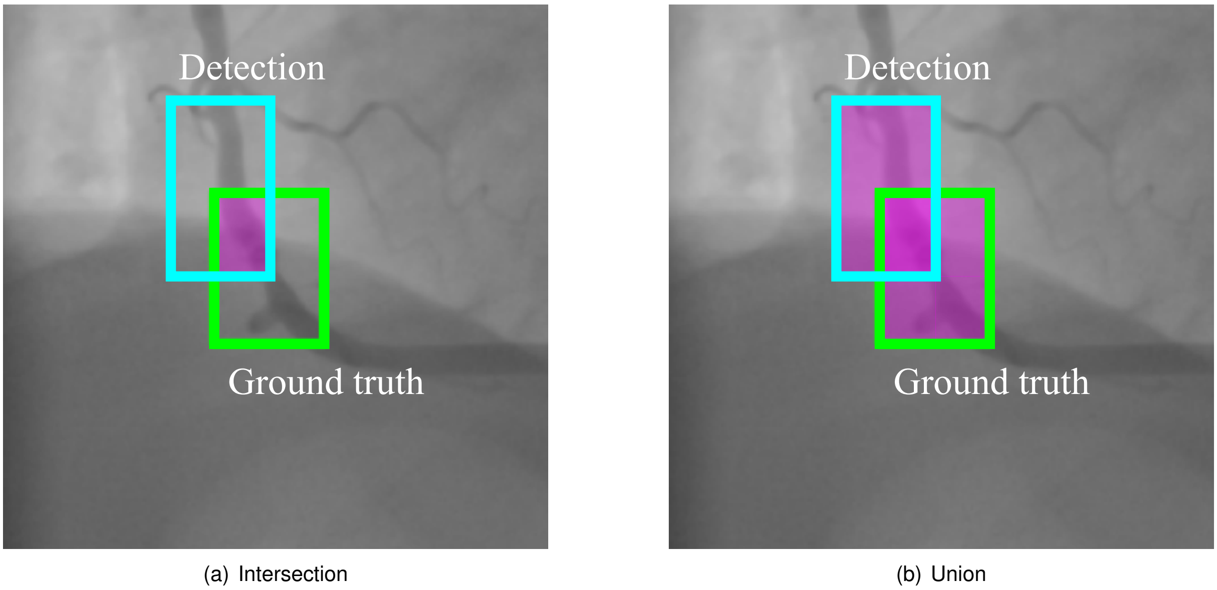
For bounding box classification and regression, translation-invariant anchor boxes (pre-defined bounding boxes) are generated in each pixel of the feature map for every pyramid level, P_3 through P_7, having areas of 32^2 to 512^2 respectively.



(a) Feature Pyramid Network (Top-Down Path way)
(b) Anchor boxes in Feature Map
(c) Anchor aspect ratios and scales

Figure 5.11: The process of generating anchor boxes from (a) Feature Pyramid Network. For each pixel in the (b) feature map with dimensions $H \times W$ (c) distinct aspect ratios and scales are created and assigned to their respective targets.

To improve bounding box coverage, several aspect ratios $\{1{:}1,\ 1{:}2,\ 2{:}1,\ 4{:}1\}$ and scales $\{2^0, 2^{1/2}, 2^1\}$ are created for each anchor box (see Figure 5.11), resulting in $A = 12$ anchors per feature map pixel in each pyramid level. Each anchor is assigned 4 length vectors of box regression values and a one-hot vector with length $K = 1$ of classification targets, with only one target to detect, i.e., the stenosis.



(a) Intersection

(b) Union

Figure 5.12: Illustration of the (a) intersection between the areas of the ground truth bounding box (green) and the generated bounding box (cyan) (valid for anchors and further detections), and (b) the union of areas.

The assignment of ground-truth bounding boxes to each generated anchor is made by setting a matching Intersection-over-Union (IoU)

$$\text{IoU} = \frac{\text{Area of intersection}}{\text{Area of union}} \tag{5.19}$$

threshold $m_{th}^{IoU} = 0.2$ between the anchor and the ground truth (see Figure 5.12). If the IoU is below the threshold, it is considered background. Otherwise, it matches the stenosis target. This assignment will then be compared with the respective classification results and further regressed.

Classification sub-network

To identify the presence of a stenosis and regress the bounding box coordinates, two sub-networks are created and attached to each pyramid level (P_3 through P_7), sharing weight parameters across all levels.

For the classification sub-network, the resulting pyramid feature map with dimensions $W \times H$ and depth C is convolved with 3 by 3 kernels four times each, followed by ReLU activations. Each pixel of the feature map is assigned $A = 12$ anchors and $K = 1$ targets. The last layer ends in $W \times H \times K \times A$ units with sigmoid activation functions

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{5.20}$$

for classification. For consistency across all pyramid levels where feature map dimensions and channel depth differ, 256 channels are defined as the depth input for the following convolutions, performed by reshaping the feature maps to the desired depth.

A class imbalance occurs between background and stenosis during the training process due to the number of anchors being generated at each pixel of the feature map. Easy background examples (smaller error), with estimated probability $p \ll 0.5$, weight heavily on the gradient updates, while hard and well-classified stenosis examples (less frequent) don't contribute as much when using binary cross-entropy loss

$$\text{CE}(p_t) = \log(p_t), \tag{5.21}$$

where

$$p_t = \begin{cases} p & \text{, if} \quad y = 1 \\ 1 - p & \text{, otherwise.} \end{cases} \tag{5.22}$$

Where $p \in [0, 1]$ denotes the class probability and y the respective class. To address the occurring class imbalance between background and stenosis, the α-balanced Focal Loss function

$$\text{FL} = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \tag{5.23}$$

with an additional focusing parameter $\gamma \geq 0$ and a class balancing hyperparameter $\alpha \in [0, 1]$

$$\alpha_t \in [0, 1] = \begin{cases} \alpha & \text{, if} \quad y = 1 \\ 1 - \alpha & \text{, otherwise} \end{cases} \tag{5.24}$$

is applied to reduce the loss contribution of well-classified detections. With $\gamma = 0$, the loss would be equal to standard cross-entropy. By defining the modulating factor $(1 - p_t)^{\gamma}$, miss-classified detections with a high confidence score p (low p_t), the modulating factor tends to one, and the loss is similar to cross-entropy. With well-classified detections and a high confidence score p (high p_t) the modulating factor tends to zero, adding less contribution to the loss. By setting the parameter $\gamma \geq 0$, the down-weighted loss contribution from easy and well-classified detections can thus be regulated. The α parameter offsets the background/stenosis imbalance presence.

For each 512 by 512 pixel, ≈45 thousand anchor boxes are generated. The loss is calculated for all bounding boxes and is normalized by the number of previously assigned anchors to ground truth stenosis N_g

$$\mathcal{L}_{det}^{cls} = -\frac{1}{N_g}\left[y_i \log(p_i)^{\gamma}\alpha + (1 - y_i)\log(1 - p_i)p_i^{\gamma}(1 - \alpha)\right], \tag{5.25}$$

giving the classification loss between the estimated stenosis probability p_t and the ground truth label class y_i. $\alpha = 0.25$ and $\gamma = 2$ were defined by experimentation for this stenosis detection task.

Bounding box regression sub-network

For the bounding box coordinates regression, a second architecture is attached to each pyramid level to estimate the offset between the predicted and the ground-truth coordinates. The architecture is the same as the classification sub-network except for the last layer, which ends in $W \times H \times 4 \times A$ units with linear activations.

The sub-networks objective is to estimate the relative offset between the predicted anchor $\hat{A}$ and the matched ground-truth bounding box G. First, a parameterized regression target T is calculated [55] for each matched pair $(\hat{A}, G)$ as

$$t_x = \left(G_x - \hat{A}_x\right) / \hat{A}_w, \tag{5.26}$$

$$t_y = \left(G_y - \hat{A}_y\right) / \hat{A}_y, \tag{5.27}$$

$$t_w = \log\left(G_w / \hat{A}_w\right), \tag{5.28}$$

$$t_h = \log\left(G_h / \hat{A}_h\right), \tag{5.29}$$

where (t_x, t_y) denotes a center scaling invariant translation and (t_w, t_h) represent logarithmic space translations of the width and height of the estimated anchor $\hat{A}$. The network is then trained to estimate this parameterized coordinates offsets T under the smooth L_1 loss function [56]

$$\text{smooth}_{L_1} = \begin{cases} 0.5x^2 & \text{, if } \quad |x| < 1 \\ |x| - 0.5 & \text{, otherwise,} \end{cases} \tag{5.30}$$

which combines L_1 loss, having a constant gradient when x is large, with L_2 loss, adding linear-gradient updates. This makes the model more robust to outlier detections giving a total regression loss

$$\mathcal{L}_{det}^{reg} = \sum_{j \in \{x,y,w,h\}} \text{smooth}_{L1}\left(T_j - \hat{T}_j\right) \tag{5.31}$$

between the estimated $\hat{T}$ and ground-truth parameterized offset coordinates T.

