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

Dinis Lourenço Tavares Rodrigues
dinis.rodrigues@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

February 2020

***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*. 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.**

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

1. Introduction

Coronary artery disease (CAD), characterized by plaque buildup inside the coronary arteries, is the leading non-communicable disease in global mortality. This buildup leads to stenosis, partially or totally blocking 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. Current standard diagnosis methods rely on an expert physician to assess the issue, off or on-site, using non-invasive or invasive procedures. [1, 2] Although several resources have been invested in prevention, proper available CAD assessment, and treatment procedures still aren't reachable to the most general public.

A contribution is made to the research field by providing a new curated medical dataset of X-ray invasive coronary angiography labeled with optimal interval frames and annotated with stenosis bounding boxes and its respective 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.

2. State-of-the-Art

Antczak and Liberadzki [3] generated and trained upon thousands of artificial 32 by 32 pixel patches mimicking the presence of stenosis, followed by convolutions through a sliding window on the original frame lead to an improved detection performance, but real test images were very few and were also scaled down.

To deal with the lack of public ICA datasets, Antczak and Liberadzki [3] 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. [4] 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 par-

			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

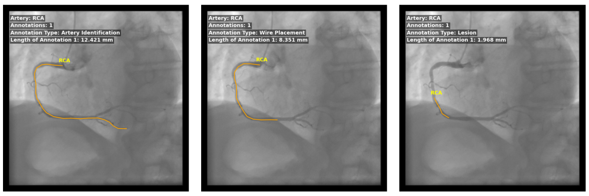
Table 1: Processed dataset, with all optimal sequences, frame intervals, stenosis annotations, and iFR values count.

ticipants using only RCA viewing angles and reference frames. A detector variant of the single-shot detector YOLO [5] 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 [6], 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. [7] also developed a three-stage end-to-end workflow for stenosis characterization. The process of viewing angle selection is initialized with transfer learning and fine-tuning of the InceptionV3 [8]. Features extracted from the last convolution layer of the InceptionV3, are used to train a bidirectional 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 [9] 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 only in the detection task of the stenosis Wu et al. [10] 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.

3. Medical Data

3.1. Overview

An expert cardiologist firstly curated the available data for this work 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 patients, ranging from 2015 until 2019. The data was properly de-identified to preserve participant privacy, and each subject was over the age of eighteen.


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 (b) Wire (c) Stenosis

Figure 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 each subject, *iFR Value* and *Coronary Artery Stenosis Location* was included at the patient level. 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* [11], an image processing software (see Figure 1).

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.

3.2. Data Treatment & Annotation Procedure

The provided ICA image sequences were in the well known and documented DICOM format protocol. Only sequences with frame dimensions of 512 by 512 pixel were allowed with pixel values ranging in the $[0, 255]$ monochromatic scale (1 channel). 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* [12] was implemented using the OpenCV image processing library. The propagation is made through a three-step process: (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. Misplaced bonding boxes due to rapid shifts from frame to frame and/or occlusions were *a posteriori* manually addressed to have a perfect fit to the stenosis.

For each sequence four different frame intervals were also labelled as: (a) No radio-opaque contrast; (b) The radio-opaque contrast is being introduced; (c) The radio-opaque contrast has been fully introduced (Optimal frame); (d) The radio-opaque contrast is vanishing. Sequences were also grouped by their respective angles (obtained by the DICOM metadata) and then manually filtered to the most common viewing angle names. Sequences with metal implants, pacemakers and with the iFR medical suited wires were discarded. The 1593 annotated sequences from 498 patients were processed with the described steps, resulting in 1294 optimal sequences with 20819 optimal frames, where the bounding box was placed (see Table 1).

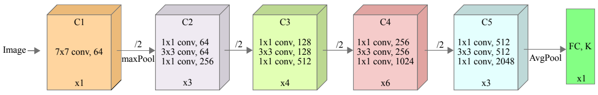


Figure 2: ResNet-50 simplified architecture, bottom of each block denotes the repeated set of layers (Kernel size, operation, number of channels) in each convolution block (C), after each block ($C1$ to $C5$) spatial dimension is reduced. Ending in a fully connected layer with K units. Shortcut connections occur every two layers but are omitted for readability.

