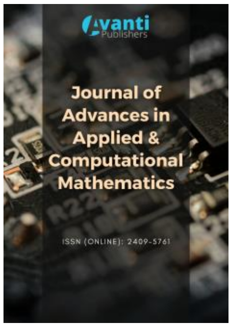
Published by Avanti Publishers

Journal of Advances in Applied & Computational Mathematics

ISSN (online): 2409-5761



Attention Mechanisms Evaluated on Stenosis Detection using X-ray Angiography Images

Emmanuel Ovalle-Magallanes[1,*], Dora E. Alvarado-Carrillo[2], Juan Gabriel Avina-Cervantes[1], Ivan Cruz-Aceves[2], Jose Ruiz-Pinales[1] and Jose Luis Contreras-Hernandez[1]

[1] *Telematics and Digital Signal Processing Research Groups (CAs), Engineering Division, Campus Irapuato-Salamanca, University of Guanajuato, Carretera Salamanca-Valle de Santiago km 3.5 + 1.8km, Comunidad de Palo Blanco, Salamanca 36885, Mexico.*

[2] *Center for Research in Mathematics (CIMAT), A.C. Jalisco S/N, Col. Valenciana, Guanajuato 36000, Mexico.*

ARTICLE INFO

Article Type: Research Article

Keywords:
Medical Imaging
Stenosis Detection
Attention Mechanisms
X-ray Coronary Angiography
Convolutional Neural Networks

Timeline:
Received: January 06, 2022
Accepted: March 07, 2022
Published: May 24, 2022

Citation: Ovalle-Magallanes E, Alvarado-Carrillo DE, Avina-Cervantes JG, Cruz-Aceves I, Ruiz-Pinales J, Contreras-Hernandez JL. Attention Mechanisms Evaluated on Stenosis Detection using X-ray Angiography Images. J Adv App Comput Math. 2022; 9: 62-75.

DOI: https://doi.org/10.15377/2409-5761.2022.09.5

ABSTRACT

Coronary stenosis results from unnatural narrowing of the heart arteries due to the accumulation of adipose depots, leading to different heart diseases and yielding top mortality worldwide. Thus far, deep learning-based methods for automatic stenosis over X-ray Coronary Angiography (XCA) have employed state-of-the-art architectures to solve the ImageNet challenge. With the advance of deep learning, contemporary architectures incorporated a variety of attention mechanisms to improve performance. Therefore, this paper presents a study of three attention mechanisms for stenosis detection in XCA images. Extensive experiments and comparisons over different Residual backbone networks are presented to verify the effectiveness of including such attention modules. An improvement of 4%, 10%, and 10% on the accuracy, recall, and F1-score was achieved using the approach, reaching mean values of 0.8787, 0.8610, and 0.8732, respectively.

*Corresponding Author
Email: e.ovallemagallanes@ugto.mx
Tel: +52 4646479940 ext. 2400

1. Introduction

Convolutional Neural Networks (CNNs) have been applied successfully in the medical imaging domain for different tasks, such as segmentation and classification [1, 2]. However, a CNN requires a sizeable training dataset to perform effectively. One of the main limitations of medical image datasets is that they are time-consuming, requiring professional expertise to label them. For such a reason, the amount of label data is restricted [3,4].

Data augmentation [5] is a widely used method as an effective way to generate new training samples to improve performance. Also, Transfer Learning is another important method [6], where the network is previously trained on a large dataset (data source) instead of training from scratch with the target dataset. Recently, attention modules [7-9] are being developed and included in the network architectures to enhance intermediate feature maps by learning channel and spatial information.

Detecting stenosis plays an essential role in cardiology because Coronary Heart Disease (CHD), including stenosis, is the leading cause of death worldwide [10]. X-ray Coronary Angiography (XCA) remains the gold-standard vascular imaging modality despite other available angiographic techniques (*e.g.,* magnetic resonance imaging, computed tomography, and ultrasound angiography). In comparison, XCA provides the ability of *in vivo* assessment and treatment, such as angioplasty, where a catheter is used to clear a stenotic artery (narrowed or blocked artery) [11].

Fig. **1** shows a representative XCA image where stenosis and non-stenosis areas are highlighted. Nevertheless, manual stenosis detection requires an exhaustive visual examination by the specialist. Besides, automatic stenosis detection is considered a challenge because of corrupted or noisy data (resolution, contrast, signal-to-noise), typically caused by image acquisition [12, 13].

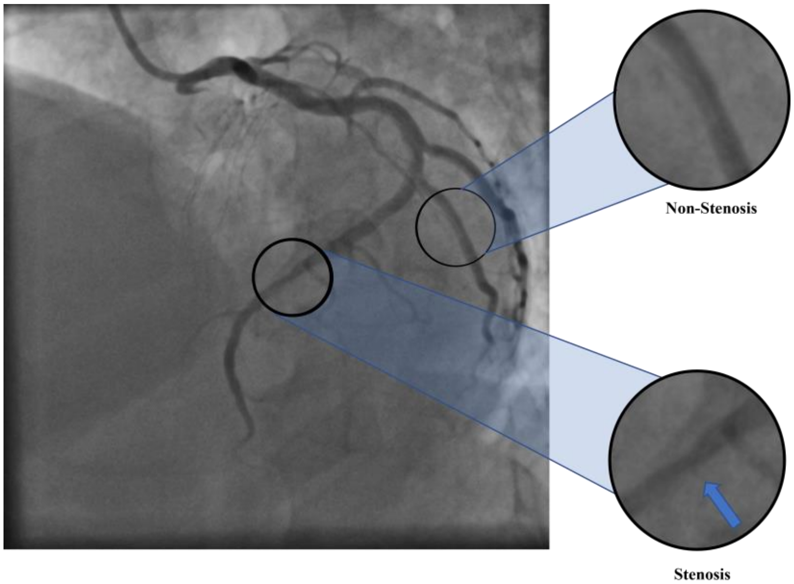


Figure 1: X-ray coronary angiography image within two marked regions corresponding to a non-stenosis and stenosis case. The stenosis case presents a partial artery blockage.