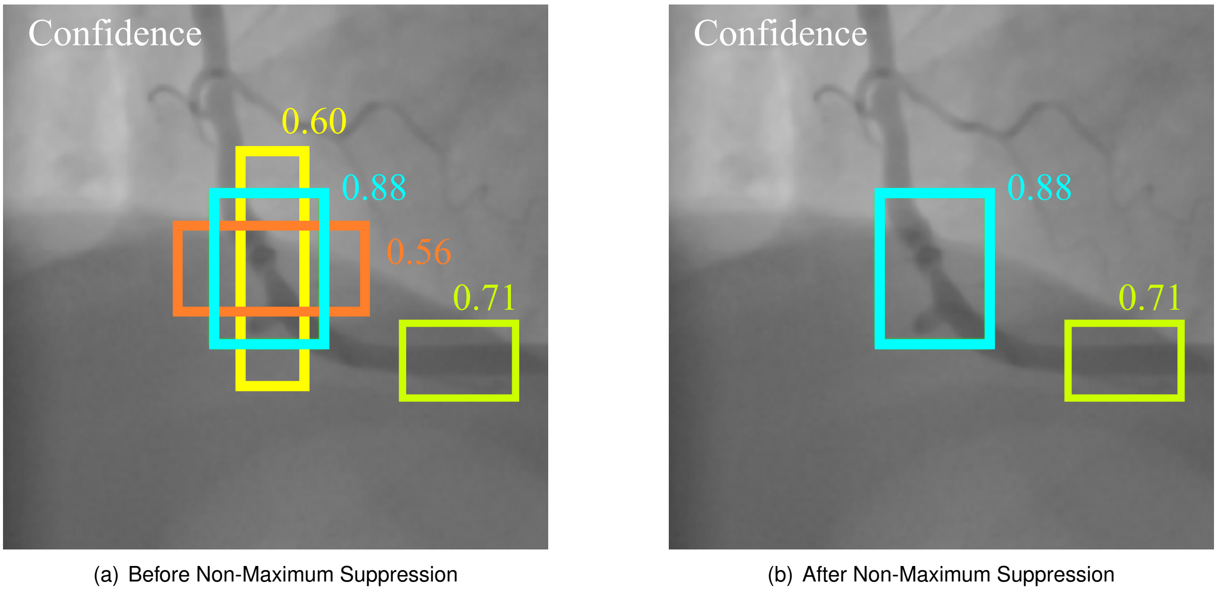
Non-Maximum Suppression (NMS)

Many overlapping situations can occur with different confidence scores due to the amount of generated candidate bounding boxes at inference. Only the greatest confidence is desired after the classification and regression estimation. For example, two bounding boxes can have the same center coordinates with different widths and heights, but one has the greatest confidence for stenosis detection. These overlapping situations can negatively impact the loss and performance of the model.

To deal with this specific issue, the non-maximum suppression algorithm is applied only in inference to filter overlapping occurrences of the set of candidate bounding boxes B with different confidence scores.

The iterative process of suppression can be defined as: (1) from the set of candidate bounding boxes B, the box $b \in B$ with the highest confidence and above the threshold NMS_{th}^{cls} score, is moved to the final proposed set D and is defined as b_h; (2) with b_h, the IoU is computed against every $b \in B$, and if the IoU is greater than a predefined threshold NMS_{th}^{IoU}, the box b is excluded; (3) the process is repeated starting from (1) until no bounding boxes b are left in B, thus leaving the final set of stenosis bounding boxes D.

The confidence threshold was set to $NMS_{th}^{cls} = 0$, allowing the suppression of every overlapping case. The IoU threshold was defined as $NMS_{th}^{IoU} = 0.5$.



(a) Before Non-Maximum Suppression

(b) After Non-Maximum Suppression

Figure 5.13: With a set of (a) candidate bounding boxes with difference confidence scores, the non-maximum suppression algorithm is applied resulting in the (b) final set of bounding boxes.

Only the top 100 bounding boxes from the classification sub-network are selected to go through the NMS post-processing algorithm to improve inference speed (see Figure 5.13).

Additional training details

The model was trained and evaluated under 5-fold stratified cross-validation at patient-level, for 3500 steps ($\approx$ 20 epochs) with a batch size of 32, under stochastic gradient descent with an initial learning rate of $\eta = 8.10^{-4}$ and a momentum term $\gamma = 0.9$. Since high weight values increase chances of overfitting, L_2 regularization was implemented, applying a penalty for the networks weight values

$$\mathcal{L}^2_{det} = \lambda \sum_{i=1}^{W} w_i^2. \tag{5.32}$$

with a weight factor $\lambda = 4.10^{-4}$. The total cost function for the RetinaNet is then a combination of the classification, regression, and regularization loss resulting in

$$\mathcal{L}_{det} = \mathcal{L}^{cls}_{det} + \mathcal{L}^{reg}_{det} + \mathcal{L}^2_{det}. \tag{5.33}$$

The weights being updated according to

$$w_{ij}(t+1) = \gamma w_{ij}(t) + \nabla w_{ij}(t), \quad \nabla w_{ij}(t) = -\eta \left. \frac{\partial \mathcal{L}_{det}}{\partial w_{ij}} \right|_{w(t)}. \tag{5.34}$$

The learning rate was then lowered with a factor of 0.2 on intervals of 1250 steps.

Data augmentation techniques were also implemented during training, enhancing our dataset's quality and size by generating additional modified versions of the original frames to reduce overfitting and aid the model at generalization. Early experiments showed that this technique must be implemented with caution. For example, if a horizontal/vertical flip transformation is applied, a scenario that originally does not exist, the model will learn and adapt to that transformation, possibly harming its performance in regards to the validation set. As such, only brightness and contrast variations were made to the original frame f_o, generating a new augmented frame

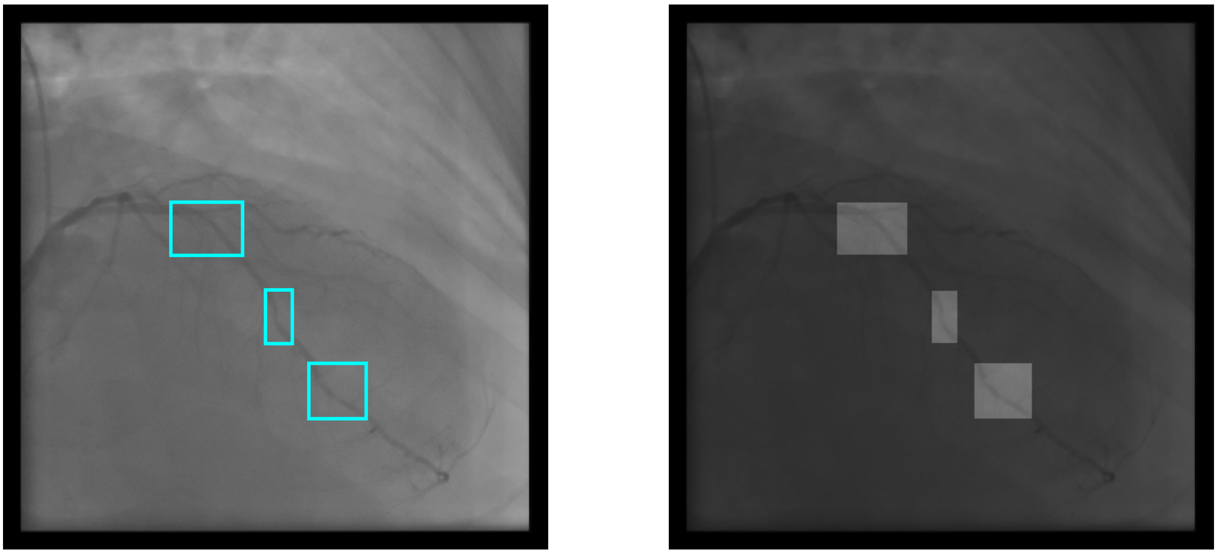
$$g_a = \lambda_c f_o + \lambda_b, \tag{5.35}$$

during training, with varying contrast $\lambda_c \in [0.2, 0.5]$ and brightness $\lambda_b \in [0, 10]$.