4. Implementation

4.1. Angle View Selection

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. 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 [13]. 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. For each layer of the ResNet network, except the fully connected layer, the ReLU activation function [14] is used with batch normalization layers [15]. In our application, the ResNet-50 was initially pre-trained on ImageNet with 224 by 224 images. A fully connected layer replaced the last layer with two output units, ending with a softmax activation function. It was then trained under categorical cross-entropy with loss defined as

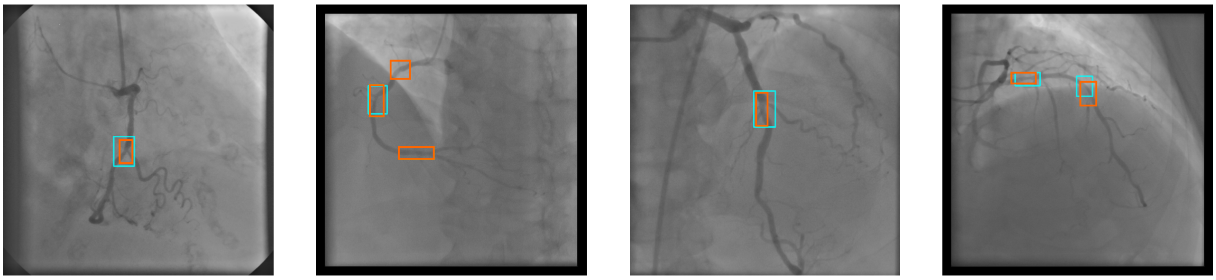
$$L_{cls}^{angle} = \text{CCE} = -\sum_{i=1}^{K} y_i \log(\hat{y}_i). \qquad (1)$$

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* [16] optimizer was performed. The initial learning rate was set at $\eta = 10^{-5}$, being reduced by a factor of 0.2 on loss *plateau*. 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.

4.2. 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.

We decided to adopt a state-of-the-art architecture for object detection, the RetinaNet [17], which was first pre-trained on the COCO dataset. The RetinaNet's architecture is based on the unified single-shot detector architecture, composed of a



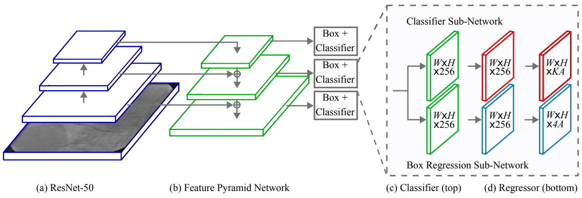
RCA 1

RCA 2

LCA 1

LCA 2

Figure 3: Stenosis detection examples in validation set for RCA and LCA viewing angles with cyan bounding boxes denoting ground truth annotations and oraange representing the estimated ones.



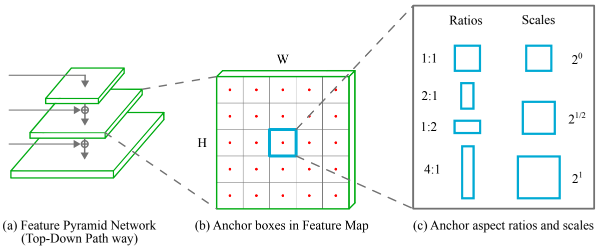
(a) ResNet-50

(b) Feature Pyramid Network

(c) Classifier (top)

(d) Regressor (bottom)

Figure 4: 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.



(a) Feature Pyramid Network (Top-Down Path way)

(b) Anchor boxes in Feature Map

(c) Anchor aspect ratios and scales

Figure 5: 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

backbone and two additional sub-networks (see Figure 4). 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.

In the backbone, the ResNet-50 is first applied for deep image feature extraction. On top of the ResNet-50, Feature Pyramid Networks (FPN) [18] is also adopted as a top-down pathway to complete the RetinaNet architecture's backbone. 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.

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. 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), 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. The assignment of ground-truth bounding boxes to each generated anchor is made by setting a matching Intersection-over-Union (IoU) threshold $m_{th}^{IoU} = 0.2$ between the anchor and the ground truth. 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.

To identify the presence of 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 function 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. For each 512 by 512 pixel, ≈45 thousand anchor boxes are generated. To deal with the amount of generated bounding boxes and occurring class imbalance between background and stenosis assignment, the α-balanced Focal Loss function [17] is used. It is normalized by the number of previously assigned anchors to ground truth stenosis N_g.