The main contribution of this paper is a comparative evaluation of different state-of-the-art attention modules to improve stenosis detection. This work is the first comparative study of attention modules for stenosis detection on XCA images to the best of our knowledge. Plus, the hyperparameters' selection for such modules is studied and discussed. So, data augmentation is set in place to deal with a limited amount of training data: random horizontal flip, random vertical flip, and discrete rotations.

Experimental results prove the significant impact of including an attention module into the CNN, boosting accuracy up to 4%.

2. Related Work

Recent works attempt to automatically detect coronary artery stenosis in XCA employing Deep Learning models [14-20]. One of the main advantages of these algorithms rely on feature extraction, feature selection, and classification, requiring traditional machine learning-based methods [21-23], which is realized within the optimization step of the same deep architecture.

Zhao *et al.* [14] developed an automatic approach to extract coronary arteries and detect arterial stenosis from a collection of 314 XCA images. First, a UNet++ [15] with Feature Pyramid techniques (FP-UNet) performs segmentation of coronary arteries. Second, centerlines of arterial segments are extracted to measure the diameter of each segment. Finally, according to a stenotic threshold, the segment is classified.

Some approaches take advantage of the availability of the full screening test. For instance, Cong *et al.* [16] put forward a deep-learning-based workflow divided into three steps: 1) artery/angle view choice, 2) candidate frame selection, and 3) stenosis classification localization. These steps employ a pre-trained Inception-v3 [24] with the ImageNet [25] dataset as the backbone network. First, an XCA video is classified into a left or right coronary artery employing the backbone network. Then, candidate frames (with clearly contrasted vessel borders) and redundancy frames (background frames) are chosen. To do so, the fully-connection layer of the Inception-v3 network feeds a bi-directional Long-Short-Term Memory (LSTM). Finally, stenosis classification is performed using the backbone network over the final candidate frames. Redundant frames are employed as additional training data.

Wu *et al.* [17] introduced a coronary artery stenosis deep learning-based object detection network working as follows: first, from the input XCA sequence, only the contrast-filled frames are selected based on U-Net segmentation results. Next, a Deconvolutional Single-Shot multi-box Detector (DSSD) with a Visual Geometry Group (VGG) [26] backbone model provides potential stenosis regions. Finally, a sequence-non-maximum suppression algorithm removes false positives stenosis cases.

Similarly, Pang *et al.* [18] disposed of the complete video to propose the Stenosis-DetNet, an end-to-end network for stenosis detection. The method consists of a Sequence Feature extraction and Fusion (SFF) module and a Sequence Consistency Alignment (SCA) module. First, the SFF fuses candidate object bounding box and feature maps obtained by a ResNet50 [27] from the continuous frames of the XCA test. Second, the SCA module optimizes the initial bounding boxes by prior coronary artery displacement information and adjacent image features. This two-step workflow improves the final detection.

These methods exploit the availability of the complete angiography image sequence acquired during the screening test. However, the last step detects stenosis from candidates' bounding boxes seen as an image patch classifier in the previous two approaches. Antczak and Liberadzki [19] introduce a shallow patch-based CNN only five-layer deep for small XCA image patches stenosis classification. Additionally, to deal with the limited training dataset and improve classification performance, the network is pre-trained by synthetic images where vessels are modeled as Bezier curves. Ovalle *et al.* [20] evaluate three CNN pre-trained on the ImageNet dataset (VGG16, ResNet50, and Inception-v3) following a patch-based classification. Hence, the number of layers needed to fine-tune is determined by an exhaustive search. Additionally, it was proven that the last convolutional layers can be cut from the network. Therefore, the training time and the number of parameters are reduced without compromising the detection rates.

3. Methods

Recently, different state-of-the-art networks for natural image classification have leveraged attention mechanisms to improve network performance [7-9]. The attention mechanisms are focused on refining the feature maps learning channel attention or spatial attention relationships of features. This section analyzes three attention modules that will be evaluated in XCA images for stenosis detection.

3.1. Squeeze-and-Excitation Attention

Squeeze-and-Excitation Networks (SENets) [7] investigate the relationships between channels, learn global information weighting the relevance of the features, emphasize discriminant features and remove less useful ones.

SENets have a channel attention mechanism consisting of two components: a Squeeze Block and an Excitation Block. Given the input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, the Squeeze Block collects global information for each channel x_c, with c indicating the channel index in the input map, *i.e.,* $c \in \mathbb{N}, c \leq C$, is expressed as follows

$$z_c = g(\mathbf{X}) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j), \tag{1}$$

where z_c is the channel-wise Global Average Pooling (GAP) associated with the c-th input feature channel. As such, a statistic $\boldsymbol{z} \in \mathbb{R}^{C\times 1\times 1}$ is generated, noticing that the number of channels remains the same.

The Excitation block is a Multi-Layer Perceptron (MLP) bottleneck responsible for capturing channel-wise dependencies, *i.e.,* learning the channel attention weights. The MLP comprises two Fully-Connected (FC) layers around the non-linearity acting as a dimensionality-reduction layer with a reduction ratio r; a dimensionality-increasing layer returns to the original channel dimension. Formally, the Squeeze-Excitation (SE) block takes the form:

$$\boldsymbol{\omega} = \sigma(f(\boldsymbol{w}_1, \boldsymbol{w}_2)(\boldsymbol{z})) \tag{2}$$

$$= \sigma(\boldsymbol{W}_2\delta(\boldsymbol{W}_1\boldsymbol{z})), \tag{3}$$

where σ refers to sigmoid and δ to the ReLU activation function, $\boldsymbol{W}_1 \in \mathbb{R}^{C\times(\frac{C}{r})}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{(\frac{C}{r})\times C}$. Therefore, the total parameters of the SE attention module are $\frac{2C^2}{r}$.