5.4 Stenosis Severity Regression

The final desired outcome is to determine the quantitative value of iFR. This is not a straightforward classification/regression problem. Revisiting chapter 3 and chapter 4, at any given frame, more than one coronary artery can be seen. Additionally, for any given coronary artery, multiple stenosis occurrences may exist. Thus it's not possible to evaluate the bounding boxes independently as iFR represents the stenosis's contribution as a whole.

Attention mechanisms applied in natural language processing fields, image recognition, and image captioning, in a broad manner, use an attention score a to learn which areas of the image contribute the most to the desired outcome. This approach can further be developed into *Hard* and *Soft Attention*. *Hard Attention* chooses the most contributing location by taking the highest attention score a, corresponding to a memory state. On the other hand, *Soft Attention* can take advantage of all the image locations by making a learnable weighted sum of the attention scores, resulting in all locations adding value to the contribution [57–59].



(a) Original set of bounding boxes

(b) Outer region regulation with $\beta = 0.5$

Figure 5.14: Proposed approach of contrast regulation from the (a) the original image with the bounding boxes follows (b) the modification of outer regions RGB values.

The already determined bounding boxes, which are the main contributing regions to the quantitative iFR severity assessment, indicate where to look. The problem relies on determining how much the bounding box's outer regions contribute to the outcome, since other parts of the coronary artery still add value to the contribution, even if small.

The approach used is based on *Hard* and *Soft Attention* principles as well as on image segmentation techniques where the region of interest is converted to binary values and commonly takes polygon shapes. An intuitive approach defined β-method was used, such that the main contributing regions, i.e., the regions of interest, remain invariant, and a variation is made to the outer region pixels corresponding to regulation in contrast levels (see Figure 5.14). Outside the bounding boxes, each pixel is multiplied by β

$$g(x^*, y^*) = \beta f(x^*, y^*) \tag{5.36}$$

with $f(.)$ representing the three-channel RGB original image, where x^* and y^* are the pixel coordinates outside the bounding boxes.

Regression Network

To quantify the iFR assessment value, the InceptionV3 [22] was chosen.

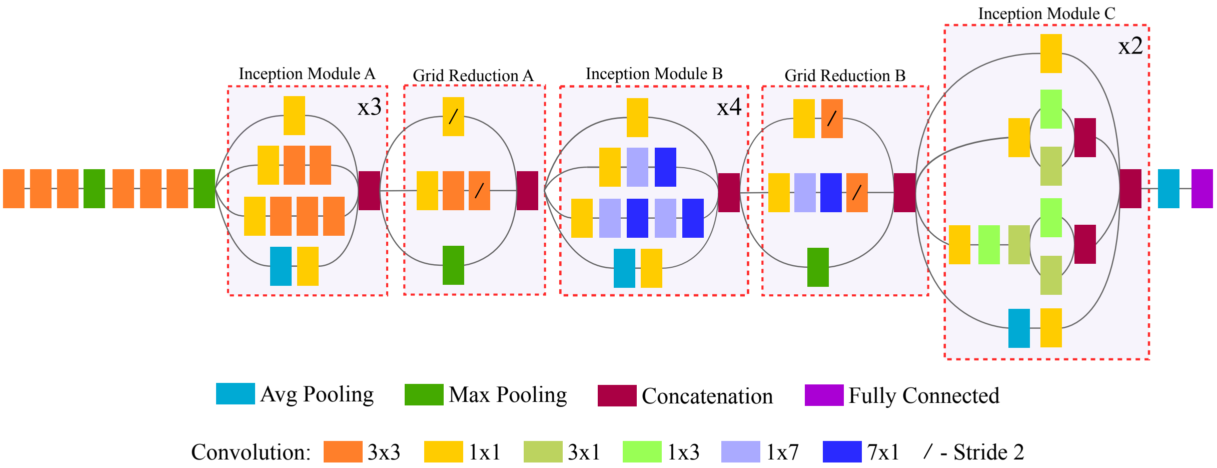


Figure 5.15: InceptionV3 simplified architecture illustrating the main Inception A through C blocks and Grid Reduction A,B blocks.

As the name implies, InceptionV3 derives from previous architectures while maintaining its core ideas. The GoogleLeNet (InceptionV1) [60], which introduced the Inception blocks, contains multiple filters of different dimensions engaging on the same level adding more width and depth to the network. From previous state-of-the-art architectures, it is known that by adding width to the network, it is possible to obtain more fine-grained features of the input. By adding depth, richer, and more complex representations of the input features are obtained.

InceptionV2 was presented along with InceptionV3, and they are similar, introducing Inception blocks A through C, as well as Grid Reduction A and B blocks (see Figure 5.15). These new inception blocks introduce convolution factorization, reducing weight parameters and computation from the original Inception block while maintaining performance. The 5×5 convolutions were replaced by two 3×3 convolutions, giving a close representation to the original. The convolution factorization is additionally propagated to more $n \times n$ convolutions which are replaced by $1 \times n$ followed by $n \times 1$ convolutions.

Max pooling layers reduces computational complexity at the cost of lowering feature representation and dimensionality. In InceptionV3 these are replaced by grid reduction modules that achieve the same purpose but maintaining richer representations through convolutional stride and concatenation.

Inception block A is a deeper variation of the original, replacing 5×5 convolutions through convolutional factorization with 3×3 convolutions to obtain richer features in early layers of the network. Inception block B is the factorized version of the original, replacing 7×7 with 1×7 and 7×1 convolutions. In Inception block C, more width is added from the original, to promote more complex features, replacing 3×3 convolutions with 1×3 and 3×1.

The last fully connected layer is replaced by one unit fully connected layer with Xavier weight initialization and linear activation.

$$f(x) = x. \tag{5.37}$$

It is trained under the mean square error loss function

$$\mathcal{L}_{iFR}^{reg} = MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.38}$$

to linearly estimate the iFR value.

Additional training details

A 5-fold stratified cross-validation at the patient-level was adopted, and all architectures were trained for a total of 50 epochs with a batch size of 32 using the *Adam* optimizer. The learning rate started at $\eta = 10^{-4}$ and was lowered by a factor of 0.6 on loss *plateau* or if the loss did not decrease after 5 epochs.

Chapter 6

Results

6.1 Metrics

6.1.1 Detection

To evaluate the performance of the detection models (see section 5.3), the mean average precision (mAP) and mean average recall (mAR) defined for the COCO dataset metrics [52] are adopted.

Both confidence scores and intersection-over-union (IoU) contribute to the criteria determining if a stenosis detection is a true positive (TP) or a false negative (FN). A detection is considered a TP only if it validates the following three conditions: (1) the estimated confidence score is above 0.5; (2) the estimated class matches the ground truth and (3) if the estimated bounding box as an IoU greater than a threshold IoU_{th}. If either of the latter two conditions fails to check, the detection is considered a false positive (FP). COCO metrics use $IoU_{th} = 0.5$ as a baseline. In our experiments it is set to $IoU_{th} = 0.2$. When the confidence score of the estimated detection that should correspond to ground truth is lower than 0.5 (condition (1)), it is considered a false negative (FN). If the detection's confidence score is lower than 0.5 and does not correspond to any ground truth (background), it is considered a true negative (TN).

Precision and Recall defined as

$$Precision = \frac{TP}{TP + FP}, \tag{6.1}$$

$$Recall = \frac{TP}{TP + FN}, \tag{6.2}$$

were used as performance metrics. By varying the IoU_{th} across a given interval, e.g. $IoU_{th} \in [0, 1.0]$, it is possible to achieve different precision/recall performances over IoU for the model, and a precision-recall (y-x axis) curve can be drawn. To avoid possible high variations of the curve, the precision is interpolated at 101 recall values, and average precision (AP) can is defined as the area under the interpolated curve (p_{int})

$$AP = \sum_{i=1}^{N} (r_{i+1} - r_i)\, p_{int}(r_{i+1}) \tag{6.3}$$

where r_i is the recall value at the i^{th} interpolated step, and average recall (AR) is defined as the recall averaged across the specified recall IoU interval.

Since AP and AR correspond to a single class, mean average precision (mAP) and mean average recall (mAR) are the mean across all K existing classes

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K}, \tag{6.4}$$

$$mAR = \frac{\sum_{i=1}^{K} AR_i}{K}, \tag{6.5}$$

where $K = 1$ for the stenosis detection.

6.1.2 Classification

To evaluate the performance of the viewing angle classification, both accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}, \tag{6.6}$$

and the F1 score

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{6.7}$$

were used. Accuracy is a more generic score, and F1 is a relation between precision and recall. Both values combined can give a more critical analysis of the performance since the dataset is not balanced.