$$L_{cls}^{det} = -\left(y_i \log(p_i)^{\gamma}\alpha + (1 - y_i)\log(1 - p_i)p_i^{\gamma}(1 - \alpha)\right) \tag{2}$$

with

$$p_t = \begin{cases} p & \text{, if} \quad y = 1 \\ 1 - p & \text{, otherwise,} \end{cases} \tag{3}$$

where $\alpha = 0.25$ and $\gamma = 2$ were defined by experimentation for this stenosis detection task. 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 regression 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 [19] for each matched pair $(\hat{A}, G)$ as

$$t_x = \left(G_x - \hat{A}_x\right) / \hat{A}_w \tag{4}$$

$$t_y = \left(G_y - \hat{A}_y\right) / \hat{A}_y \tag{5}$$

$$t_w = \log\left(G_w / \hat{A}_w\right) \tag{6}$$

$$t_h = \log\left(G_h / \hat{A}_h\right) \tag{7}$$

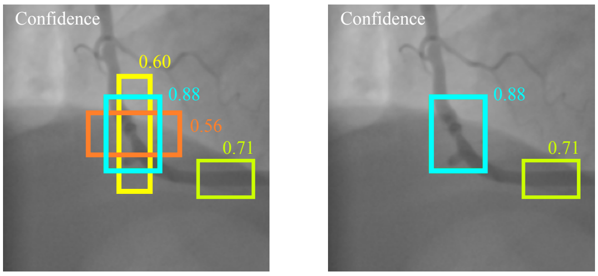
(t_x, t_y) denotes a center scaling invariant translation, and (t_w, t_h) represent logarithmic space translations of the estimated width and height anchor $\hat{A}$. The network is then trained to estimate these parameterized coordinates offset T under the smooth L_1 loss function [20]

$$\text{smooth}_{L_1} = \begin{cases} 0.5x^2 & \text{, if} \quad |x| < 1 \\ |x| - 0.5 & \text{, otherwise,} \end{cases} \tag{8}$$

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

$$\mathcal{L}_{det}^{reg} = \sum_{j \in \{x,y,w,h\}} \text{smooth}_{L1}\left(T_j - \hat{T}_j\right) \tag{9}$$

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



(a) Before NMS

(b) After NMS

Figure 6: 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.

To deal with the amount of generated candidate bounding boxes in at inference, where overlapping occurrences exist, the non-maximum suppression algorithm is applied with a score threshold $NMS_{th}^{cls} = 0$ and an IoU threshold defined as $NMS_{th}^{IoU} = 0.5$.

The model was trained and evaluated under 5-fold stratified cross-validation at patient-level, for 3500 steps (≈ 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

$$\mathcal{L}_{det}^{2} = \lambda \sum_{i=1}^{W} w_i^2. \tag{10}$$

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

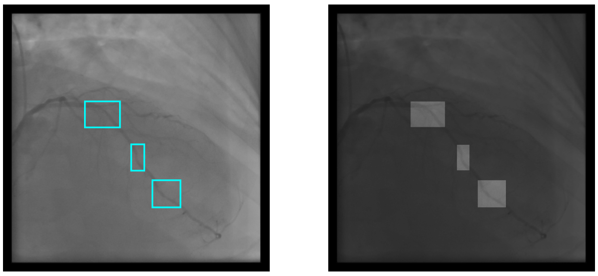
$$\mathcal{L}_{det} = \mathcal{L}_{det}^{cls} + \mathcal{L}_{det}^{reg} + \mathcal{L}_{det}^{2}. \tag{11}$$

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 in brightness and contrast from the original frames to reduce overfitting and aid the model at generalization.

4.3. Stenosis Severity Regression

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



(a) Original set of bounding boxes

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

Figure 7: Proposed approach of contrast regulation from the (a) the original image with the bounding boxes follows (b) the modification of outer regions RGB values.

With inspiration from the *Hard* and *Soft Attention* principles [21–23], An intuitive approach defined β-method was used. 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. Outside the bounding boxes, each pixel is multiplied by β

$$g(x^*, y^*) = \beta f(x^*, y^*) \tag{12}$$

with $f(.)$ representing the three-channel RGB original image, where x^* and y^* are the pixel coordinates outside the bounding boxes. To quantify the iFR assessment value, the InceptionV3 [8] was chosen.

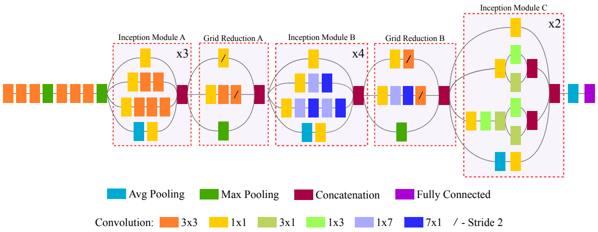


Figure 8: InceptionV3 simplified architecture illustrating the main Inception A through C blocks and Grid Reduction A,B blocks.

It is trained under the mean square error loss function

$$L_{reg}^{iFR} = MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13}$$

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.