Fig. **2** summarizes the SE module procedure from the given intermediate feature maps, generally obtained by a previous convolutional layer. Finally, the refined feature maps are obtained by

$$\widehat{\mathbf{X}} = \mathbf{X} \otimes \boldsymbol{\omega}, \tag{4}$$

where $\otimes$ denotes channel-wise multiplication.

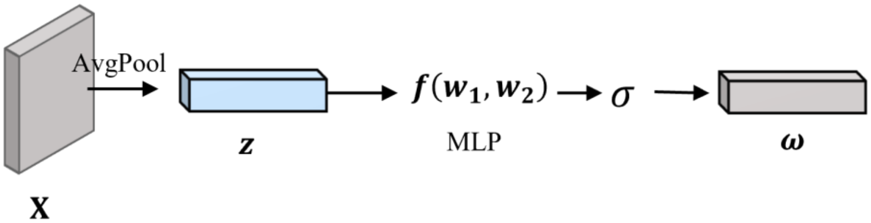


Figure 2: Squeeze-and-Excitation Attention module. Global information for each channel is obtained by the Squeeze Block (AvgPool), then the Excitation Block is an MLP capturing the channel-wise dependencies.

3.2. Efficient Channel Attention

Efficient Channel Attention Networks (ECANet) [9] introduced an attention mechanism based on SE blocks without dimensionality reduction. By a local cross-channel interaction strategy, each channel from a given input feature map $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$ has interdependence on every other channel in a small local group. Thus, once computed the feature vector $\boldsymbol{z} \in \mathbb{R}^{C\times 1\times 1}$ (GAP outcome), the local cross-channel interaction-based attention can be written as

$$\omega_c \; = \; \sigma\left(\sum_{j=1}^{k} w_c^j z_c^j\right), z_c^j \in \Omega_c^k \,, \tag{5}$$

where Ω_c^k is the set of k adjacent channels of the query channel w_c, *i.e.,* a neighborhood of size k of the c-th channel and σ stands for the sigmoid activation function.

Furthermore, if the channels share the same learning weights w_c, the parameters can be reduced from $C \cdot k$ to k. Thus, the above equation can be re-written as follows

$$\boldsymbol{\omega}_c \; = \; \sigma\left(\sum_{j=1}^{k} w^j z_c^j\right), z_c^j \in \Omega_c^k \,. \tag{6}$$

This local interaction can be seen as a 1D convolution (Conv1D) with kernel size k. In such a way, the Efficient Channel Attention (ECA) can be defined as:

$$\boldsymbol{\omega} = \; \sigma(Conv1D(k)(\mathbf{z})). \tag{7}$$

The size of the kernel is obtained adaptive and proportional to the number of channels C as follows,

$$k \; = \; \left|\frac{log2(C)}{\gamma} + \frac{\beta}{\gamma}\right|_{odd} \,, \tag{8}$$

here $|\cdot|_{\text{odd}}$ indicates the nearest odd number, $\beta = 1$ and $\gamma = 2$. Fig. **3** shows the ECA module, where one can clearly see that a 1D convolution layer now obtains the channel-wise dependencies.

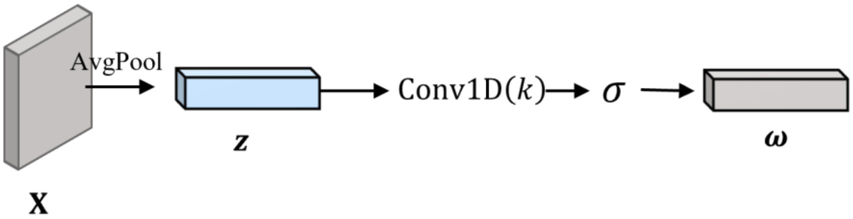


Figure 3: Efficient Channel Attention module. After the global Average Pooling (AvgPool) infer global information, the local interaction between channels is acquired by a 1D convolution of adaptive kernel size k.

3.3. Convolutional Block Attention

Convolutional Block Attention Module (CBAM) [8] contains two sequential sub-modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). From a given input feature map $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$, the CBAM mechanism infers attention maps along the channel and spatial dimensions to generate a refined feature map $\widehat{\boldsymbol{X}} \in \mathbb{R}^{C\times H\times W}$.

First, the CAM is based on the SE module; wherein besides using the global average-pooled features, a max-pooling operation is carried out to generate two different spatial feature vectors: $\boldsymbol{z^c}_{\text{avg}}, \boldsymbol{z^c}_{\text{max}} \in \mathbb{R}^{C\times 1\times 1}$, respectively. Both feature vectors are forwarded through an MLP sharing the weights with each input. This shared network comprises dimensionality-reduction and dimensionality-increasing layers, with a ReLU activation function between them. Hence, the channel attention map $\boldsymbol{\omega^c} \in \mathbb{R}^{C\times 1\times 1}$ is computed as

$$\boldsymbol{\omega^c} = \sigma\big(f(\boldsymbol{w}_1, \boldsymbol{w}_2)(\boldsymbol{z^c}_{\text{avg}}) + f(\boldsymbol{w}_1, \boldsymbol{w}_2)(\boldsymbol{z^c}_{\text{max}})\big) \tag{9}$$

$$= \sigma\big(\boldsymbol{W}_2\delta(\boldsymbol{W}_1\boldsymbol{z}^c{}_{\text{avg}}) + \boldsymbol{W}_2\delta(\boldsymbol{W}_1\boldsymbol{z}^c{}_{\text{max}})\big). \tag{10}$$

It is noteworthy that the number of parameters of the CAM is the same as the SE attention module with $\frac{2C^2}{r}$, where r is the feature reduction ratio involving the MLP. Fig. **4** illustrates the CAM procedure. At this point, an intermediate refined feature map $\boldsymbol{X}'$ is obtained by

$$\mathbf{X}' = \boldsymbol{\omega}^c \otimes \mathbf{X}, \tag{11}$$

where $\otimes$ denotes element-wise multiplication.