Additionally, for the assessment of iFR, values of iFR are converted to a binary class, using a threshold of 0.89. This makes possible the computation of both the accuracy and the F1 score

6.1.3 Regression

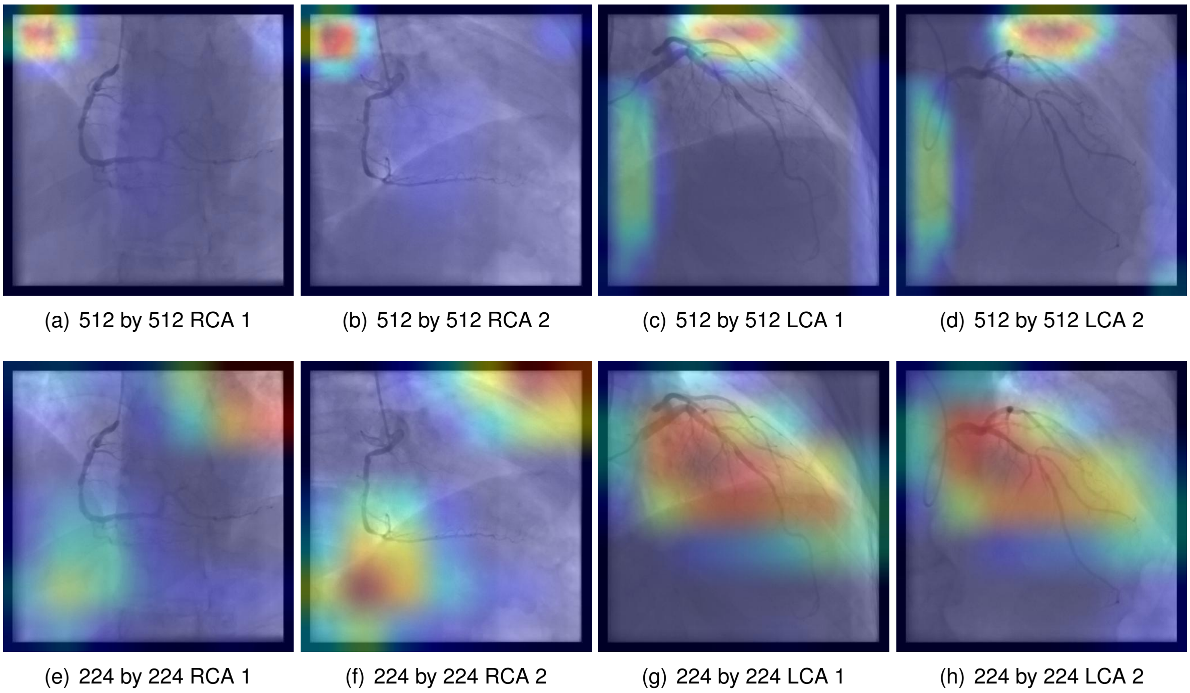
For the iFR regression task, performance is evaluated with the mean square error (MSE) (see Equation 5.38) and root mean squared error ($RMSE$)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}}. \tag{6.8}$$

Both measure how close the quantitative iFR regression estimated values are to the ground truth.

6.2 Angle selection performance

For the angle view selection task, the objective is to correctly classify to which coronary angle the respective reference frame belongs.

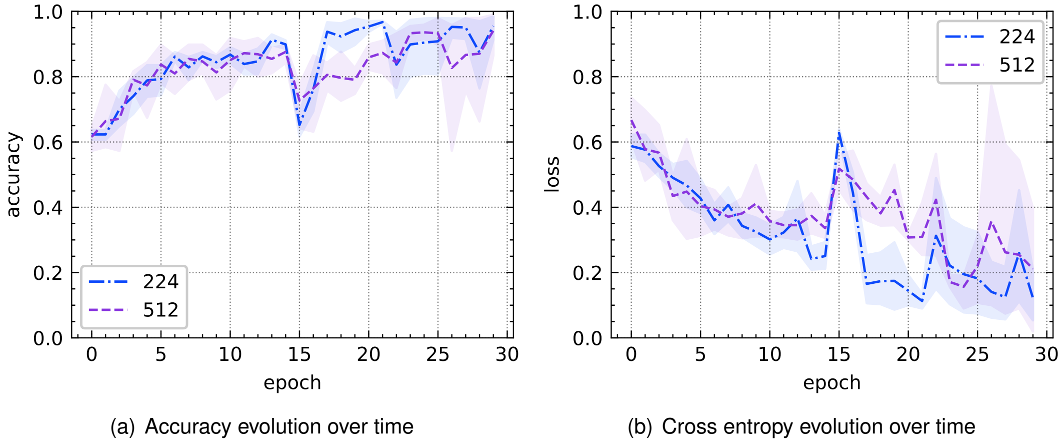


(a) 512 by 512 RCA 1 (b) 512 by 512 RCA 2 (c) 512 by 512 LCA 1 (d) 512 by 512 LCA 2

(e) 224 by 224 RCA 1 (f) 224 by 224 RCA 2 (g) 224 by 224 LCA 1 (h) 224 by 224 LCA 2

Figure 6.1: Grad-CAM visualizations in the RCA and LCA viewing angles showcasing the regions of the frame that most contribute to their correct classification.

From the original 512 by 512-pixel dimensions, a version of the input was generated by down-scaling it to 224x224. The objective was to understand if the model would still correctly classify the images with lower resolution.



(a) Accuracy evolution over time (b) Cross entropy evolution over time

Figure 6.2: 5-Fold Cross validation performance and loss evolution over time showing their respective mean and standard deviation for the viewing angle classification task for both 512 by 512 and 224 by 224-pixel images

On epoch 15 (see Figure 6.2), a spike occurs in accuracy and loss evolution due to the training implementation described in section 5.2, where part of the model was frozen until epoch 15 for the gradient updates. After epoch 15, the model recovers and continues to improve. If the freezing procedure was not applied, from early experiments, the evolution would *plateau* and would never reach such equivalent values in accuracy and loss, respectively.

To better understand which regions the model is focusing on the frame to decide the correct viewing angle, gradient-weighted class activation maps (Grad-CAM) [23] were employed to visualize the degree of contribution of specific image regions. The gradient of the detected class y^c is first computed, before the *softmax* activation function in regards to the last convolutional layer A, these are then global-average pooled over its dimensions i, j and neuron importance weights are obtained

$$\alpha^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}}, \tag{6.9}$$

with Z being the number of feature map pixels. A weighted operation to obtain the relevant regions is then followed with ReLu to eliminate negative influences on the class of interest

$$L^c = ReLU\left(\sum_k \alpha^c A\right), \tag{6.10}$$

giving the final coarse region of importance.

It is possible to observe (see Figure 6.1) that the larger 512 by 512-pixel resolution model, by having more parameters and larger feature map dimensions, instead of learning to focus on the coronary artery themselves to differentiate the designated viewing angle, it focuses on more fine-grained patterns of the human morphology, as a result from the variations of the C-arm X-ray unit. On the other hand, the 224 by 224 resolution model, due to the scaled-down resolution (resulting in lower-dimensional feature maps), captures the more broad patterns of the LCA whilst still detecting human morphology patterns in the RCA.