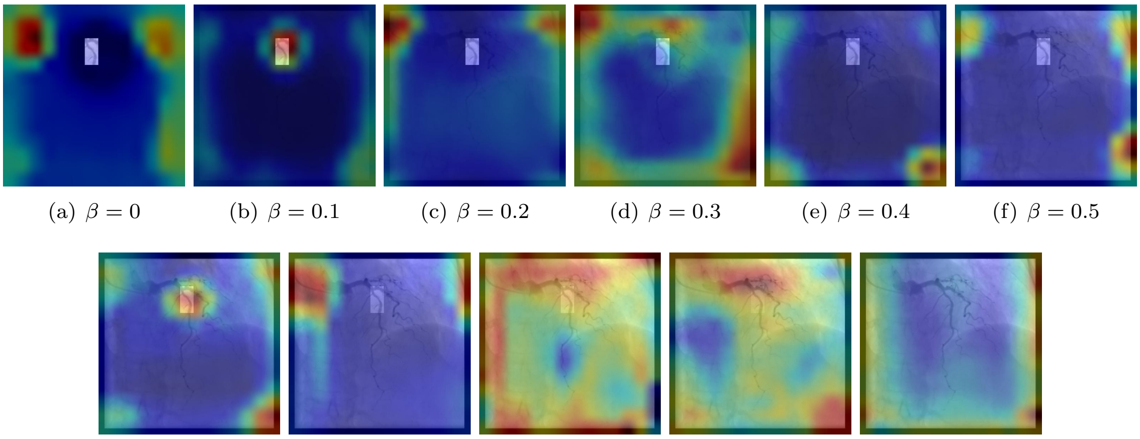
5. Results

5.1. Angle selection performance

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

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.

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) [9] were employed to visualize the degree of contribution of specific image regions. It's possible to observe (see Figure 10) 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.


(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 9: 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.

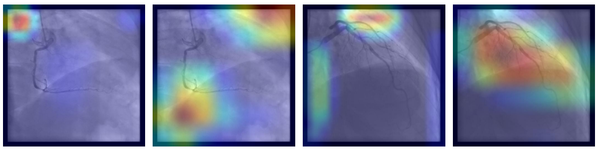
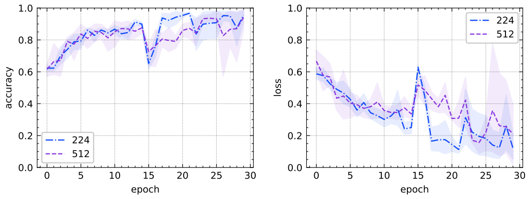

(a) 512 RCA (b) 224 RCA (c) 512 LCA (d) 224 LCA

Figure 10: Grad-CAM visualizations in the RCA and LCA viewing angles showcasing the regions of the frame that most contribute to their correct classification.



(a) Accuracy (b) Cross entropy

Figure 11: 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 image input dimensions variations in training for the models.

Image Dimensions	Accuracy	F1 Score	Cross Entropy
512	0.96±0.01	0.96±0.01	0.14±0.27
224	0.97±0.0	0.97±0.01	0.08±0.31

Table 2: Coronary viewing angles classification metrics performance.

From the observed performance shown in Table 2, 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.

5.2. 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. 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 under and IoU of 0.2. This is a more strict measure of performance. Following Cong et al. [7], *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.

Our baseline (B) is defined as having all frames from the full radio-opaque contrast interval 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.

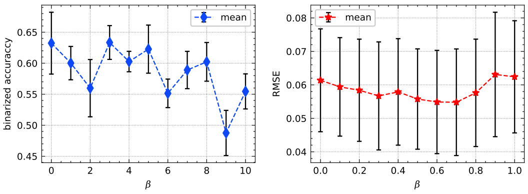
Method	Stenosis Detection Performance @ 0.2 IoU [mean ± std)]			
	Sensitivity	At least One	mAP max 5 det	mAR max 5 det
B	**0.72±0.03**	**0.81±0.01**	**0.61±0.03**	0.82±0.02
B w/ BG	0.64±0.03	0.74±0.03	0.57±0.02	0.82±0.02
B w/ NL	0.68±0.04	0.74±0.04	0.59±0.02	**0.83±0.01**
B w/ BG w/ NL	0.65±0.03	0.71±0.02	0.59±0.02	0.83±0.01
Cong et al. [7]	0.71	-	-	-
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. [7] *	0.60	-	-	-

Table 3: Stenosis detection metrics comparison on reference frames against different authors, (*) Denotes LCA viewing angle model, RCA otherwise

We compare the performance of all our models against themselves and with different authors (see Table 3). From visual validation of the model's performance (see Figure 3), 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. 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.