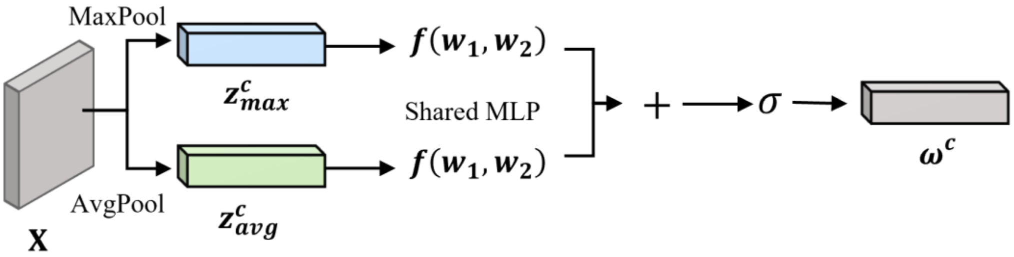


Figure 4: The Channel Attention sub-module includes two pooling layers and a shared MLP to exploit the inter-channel relationship of features from a given input $\boldsymbol{X}$.

Secondly, the SAM is generated by using the refined intermediate features as input. In order to obtain the spatial attention feature map, two pooling operations along the channel axis were applied: an average-pooling and max-pooling generating two 2D maps $\boldsymbol{z}^s{}_{\text{avg}}, \boldsymbol{z}^s{}_{\text{max}} \in \mathbb{R}^{1\times H\times W}$, respectively. Subsequently, those feature maps are concatenated and convolved by a 2D convolution layer (Conv2D). Therefore, the spatial attention is computed as

$$\boldsymbol{\omega}^s = \sigma(\text{Conv2D}(k)(\boldsymbol{z}^s{}_{\text{avg}};\ \boldsymbol{z}^s{}_{\text{max}})), \tag{12}$$

where k is the filter size, set as $k = 7 \times 7$ by default.

Fig. **5** depicts the computation process of this sub-module, requiring $2k^2$ parameters.

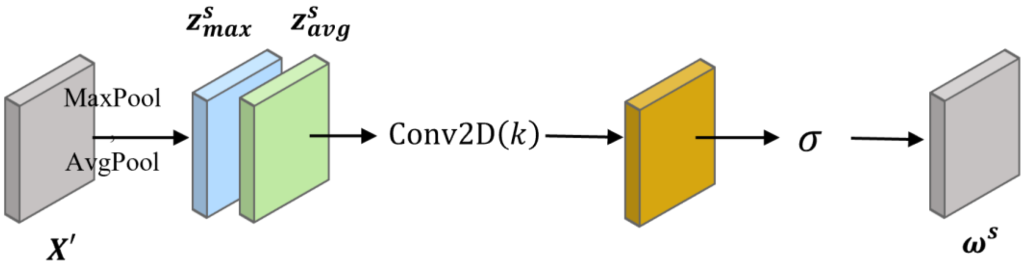


Figure 5: Spatial Attention sub-module. Similarly, two pooling operations are required. Their outputs are concatenated and sent to a 2D convolutional layer.

Finally, the overall CBAM process is expressed as an element-wise multiplication between the intermediate feature map and the spatial attention map,

$$\widehat{\mathbf{X}} = \boldsymbol{\omega}^s \otimes \mathbf{X}'. \tag{13}$$

Similar to the previous attention mechanisms, the output refined feature map shares the exact dimensions concerning the input feature map, *i.e.*, $\widehat{\boldsymbol{X}} \in \mathbb{R}^{C\times H\times W}$.

It is essential to point out that the attention modules mentioned above can be easily plugged into a CNN architecture. More commonly, the attention module is included after a convolutional or residual block (before subsampling) to improve the feature relevance at each scale level.

4. Results

This work conducted extensive experiments to evaluate the relevance of including an attention module into a ResNet backbone network for stenosis detection in a limited XCA dataset. The analysis includes the effect of the reduction ratio and the choice of the ResNet depth (*i.e.,* 18, 50, and 101).

4.1. Dataset

The three attention modules were evaluated with the public stenosis dataset presented by Antczak and Liberadzki [19]. This dataset consists of 250 image patches produced from real XCA images acquired from internal and internet sources. The data are labeled as stenosis and non-stenosis, containing 125 images for each class. The dataset is split in a stratified manner into three subsets: training (100 images), validation (25 images), and testing (125 images). This amount of data is really limited for a robust CNN training process. For this reason, a combination of random horizontal and vertical flip and discrete rotation as data augmentation is conducted during the training procedure. These particular transformations do not modify the geometrical structure of the images (*i.e.,* vessels), being crucial not to create stenosis structures where is a negative case or vice-versa.

Fig. **6** shows XCA real samples (first column), followed by the output images generated by the data augmentation procedure.