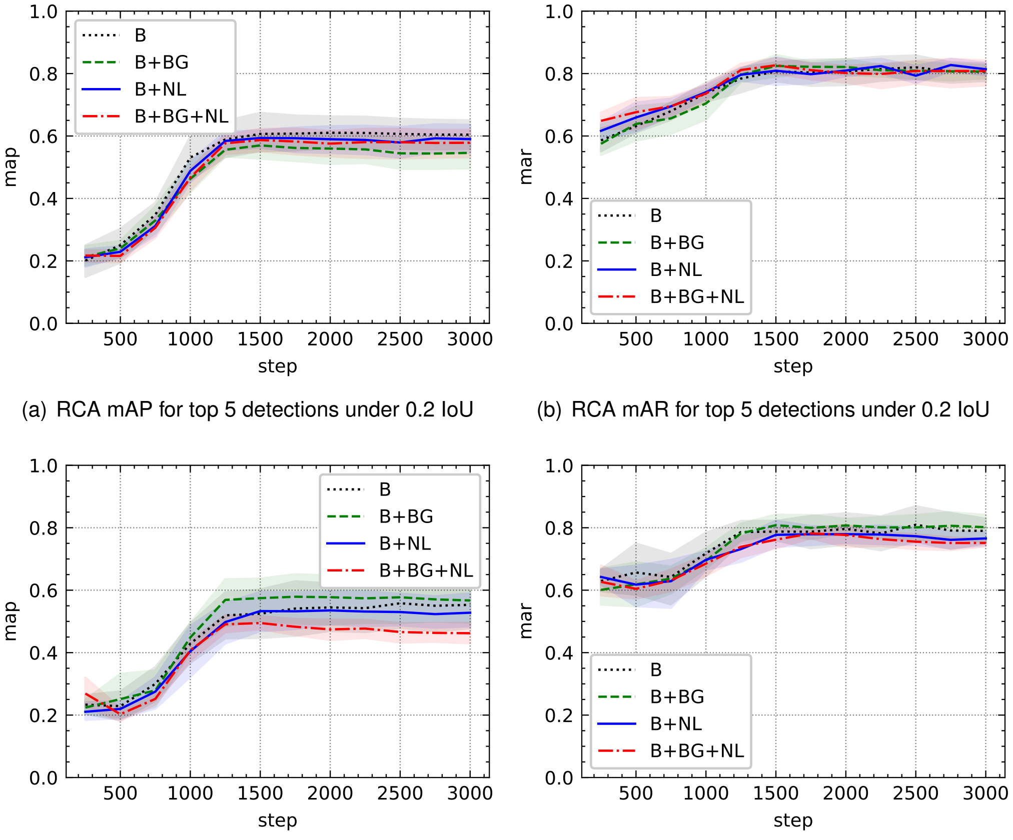
Image Dimensions	Accuracy	F1 Score	Cross-Entropy Loss
512	0.96±0.01	0.96±0.01	0.14±0.27
224	0.97±0.01	0.97±0.01	0.08±0.31

Table 6.1: Coronary viewing angles classification metrics performance

From the observed performance shown in Table 6.1, it is clear that even with a scaled-down version of the image, the model can correctly relate the images to their respective viewing angles. This is important since decreased dimensions significantly improve training and inference time. The scaled-down image input model shows marginal increases in accuracy and F1 score but a considerably lower value in the cross-entropy loss, which corresponds to more confidence in the viewing angles predictions.

6.3 Stenosis detection performance

For the stenosis detection task, the objective is to generate bounding box proposals with a high IoU and confidence score with reference to ground truth annotations. Two separate models with equal configurations were trained, specifically for the RCA viewing angle and another for the LCA. The performance is evaluated using COCO mean average precision (mAP) and recall (mAR) [52] for a maximum of five detections under an IoU of 0.2. Following Cong et al. [21], *sensitivity* is also defined as the recall rate of detection for a maximum of one detection, our highest confidence score detection, at an IoU threshold of 0.2 and confidence score over 0.5. Additionally, the performance of at least one candidate bounding box per sequence with a confidence score over 0.5, corresponding to a ground truth, is also shown.



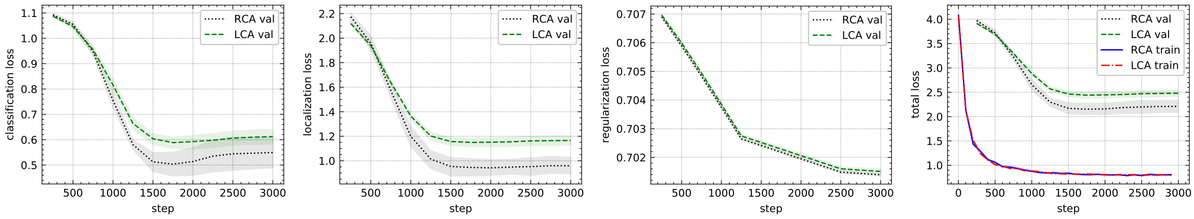
(a) RCA mAP for top 5 detections under 0.2 IoU

(b) RCA mAR for top 5 detections under 0.2 IoU

(c) LCA mAP for top 5 detections under 0.2 IoU

(d) LCA mAR for top 5 detections under 0.2 IoU

Figure 6.3: 5-Fold Cross-validation with standard deviation detection performance over time of the mean average precision (mAP), recall (mAR) for the RCA and LCA viewing angles. *B* denotes our baseline, *BG* denotes the addition of background (i.e., frames with no contrast) and *NL* denotes the addition of full contrast healthy RCA/LCA (no lesion/stenosis) frames



(a) Classification Loss

(b) Box Regression Loss

(c) Regularization Loss

(d) Total Loss

Figure 6.4: 5-Fold Cross-validation losses evolution with mean and standard deviation of the baseline RCA/LCA models (B) through all steps of the training procedure. Total loss is the sum of classification, regression and regularization loss (see section 5.3 for loss details)

Our baseline (*B*) is defined as having all frames from the full radio-opaque contrast interval (see chapter 4) included in training. Performance is evaluated only in reference frames. In attempts to improve the model's capacity at differentiating positive examples (stenosis) from negative ones (background/healthy coronaries), experiments were made including background (*BG*, frames without any contrast) and healthy coronary frames (*NL*) to the RCA and LCA model.

Analyzing the evolution of the baseline losses, the original data without the addition of background and healthy frames, for both RCA and LCA, a *plateau* in validation loss occurs while training loss continues to decrease (see Figure 6.4). The RCA loss shows a better performance than the LCA, because the training set was larger. Comparing the train set evolution with validation, we are in the situation of slight and healthy overfitting as the order of magnitude difference is not extreme, and the validation loss is not increasing. This loss difference was present in all variations and was common throughout all hyperparameter fine-tuning.

The faster-RCNN [56] architecture was also trained in early experiments but was not re-trained because of the amount of overfitting in addition to the lack of performance. More recently published state-of-the-art architectures for object detection EfficientDet [61] and CenterNet [62] were also tested in our experiments. Although they performed well, they were inferior to the tested RetinaNet variations. Lastly, the DETR architecture [63], which uses the transformer [59] with attention mechanisms for object detection was also tested. Still, due to the extensive amount of parameters, fine-tuning was not feasible.

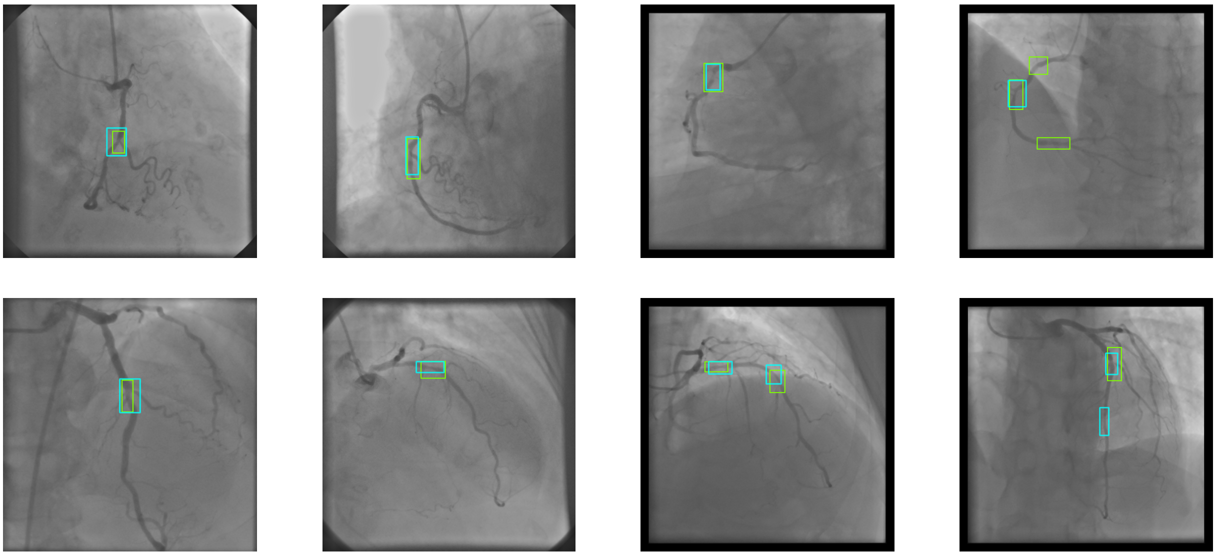


Figure 6.5: Stenosis detection examples in validation set for RCA (top) and LCA (bottom) viewing angles, with cyan bounding boxes denoting ground truth annotations and green representing the estimated ones.