5.3. iFR regression performance

For the stenosis assessment task, the objective is to estimate its corresponding iFR value. We evaluate the performance under binarized accuracy 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.

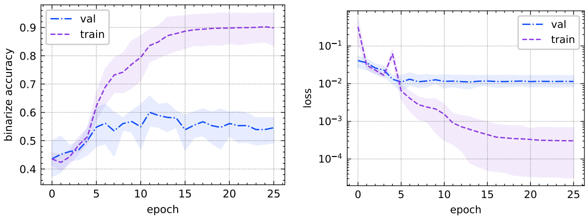


(a) Binarized accuracy

(b) RMSE

Figure 12: 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.

For additional validation, Grad-CAM was again applied. From the performance metrics evolution (see Figure 12), 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 uptrends for the last values (see Figure 12), indicating that the models failed to learn the task. With the Grad-CAM visualizations (see Figure 9) 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

(b) RMSE

Figure 13: 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 13), 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.

6. Conclusions

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

A three stage framework based on convolutional neural networks was assembled to automate stenosis assessment. Starting with viewing angle selection where the objective was to precisely classify reference frames as belonging to the right coronary artery (RCA) or to the left coronary artery (LCA) with previously viewing angles 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 additionally demonstrating the feasibility of frame down-scaling to increase inference time memory optimization.

With the the RCA and LCA viewing angles, two distinct models, based on the single shot detector RetinaNet architecture were assembled as the second stage of the framework to automatically detect stenosis by bounding box placement. Comparisons with different authors confirm the relevance of our work 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 considerably well at stenosis single and multiple detection but leaves room for precision improvement as many background regions were detected as being stenosis.

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

7. Future Work

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

References

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

[2] Gregory A Roth, Degu Abate, Kalkidan Hassen 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. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(18)32203-7. URL http://www.sciencedirect.com/science/article/pii/S0140673618322037.

[3] Karol Antczak and Lukasz Liberadzki. Stenosis detection with deep convolutional neural networks. *MATEC Web of Conferences*, 210: 04001, 01 2018. doi: 10.1051/matecconf/201821004001.

[4] Benjamin Au, Uri Shaham, Sanket Dhruva, Georgios Bouras, Ecaterina Cristea, Alexandra Lansky, Andreas Coppi, Fred Warner, Shu-Xia Li, and Harlan M. Krumholz. Automated characterization of stenosis in invasive coronary angiography images with convolutional neural networks. *CoRR*, abs/1807.10597, 2018. URL http://arxiv.org/abs/1807.10597.

[5] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. URL http://arxiv.org/abs/1506.02640.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

[7] 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. doi: 10.1109/BIBM47256.2019.8983033.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[9] 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. doi: 10.1109/ICCV.2017.74.

[10] Wei Wu, Jingyang Zhang, Hongzhi Xie, Yu Zhao, Shuyang Zhang, and Lixu Gu. Automatic detection of coronary artery stenosis by convolutional neural network with temporal constraint. *Computers in Biology and Medicine*, 118:103657, 2020. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2020.103657. URL http://www.sciencedirect.com/science/article/pii/S0010482520300512.

[11] Antoine Rosset, Luca Spadola, and Osman 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, 10 2004. doi: 10.1007/s10278-004-1014-6.

[12] Alan Lukežič, Tomáš Vojíř, Luka Čehovin Zajc, Jiří Matas, and Matej Kristan.

Discriminative correlation filter tracker with channel and spatial reliability. *International Journal of Computer Vision*, 126(7): 671–688, Jan 2018. ISSN 1573-1405. doi: 10.1007/s11263-017-1061-3. URL `http://dx.doi.org/10.1007/s11263-017-1061-3`.

[13] 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. doi: 10.1109/CVPR.2016.90.

[14] Vinod Nair and Geoffrey 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. ISBN 9781605589077.

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL `http://proceedings.mlr.press/v37/ioffe15.html`.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

[17] 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. doi: 10.1109/ICCV.2017.324.

[18] 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. doi: 10.1109/CVPR.2017.106.

[19] 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. doi: 10.1109/CVPR.2014.81.

[20] R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. doi: 10.1109/ICCV.2015.169.

[21] Thang Luong, Hieu Pham, and Christopher 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, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL `https://www.aclweb.org/anthology/D15-1166`.

[22] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. URL `http://proceedings.mlr.press/v37/xuc15.html`.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia 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. ISBN 9781510860964.