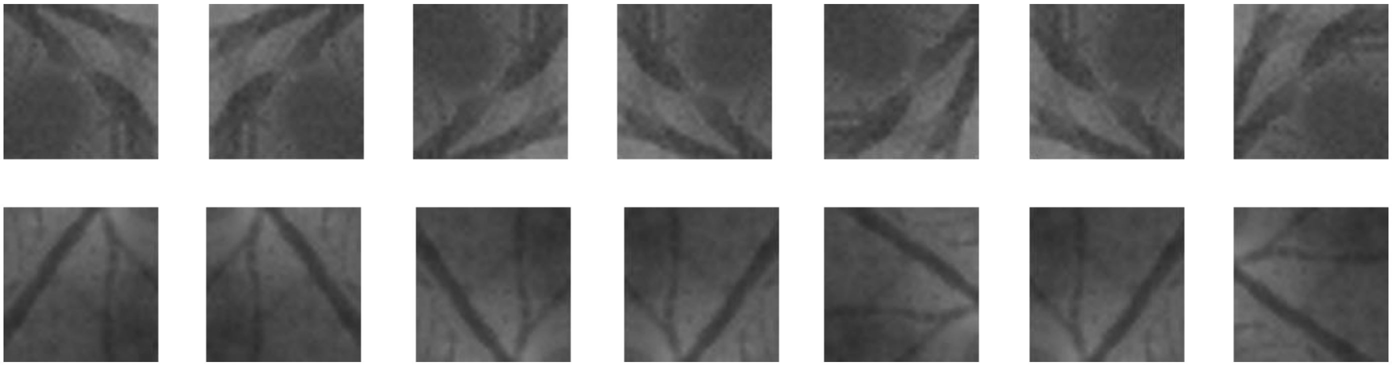


Figure 6: Two examples of XCA images and their corresponding data-augmentation-generated images (Top: a stenosis case, bottom: a non-stenosis case). For each row, the first image corresponds to the original image.

4.2. Implementation Details

Three different backbone networks (part of the family of the ResNet) were evaluated: the ResNet18, ResNet50, and ResNet101. To plug the different attention modules and avoid the input for the last residual block from being minimal (a feature map in $\mathbb{R}^{C\times1\times1}$), two changes are made on the ResNets. First, the last residual block is ignored, *i.e.,* changed by an Identity block, reducing the network depth, therefore, the number of parameters. Secondly, the kernel size of the first convolutional layer was set to 3×3 instead of 7×7. Besides, the stride was changed from 2 to 1 to avoid subsampling.

Table **1** shows the backbone architectures evaluated and their configuration characteristics, including the kernel size, number of filters, and output of the layers. Notice that the attention module included in each residual block does not change the layer output size. Layer removal has been employed to improve accuracy and speed the inference up in pre-trained networks used in transfer learning and pre-defined architectures employed for the ImageNet challenge [20, 28]. Such an approach is based on the intuition that features at the bottom of a network extract high-level features (*i.e.,* edges) [6].

Conversely, the top layers (last layers in the models) extract more abstract features being not necessarily useful enough when are extracted from small feature maps. For instance, the convolution kernel is bigger than the input feature map. Therefore, it makes the last layers more suitable for removal or change, *i.e.,* for an identity layer.

Table 1: Backbone architectures. Building blocks are shown in the matrix notation, with the numbers of stacked blocks. The input sample size is a 32×32 image patch.

Layer	Output Size	18-Layer	50-Layer	101-Layer
Conv1	32×32	3×3, Stride 1		
Pool1	16×16	3×3 Max-pooling, Stride 2		
Residual1	8×8	$\begin{bmatrix} 3\times3, & 64 \\ 3\times3, & 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix} \times 3$
Residual2	4×4	$\begin{bmatrix} 3\times3, & 128 \\ 3\times3, & 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix} \times 4$
Residual3	2×2	$\begin{bmatrix} 3\times3, & 256 \\ 3\times3, & 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix} \times 23$
Identity1	2×2	–	–	–
GAP	–	Global Average Pooling		
Class	2	SoftMax		

The training process employs the Adam [29] optimizer with an initial learning rate of 1×10^{-3}. The learning rate follows a cosine annealing schedule [30] with a restart period of 10 epochs, a minimum learning rate of 1×10^{-4}, and a batch size of 4. A total of one hundred epochs were executed, saving the best weights of the model only if the validation loss manifested substantial improvement. Furthermore, the experiments were implemented employing the Pytorch framework and the Timm library [31] and running on Google's cloud servers, including a Tesla P4 GPU.

4.3. Evaluation Metrics

All the network configurations were evaluated under different metrics based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN) measures. The considered metrics are Accuracy (Acc), Precision (Pr), Recall (Rec), F-score (Fβ), and Specificity (Sp), defined as follows

$$\text{Acc} = \frac{TP + TN}{TP + FP + TN + FN}, \tag{14}$$

$$\text{Pr} = \frac{TP}{TP + FP}, \tag{15}$$

$$\text{Rec} = \frac{TP}{TP + TN}, \tag{16}$$

$$F_\beta = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2(FN + FP)}, \tag{17}$$

$$\text{Sp} = \frac{TN}{TN + FP}. \tag{18}$$

A particular factor, $\beta = 1$, was chosen for the F-score, given the harmonic mean of the precision and recall known as F1-score.

4.4. Stenosis Detection Results

Experiments are conducted with three different backbone residual networks and three attention modules. Each network was trained using the same hyperparameters. Besides, the experiments were repeated five times, using different seeds for the weights' initialization.

Table **2** shows the results for the ResNet18, where the best accuracy, recall, and F1-score are obtained within an SE attention module with a reduction ratio of $r = 8$.

Table 2: Performance comparison on ResNet18 as backbone network.

Attention	r	Accuracy	Precision	Recall	F1-score	Specificity
N/A	N/A	0.8361±0.046	0.8903±0.0592	0.7627±0.1216	0.8140±0.0684	0.9048±0.0561
SE	1	0.8672±0.0359	**0.8963±0.0123**	0.8203±0.0809	0.8550±0.0457	**0.9111±0.0142**
	2	0.8738±0.0299	0.8955±0.0173	0.8373±0.0735	0.8639±0.0383	0.9079±0.0207
	4	0.8492±0.0396	0.8630±0.0513	0.8203±0.0594	0.8400±0.0439	0.8762±0.0531
	8	**0.8787±0.0280**	0.8873±0.0474	**0.8610±0.0303**	**0.8732±0.0269**	0.8952±0.0522
	16	0.8525±0.0353	0.8488±0.0392	0.8508±0.1020	0.8457±0.0476	0.8540±0.0543
CBAM	1	0.8721±0.0408	0.8784±0.0455	0.8576±0.0877	0.8650±0.0480	0.8857±0.0543
	2	0.8429±0.0286	0.8467±0.0272	0.8271±0.0939	0.8336±0.0402	0.8571±0.0435
	4	0.8246±1.0197	0.8762±0.0246	0.7424±0.0251	0.8036±0.0225	0.9016±0.0207
	8	0.8344±0.0340	0.8688±0.0444	0.7763±0.0528	0.8190±0.0384	0.8889±0.0435
	16	0.8508±0.0135	0.8593±0.0402	0.8305±.0339	0.8435±0.0108	0.8698±0.0481
ECA	N/A	0.8607±0.0317	0.871±0.0568	0.8407±0.0517	0.8538±0.0327	0.8794±0.0631

Table 3: Performance comparison on ResNet50 as a backbone network.