From visual validation of the model's performance (see Figure 6.5), it is possible to observe that it performs better in frames with only one ground truth stenosis. However, in cases where more than one is present, the model struggles at the detection in its entirety. It is possible to regulate the number of detections the model predicts, whether true positives or false positives, by varying the matched threshold m_{th}^{IoU}. With a value too high, the detections decrease, harming all metrics, and with a value too low, even with non-maximum suppression, the number of detections becomes overwhelming. Balancing this hyperparameter was found to be of key importance. For aspect ratios definition, it was empirically found, the more, the better until it reaches a *plateau* in performance. However, due to memory limitations, it is not feasible to scale this hyperparameter.

Our detection model is configured to output a maximum of 100 bounding boxes at inference. But for evaluation mAP and mAR are set only to evaluate a maximum of five detections. This is a more strict measure of performance, since by increasing the maximum number of possible detections, mAR would increase as it measures the rate of true positives (correctly identified stenosis) against false negatives (stenosis detected as background). However, mAP would drastically decrease since it counts the false positives (background detected as stenosis). Both metrics combined then provide a better understanding of the model's real performance.

Viewing Angle	Method	Stenosis Detection Performance @ 0.2 IoU [mean ± std)]			
		Sensitivity	At least One	mAP max 5 det	mAR max 5 det
RCA	B	**0.72±0.03**	**0.81±0.01**	**0.61±0.03**	0.82±0.02
	B w/ BG	0.64±0.03	0.74±0.03	0.57±0.02	0.82±0.02
	B w/ NL	0.68±0.04	0.74±0.04	0.59±0.02	**0.83±0.01**
	B w/ BG w/ NL	0.65±0.03	0.71±0.02	0.59±0.02	0.83±0.01
	Cong et al. [21]	0.71	-	-	-
LCA	B	0.68±0.04	**0.77±0.03**	0.56±0.04	**0.81±0.03**
	B w/ BG	**0.70±0.04**	0.74±0.04	**0.58±0.04**	0.81±0.02
	B w/ NL	0.65±0.02	0.71±0.02	0.54±0.03	0.78±0.02
	B w/ BG w/ NL	0.58±0.03	0.65±0.03	0.49±0.01	0.78±0.01
	Cong et al. [21]	0.60	-	-	-

Table 6.2: Stenosis detection metrics comparison on reference frames against different authors

We compare all our models' performance in different settings against themselves and with previous work by other authors (see Table 6.2). The work from Au et al. [18], as the detection stage of their pipeline, only estimate regions of 192 by 192 pixel for the RCA viewing angle, are not comparable. Additionally, the results from Wu et al. [24] are left out from comparison purposes since different metrics were used, and their work only involves detections for confirmed stenosis above the QCA threshold. These are easily classified positive examples, in contrast with ours, which includes hard to classify examples, such as visually annotated stenosis above the iFR threshold.

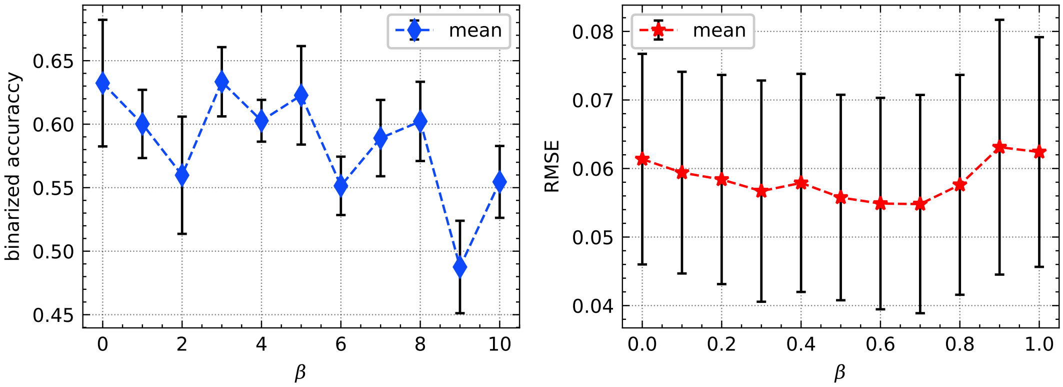
The results show that our variations on the base model do not improve performance. Nevertheless, the models can detect several stenoses per frame even with hard examples (iFR above threshold), achieving good performances.

6.4 iFR regression performance

For the stenosis assessment task, the objective is to estimate its corresponding iFR value. We evaluate the performance under binarized accuracy (see subsection 6.1.2) and analyze the error compared to their corresponding target.

Distinct models were trained under 5-Fold cross-validation for different values of the hyperparameter β, which varied from 0 to 1 with a step of 0.1, with the objective of studying the model's ability to estimate the iFR quantitative value.

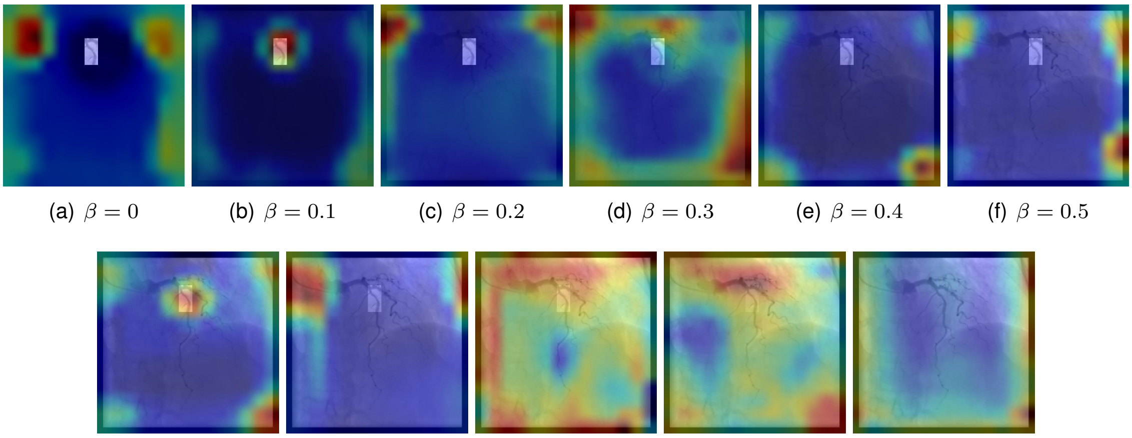
For additional validation, Grad-CAM was again applied. From the performance metrics evolution (see Figure 6.6), it is possible to observe that as β increases, there is a very slight downtrend in the binarized accuracy. Still, in the error, it only up-trends for the last values (see Figure 6.6), indicating



(a) Binarized accuraccy performance

(b) Root mean square error loss

Figure 6.6: The binarized accuracy and root mean square error evolution against the β variation of the outer regions regions of interest in the LCA frames. RMSE is shown for better error interpretability.

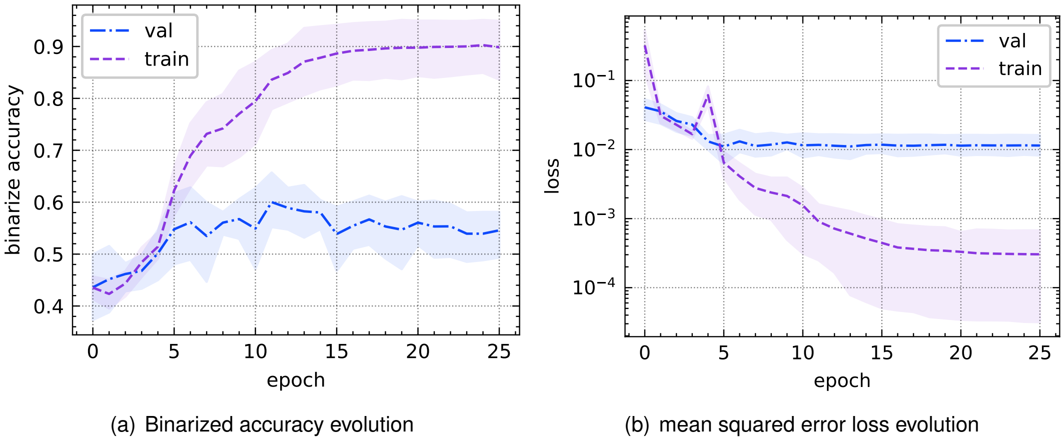


(a) $\beta = 0$ (b) $\beta = 0.1$ (c) $\beta = 0.2$ (d) $\beta = 0.3$ (e) $\beta = 0.4$ (f) $\beta = 0.5$

(g) $\beta = 0.6$ (h) $\beta = 0.7$ (i) $\beta = 0.8$ (j) $\beta = 0.9$ (k) $\beta = 1$

Figure 6.7: Grad-CAM visualizations through all β variations where it's possible to see the highlighted stenosis and the model's main focus of interest to estimate the iFR.

that the models failed to learn the task. With the Grad-CAM visualizations (see Figure 6.7) the $\beta = 0.1$ and $\beta = 0.6$ models started to take more information from the highlighted stenosis region, but all others failed. From $\beta = 0.8$ to $\beta = 1$ the model loses its focus and takes random guesses at every region, since outer regions were given more weight. All other models focus on all regions, but the highlighted one, again indicating randomness.