Attention	r	Accuracy	Precision	Recall	F1-score	Specificity
N/A	N/A	0.8344±0.0303	**0.9106±0.0500**	0.7322±0.0639	0.8096±0.0390	**0.9302±0.0414**
SE	1	0.8574±0.0310	0.8760±0.0402	0.8237±0.0663	0.8474±0.0359	0.8889±0.0405
	2	0.8426±0.0354	0.8928±0.0312	0.7661±0.0554	0.8242±0.0434	0.9143±0.0241
	4	0.8361±0.0192	0.8955±0.0206	0.7492±0.0470	0.8149±0.0269	0.9175±0.0207
	8	0.8525±0.0284	0.8758±0.0200	0.8102±0.0627	0.8406±0.0353	0.8921±0.0207
	16	0.8361±0.0441	0.8799±0.0525	0.7661±0.0514	0.8187±0.0488	0.9016±0.0455
CBAM	1	0.8361±0.0471	0.8654±0.0495	0.7831±0.0661	0.8214±0.0530	0.8857±0.0440
	2	0.8295±0.0187	0.8594±0.0584	0.7831±0.0682	0.8158±0.0210	0.8730±0.0753
	4	0.8574±0.0250	0.8863±0.0322	0.8102±0.0470	0.8457±0.0297	0.9016±0.0344
	8	0.8508±0.0212	0.8595±0.0313	0.8102±0.0733	0.8388±0.0310	0.8889±0.0337
	16	**0.8689±0.0301**	0.8910±0.0213	**0.8305±0.0623**	**0.8588±0.0368**	0.9048±0.0194
ECA	N/A	0.8639±0.0351	0.8895±0.0172	0.8203±0.0735	0.8523±0.0435	0.9048±0.0159

Moreover, with this attention module, the best precision and specificity are achieved, but with a reduction ratio of $r = 1$ (*i.e.,* with no reduction). Incorporating an attention module into the ResNet18 boosts up to four points in the accuracy concerning the model without attention.

Table **3** compares the ResNet50 performance; the best accuracy, recall, and F1-score are reached when the CBAM (with a reduction ratio of $r = 16$) is plugged in the network.

Table **4** shows a deeper ResNet, where the ECA module obtains the best accuracy, recall, and F1-score. The CBAM (with $r = 8$) has the best precision and specificity. Notice that incorporating the attention module into the ResNet18 and the ResNet101 improves the values of five evaluation metrics. Contrarily, in the ResNet50, the precision and recall were not surpassed by plugging the attention module into the network. The numerical result shows that there is no need to train a very deep network to get high performance in stenosis detection, where three of the best overall metrics are achieved by the ResNet18 + SE with $r = 8$.

Table 4: Performance comparison on ResNet101 as a backbone network.

Attention	r	Accuracy	Precision	Recall	F1-score	Specificity
N/A	N/A	0.8344±0.0340	0.8620±0.0423	0.7864±0.0853	0.8196±0.0439	0.8794±0.0429
SE	1	0.8509±0.0554	0.8804±0.0372	0.8068±0.1316	0.8424±0.0753	0.9079±0.0344
	2	0.8328±0.0250	0.8633±0.0539	0.7864±0.0978	0.8176±0.0409	0.8762±0.0658
	4	0.8557±0.0160	0.8881±0.0382	0.8068±0.0652	0.8431±0.0234	0.9016±0.0440
	8	0.8230±0.0539	0.8527±0.0667	0.7695±0.0663	0.8077±0.0569	0.8730±0.0673
	16	0.8639±0.0293	0.9062±0.0205	0.8034±0.0827	0.8492±0.0399	0.9206±0.0251
CBAM	1	0.8459±0.0212	0.8576±0.0443	0.8203±0.0351	0.8375±0.0205	0.8698±0.0481
	2	0.8525±0.0239	0.8535±0.0322	0.8407±0.0458	0.8462±0.0258	0.8635±0.0398
	4	0.8574±0.0321	0.8895±0.0364	0.8068±0.0663	0.8446±0.0388	0.9048±0.0355
	8	0.8721±0.0197	**0.9202±0.0310**	0.8068±0.0425	0.8589±0.0239	**0.9333±0.0284**
	16	0.8279±0.0430	0.8646±0.0550	0.7661±0.0723	0.8106±0.0524	0.8857±0.0531
ECA	N/A	**0.8754±0.0354**	0.9075±0.0336	**0.8305±0.1024**	**0.8660±0.0522**	0.9175±0.0379

The best precision and specificity are obtained without any attention module.

Additionally, Table **5** shows a comparison study with other proposed deep learning architectures employing the same dataset. Only the models trained from scratch are included to make a fair comparison. The ResNet18 + SE using $r = 8$ surpassed four over five metrics concerning the analyzed baseline models.

Table 5: Performance comparison concerning models using the same dataset and training/testing partitions. The results are taken as presented in the original paper [20].

Approach	Backbone	Accuracy	Precision	Recall	F1-score	Specificity
Antczak and Liberadzki [19]	Plain CNN	0.59	0.55	0.87	0.68	0.33
Ovalle-Magallanes *et al.* [20]	VGG16	0.50	0.50	**1.00**	0.67	0.00
	ResNet50	0.82	0.77	0.94	0.84	0.71
	Inception-v3	0.71	0.70	0.75	0.72	0.68
Best Proposed	ResNet18	**0.8787**	**0.8873**	0.8610	**0.8732**	**0.8952**

The Gradient-weighted Class Activation Map (Grad-CAM) [32] is a visualization method based on gradients designed to examine class-discriminative regions (*i.e.,* a heat-map) learned throughout the network. So, the Grad-CAM for each class c is generated as follows

$$L^c = \delta\left(\sum_k \alpha_k^c A_k\right), \tag{19}$$

where $A_k \in R^{(u \times v)}$ is the output k-feature map, generally the last convolutional layer, α_k^c is the weight concerning the k-th feature map with the c-th class, and δ is the ReLU activation function. Hence, this weight α_k^c is computed by

$$\alpha_k^c = g\left(\frac{\partial y^c}{\partial A_k}\right), \tag{20}$$

where $g(\cdot)$is the global average pooling operation, and y^c is the score for class c before the SoftMax.

For challenging test images, one non-stenosis, and one stenosis case, the Grad-CAM images are shown in Fig. **7**.

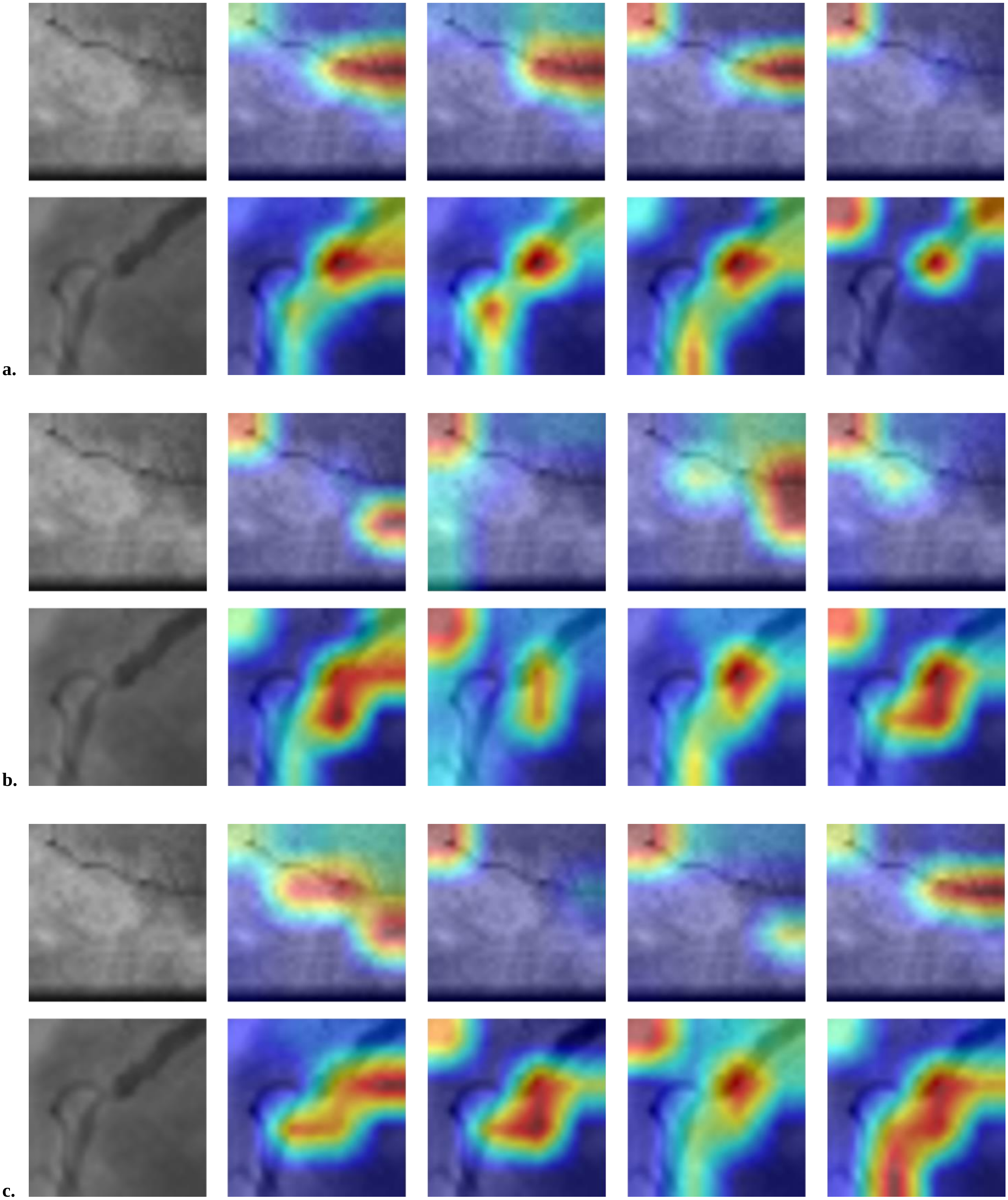


Figure 7: Grad-CAM feature visualization for negative and positive stenosis cases, respectively. Left to right, input image, backbone ResNet((a) 18, (b) 50, (c) 101), SE, CBAM, and ECA outputs.

The figure illustrates high discriminative regions for stenosis detection in hot tones (red colors) and otherwise cold tones (purple colors). In the first case, with a ResNet18 as the backbone, when the SE attention module is employed, the more discriminant regions correspond to almost all vessel pixels and areas with a particular grade of stenosis.

For the ResNet50 backbone, the artery vessel pixels present a better visual delineation when the CBAM module is incorporated. Finally, in the case of the ResNet101, the ECA network configuration set greater attention over the artery tree than the other networks variants. Notice that this visual examination agrees with the obtained numerical results, showing that the network focuses on some distinctive regions of the image, including an attention mechanism.