(a) Binarized accuracy evolution

(b) mean squared error loss evolution

Figure 6.8: The binarized accuracy and mean square error evolution of train and validation set with respective means and standard deviation for $\beta = 0.1$ from 5-Fold cross-validation

From the loss evolution during training (see Figure 6.8), it is clear that the model completely overfitted as the validation loss is orders of magnitude higher than the training loss. As in this task, it is required for the model to interpret very fine-grained features. During experimentation, it was noticeable that lower width and depth models overfitted even more and failed to improve, showing even loss increase. Many variations and hyperparameter fine-tuning were attempted but with no results. Dropout regularization was excluded as in this task adding randomness in training deteriorated, even more, the performance. Learning rate and optimizer were also key factors since Adam [50] converges quickly, and the model is more prone to overfit. Other optimizers were tried giving the same result. Standard SGD was tried and discarded since it requires many more epochs for improvements, and the time constraint was an issue. Instead of regression, classification methods were also tried by having two classes: one for iFR above the threshold and another iFR below the threshold. These models failed completely, giving just one-sided predictions. No positive nor negative conclusions can be made with this experiment as more work needs to be carried for the iFR assessment.

Chapter 7

Conclusions

From the start of this study, the objective was to conduct relevant research in coronary artery disease and automated stenosis assessment to contribute to both medical and machine learning fields.

At the time of writing, no public datasets of invasive coronary angiography (ICA) with stenosis labels and annotations are publicly available. From the unprocessed, unlabeled, and de-identified *Hospital de Santa Maria* (ICA) dataset, it was possible to create a newly organized data structure at the patient level containing every corresponding sequence, iFR values, and coronary artery annotations. For each sequence, from the coronary arteries annotations, it was possible to apply a new automated method with discriminative filters for bounding box propagation, enriching every frame of the sequence with more metadata. Additionally, intervals of optimal frames were also defined, and common viewing angles were also grouped. Contributing to an optimal and curated dataset containing 1294 sequences from 392 patients, it's sought to make it accessible to the research community enticing the creation and development of new methods for stenosis assessment.

A three-stage framework based on convolutional neural networks was assembled to automate stenosis assessment. It started with viewing angle selection. The objective was to classify reference frames as belonging to the right coronary artery (RCA) or the left coronary artery (LCA), with previously viewing angle labels. High-performance metrics of 0.97 accuracy and 0.97 F1 score were obtained with transfer learning and fine-tuning of the ResNet-50, demonstrating the feasibility of frame down-scaling to increase inference time memory optimization.

With the RCA and LCA viewing angles, two distinct models, based on the RetinaNet single-shot detector architecture, were assembled as the second stage of the framework to automatically detect stenosis by bounding box placement. Comparisons with the work of different authors confirm our work's relevance with obtained scores of 0.72/0.70 sensitivity, 0.83/0.81 mAR, and 0.61/0.58 mAP for the RCA/LCA. Our models performed well at stenosis single and multiple detections but leave room for precision improvement as many background regions were detected as stenosis.

The last stage of our framework ends with iFR quantitative assessment. Experiments were made with transfer learning and fine-tuning of InceptionV3 as a regression task. A β method approach was implemented that varies the outer regions contrast of the stenosis bounding box to boost the models focus on the region that most contributes to the iFR real value. From gradient-weighted class activation mapping visualizations and metrics evolution, almost all models failed to gain a notion showing overfitting and randomness in regression values, leaving room for improvement in this task.

Chapter 8

Future Work

Our processing method discarded a significant amount of sequences to be curated. Further medical data is also expected to be received and possibly with segmentation annotations. As neural network models improve with more information, data processing will be necessary to improve the dataset.

In all stages of the framework, transfer learning and fine-tuning were exploited since they eliminated the time constraint from training from scratch and performed almost as equal. Experiments on designing custom architectures suited for the tasks can be made to boost performance.

The assembled framework also considers that the reference frame is selected *a priori*. An additional stage can be made taking advantage of the sequence temporal dimension. Since our labeling and annotation process defined all possible intervals of the sequence together with the specific reference frame, this task can be further explored with recurrent neural network application.

One post-processing procedure in stenosis detection was the application of non-maximum suppression that checks for overlapping candidate bounding boxes at the frame level, improving precision. The application of sequence non-maximum suppression can also be explored, selecting the best candidate bounding box by taking into account detections in the entire frame interval. Different frames are also suited for stenosis representation.

The most difficult task to overcome was the iFR regression task, as our method did not quite perform to expectations. New novel experiments can be done with possible incoming segmentation annotations of stenosis and coronary arteries, for example, automatically segment and classify the regions of interest obtained from the bounding boxes. Attention mechanisms are being increasingly developed and published in the literature, demonstrating their applications in all research fields. These mechanisms can be explored for iFR estimates since the stenosis only represents a small portion of the frame. It could help the model learn that these specific regions contribute the most to the iFR assessment.

Bibliography

[1] G. A. Roth, D. Abate, K. H. Abate, and et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1736 – 1788, 2018.

[2] H. Zhang, L. Mu, S. Hu, B. K. Nallamothu, A. J. Lansky, B. Xu, G. Bouras, D. J. Cohen, J. A. Spertus, F. A. Masoudi, J. P. Curtis, R. Gao, J. Ge, Y. Yang, J. Li, X. Li, X. Zheng, Y. Li, H. M. Krumholz, and L. Jiang. Comparison of Physician Visual Assessment With Quantitative Coronary Angiography in Assessment of Stenosis Severity in China. *JAMA Intern Med*, 178(2):239–247, Feb 2018.

[3] S. K. Bhatia. *Biomaterials for Clinical Applications*. Springer, 1^{st} edition, 2010.

[4] *Ischemic Heart Disease*. National Heart, Lung, and Blood Institute, 2013. https://www.nhlbi.nih.gov/health-topics/ischemic-heart-disease.

[5] E. J. Topol and S. E. Nissen. Our preoccupation with coronary luminology. The dissociation between clinical and angiographic findings in ischemic heart disease. *Circulation*, 92(8):2333–2342, Oct 1995.

[6] P. W. Serruys, J. H. Reiber, W. Wijns, and et. al. Assessment of percutaneous transluminal coronary angioplasty by quantitative coronary angiography: diameter versus densitometric area measurements. *Am. J. Cardiol.*, 54(6):482–488, Sep 1984.

[7] D. P. Foley, J. Escaned, B. H. Strauss, C. di Mario, J. Haase, D. Keane, W. R. Hermans, B. J. Rensing, P. J. de Feyter, and P. W. Serruys. Quantitative coronary angiography (QCA) in interventional cardiology: clinical application of QCA measurements. *Prog Cardiovasc Dis*, 36(5):363–384, 1994.

[8] A. F. members, S. Windecker, P. Kolh, F. Alfonso, J.-P. Collet, J. Cremer, V. Falk, G. Filippatos, C. Hamm, S. J. Head, P. Jüni, A. P. Kappetein, A. Kastrati, J. Knuuti, U. Landmesser, G. Laufer, F.-J. Neumann, D. J. Richter, P. Schauerte, M. Sousa Uva, G. G. Stefanini, D. P. Taggart, L. Torracca, M. Valgimigli, W. Wijns, A. Witkowski, E. C. for Practice Guidelines, J. L. Zamorano, S. Achenbach, H. Baumgartner, J. J. Bax, H. Bueno, V. Dean, C. Deaton, C. Erol, R. Fagard, R. Ferrari, D. Hasdai, A. W. Hoes, P. Kirchhof, J. Knuuti, P. Kolh, P. Lancellotti, A. Linhart, P. Nihoyannopoulos, M. F. Piepoli, P. Ponikowski, P. A. Sirnes, J. L. Tamargo, M. Tendera, A. Torbicki, W. Wijns, S. Windecker, E. C. G. Committee, M. Sousa Uva, D. reviewers, S. Achenbach, J. Pepper, A. Anyanwu, L. Badimon, J. Bauersachs, A. Baumbach, F. Beygui, N. Bonaros, M. De Carlo, C. Deaton, D. Dobrev,

J. Dunning, E. Eeckhout, S. Gielen, D. Hasdai, P. Kirchhof, H. Luckraz, H. Mahrholdt, G. Montalescot, D. Paparella, A. J. Rastan, M. Sanmartin, P. Sergeant, S. Silber, J. Tamargo, J. ten Berg, H. Thiele, R.-J. van Geuns, H.-O. Wagner, S. Wassmann, O. Wendler, J. L. Zamorano, F. Weidinger, F. Ibrahimov, V. Legrand, I. Terzić, A. Postadzhiyan, B. Skoric, G. M. Georgiou, M. Zelizko, A. Junker, J. Eha, H. Romppanen, J.-L. Bonnet, A. Aladashvili, R. Hambrecht, D. Becker, T. Gudnason, A. Segev, R. Bugiardini, O. Sakhov, A. Mirrakhimov, B. Pereira, H. Felice, T. Trovik, D. Dudek, H. Pereira, M. A. Nedeljkovic, M. Hudec, A. Cequier, D. Erlinge, M. Roffi, S. Kedev, F. Addad, A. Yildirir, and J. Davies. 2014 ESC/EACTS Guidelines on myocardial revascularization: The Task Force on Myocardial Revascularization of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS)Developed with the special contribution of the European Association of Percutaneous Cardiovascular Interventions (EAPCI). *European Heart Journal*, 35(37):2541–2619, Oct 2014.

[9] S. Achenbach, T. Rudolph, J. Rieber, H. Eggebrecht, G. Richardt, T. Schmitz, N. Werner, F. Boenner, and H. Mollmann. Performing and Interpreting Fractional Flow Reserve Measurements in Clinical Practice: An Expert Consensus Document. *Interv Cardiol*, 12(2):97–109, Sep 2017.

[10] F.-J. Neumann, M. Sousa-Uva, A. Ahlsson, F. Alfonso, A. P. Banning, U. Benedetto, R. A. Byrne, J.-P. Collet, V. Falk, S. J. Head, P. Jüni, A. Kastrati, A. Koller, S. D. Kristensen, J. Niebauer, D. J. Richter, P. M. Seferović, D. Sibbing, G. G. Stefanini, S. Windecker, R. Yadav, M. O. Zembala, and E. S. D. Group. 2018 ESC/EACTS Guidelines on myocardial revascularization. *European Heart Journal*, 40(2):87–165, Aug 2018.

[11] S. Baumann, L. Chandra, E. Skarga, M. Renker, M. Borggrefe, I. Akin, and D. Lossnitzer. Instantaneous wave-free ratio (ifr®) to determine hemodynamically significant coronary stenosis: A comprehensive review. *World J Cardiol*, 10(12):267–277, 2018.

[12] M. Avendi, A. Kheradvar, and H. Jafarkhani. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri. *Medical Image Analysis*, 30:108–119, 2016.

[13] L. K. Tan, Y. M. Liew, E. Lim, and R. A. McLaughlin. Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine mr sequences. *Medical Image Analysis*, 39: 78 – 86, 2017.

[14] Z. Li, A. Lin, X. Yang, and J. Wu. Left ventricle segmentation by combining convolution neural network with active contour model and tensor voting in short-axis mri. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 736–739, 2017.

[15] R. P. K. Poudel, P. Lamata, and G. Montana. Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation. In M. A. Zuluaga, K. Bhatia, B. Kainz, M. H. Moghari, and D. F. Pace, editors, *Reconstruction, Segmentation, and Analysis of Medical Images*, pages 83–94, Cham, 2017. Springer International Publishing.

[16] M. Zreik, T. Leiner, B. D. de Vos, R. W. van Hamersvelt, M. A. Viergever, and I. Išgum. Automatic segmentation of the left ventricle in cardiac ct angiography using convolutional neural networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 40–43, 2016.

[17] K. Antczak and L. Liberadzki. Stenosis detection with deep convolutional neural networks. *MATEC Web of Conferences*, 210:04001, Jan 2018.

[18] B. Au, U. Shaham, S. Dhruva, G. Bouras, E. Cristea, A. Lansky, A. Coppi, F. Warner, S. Li, and H. M. Krumholz. Automated characterization of stenosis in invasive coronary angiography images with convolutional neural networks. *CoRR*, abs/1807.10597, 2018.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[20] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[21] C. Cong, Y. Kato, H. D. Vasconcellos, J. Lima, and B. Venkatesh. Automated stenosis detection and classification in x-ray angiography using deep neural network. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1301–1308, 2019.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[24] W. Wu, J. Zhang, H. Xie, Y. Zhao, S. Zhang, and L. Gu. Automatic detection of coronary artery stenosis by convolutional neural network with temporal constraint. *Computers in Biology and Medicine*, 118:103657, 2020.

[25] Coronary arteries, (n.d). Retrieved from https://www.texasheart.org/heart-health/heart-information-center/topics/the-coronary-arteries/.

[26] Anatomy and function of the coronary arteries, (n.d). Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/anatomy-and-function-of-the-coronary-arteries.

[27] Coronary artery disease, 2019. Retrieved from https://my.clevelandclinic.org/health/diseases/16898-coronary-artery-disease.

[28] W. N. Epidemiological studies of chd and the evolution of preventive cardiology. *Nat Rev Cardio*, 11(5):276–289, Mar 2014.

[29] Wikidoc. Coronary circulation, 2013. URL `https://www.wikidoc.org/index.php/Coronary_circulation`. [Online; accessed 31-December-2019].

[30] Wikidoc. Coronary angiography standard views, 2013. URL `https://www.wikidoc.org/index.php/Coronary_angiography_standard_views`. [Online; accessed 31-December-2019].

[31] D. P. Foley, J. Escaned, B. H. Strauss, C. di Mario, J. Haase, D. Keane, W. R. Hermans, B. J. Rensing, P. J. de Feyter, and P. W. Serruys. Quantitative coronary angiography (QCA) in interventional cardiology: clinical application of QCA measurements. *Prog Cardiovasc Dis*, 36(5):363–384, 1994.

[32] S. Achenbach, T. Rudolph, J. Rieber, H. Eggebrecht, G. Richardt, T. Schmitz, N. Werner, F. Boenner, and H. M?llmann. Performing and Interpreting Fractional Flow Reserve Measurements in Clinical Practice: An Expert Consensus Document. *Interv Cardiol*, 12(2):97–109, Sep 2017.

[33] S. Nijjer and J. Davies. *Physiologic Assessment in the Cardiac Catheterization Laboratory*, chapter 6, pages 59–70. John Wiley & Sons, Ltd, 2016.

[34] K. H. Parker. An introduction to wave intensity analysis. *Med Biol Eng Comput*, 47(2):175–188, Feb 2009.

[35] S. Sen, K. N. Asrress, S. Nijjer, R. Petraco, I. S. Malik, R. A. Foale, G. W. Mikhail, N. Foin, C. Broyd, N. Hadjiloizou, A. Sethi, M. Al-Bustami, D. Hackett, M. A. Khan, M. Z. Khawaja, C. S. Baker, M. Bellamy, K. H. Parker, A. D. Hughes, D. P. Francis, J. Mayet, C. D. Mario, J. Escaned, S. Redwood, and J. E. Davies. Diagnostic classification of the instantaneous wave-free ratio is equivalent to fractional flow reserve and is not improved with adenosine administration: Results of clarify (classification accuracy of pressure-only ratios against indices using flow study). *Journal of the American College of Cardiology*, 61(13):1409 – 1420, 2013.

[36] A. Rosset, L. Spadola, and O. Ratib. Osirix: An open-source software for navigating in multidimensional dicom images. *Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology*, 17:205–16, Oct 2004.

[37] A. Lukežič, T. Vojíř, L. Čehovin Zajc, J. Matas, and M. Kristan. Discriminative correlation filter tracker with channel and spatial reliability. *International Journal of Computer Vision*, 126(7):671–688, Jan 2018.

[38] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, Dec 1943.

[39] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*, 65(6):386–408, Nov 1958.

[40] J. S. Marques. *Reconhecimento de Padrões*. IST - Instituto Superior Técnico, Lisbon, Portugal, 2005.

[41] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., USA, 1 edition, 1997.

[42] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2(11), Nov. 2017.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[44] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, pages 3347–3357, 2019.

[45] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, page 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[46] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[47] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress.

[48] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, Jul 2015. PMLR.

[49] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, May 2010. JMLR Workshop and Conference Proceedings.

[50] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[51] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015.

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[53] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[54] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[55] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[56] R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[57] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[58] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, Jul 2015. PMLR.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[60] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[61] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10778–10787, 2020.

[62] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. Centernet: Keypoint triplets for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6568–6577, 2019.

[63] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.