5. Discussion

This work examined three attention mechanisms for stenosis detection in XCA images. In particular, while discussing the classification performance for the networks, numerical results suggest that ResNet50 and ResNet101 are unnecessarily large and more susceptible to overfitting than the ResNet18. This phenomenon can also be observed in very deep networks where attention modules are not included. A similar situation can be seen with the ResNet + SE and the ResNet + CBAM. The ResNet101 model is unnecessarily large. Only a slight improvement of going deeper was found in the ResNet + ECA. However, the ResNet18 + SE outperforms them with fewer parameters. As such, deeper ResNets learn irrelevant or repetitive features in the learning procedure for the two-class stenosis detection task, which in turn causes reductions in the classification performance. Therefore, attention modules do not contribute to the learning processes when they are incorporated into deeper architectures. Moreover, the ResNet architectures lack dropout layers which can be considered as one of the reasons for accuracy degradation.

6. Conclusions

This paper evaluates three state-of-the-art attention mechanisms and their impact on stenosis detection in X-ray coronary angiography images. As the backbone network for the detection task, the ResNet employed three deep variants with a depth of 18, 50, and 101. The method achieves a maximum accuracy, recall, and F1-score when a ResNet18 with a SE attention module is used. This network configuration significantly boosts these metrics by 4%, 10%, and 6%, respectively, concerning the basic ResNet18 (without attention). However, the SE module requires $\frac{2C^2}{r}$ additional parameters per block. In this case, a reduction ratio of $r = 8$ obtains the best results. Numerical and visual results (obtained by Grad-CAM) agree that incorporating an attention mechanism allows the network to focus on more rich-feature areas (*i.e.,* vessel pixel) to improve the performance.

Additionally, attention modules allow the generation of fine-grained attention maps, which are exploited for lesion localization. These maps enable explainable deep learning-based methods, vital for medical imaging analysis and diagnosis support. Moreover, the current approach can classify a full-size XCA image by dividing the image into patches and classifying them individually.

Future work will focus on lighted attention modules suitable for stenosis detection and localization in a full-size XCA image, with fewer parameters without compromising the network performance.

Acknowledgments

This work was supported in part by the University of Guanajuato Grant NUA 147347, Convocatoria Institucional de Investigación Científica – CIIC 2022 (DAIP), and in part by the Mexican Council of Science and Technology CONACyT under Grant No. 626154/755609 and 626155/719327, and by the Mexican National Council of Science and Technology under project Cátedras-CONACyT No. 3150-3097.

Conflicts of Interest

The authors declare that they have no conflict of interest. The funders had no role in the study's design; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

[1] Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. Evolutionary intelligence, pages 2021; 1-22. https://doi.org/10.1007/s12065-020-00540-3.

[2] Mohapatra S, Swarnkar, Das J. Deep convolutional neural network in medical image processing. In Handbook of deep learning in biomedical engineering. pages 25-60. Elsevier, 2021. https://doi.org/10.1016/B978-0-12-823014-5.00006-5.

[3] Althnian A, AlSaeed D, Al-Baity H, Samha A, Bin Dris A, Alzakari N, *et al*. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. Applied Sciences, 2021; 11(2): 796. https://doi.org/10.3390/app11020796.

[4] Hammernik K, Schlemper J, Qin C, Duan J, Summers RM, Rueckert D. Systematic evaluation of iterative deep neural networks for fast parallel MRI reconstruction with sensitivity-weighted coil combination. Magnetic Resonance in Medicine, 2021; 86(4): 1859-1872. https://doi.org/10.1002/mrm.28827.

[5] Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 2019; 6(1): 1-48. https://doi.org/10.1186/s40537-019-0197-0.

[6] Yosinski J, Clune J, Bengio Y, Lipson H. How Transferable Are Features in Deep Neural Networks? In Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014; 2(14): 3320-3328. Montreal, Canada, dec MIT Press.

[7] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018; 7132-7141. Salt Lake City, UT, USA, Jun https://doi.org/10.1109/CVPR.2018.00745.

[8] Woo S, Park J, Lee J-Y, Kweon IS. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, 2018; 3-19, Munich, Germany, https://doi.org/10.1007/978-3-030-01234-2 1.

[9] Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020; 11531-11539, Seattle, WA, USA, Jun https://doi.org/10.1109/CVPR42600.2020.01155.

[10] World Health Organization. Cardiovascular Diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/ cardiovascular-diseases-(cvds), Nov 2021.

[11] Johal GS, Goel S, Kini A. Coronary Anatomy and Angiography. In Practical Manual of Interventional Cardiology, 2021; 35-49.

[12] Manson EN, Ampoh VA, Fiagbedzi E, Amuasi JH, Flether JJ, Schandorf C. Image noise in radiography and tomography: Causes, effects and reduction techniques. Current Trends in Clinical & Medical Imaging, 2019; 2(5): 555620. https://doi.org/10.19080/CTCMI.2019.03.555620.

[13] Chang C-F, Chang K-H, Lai C-H, Lin T-H, Liu T-J, Lee W-L, *et al*. Clinical outcomes of coronary artery bifurcation disease patients underwent Culotte two-stent technique: a single center experience. BMC Cardiovascular Disorders, 2019; 19(1): 1-8. https://doi.org/10.1186/s12872-019-1192-2.

[14] Zhao C, Vij A, Malhotra S, Tang J, Tang H, Pienta D, *et al*. Automatic extraction and stenosis evaluation of coronary arteries in invasive coronary angiograms. Computers in Biology and Medicine, 2021; 136: 104667. https://doi.org/10.1016/j.compbiomed.2021.104667.

[15] Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018; 3-11. https://doi.org/10.1007/978-3-030-00889-5 1.

[16] Cong C, Kato Y, Vasconcellos HD, Lima J, Venkatesh B. Automated Stenosis Detection and Classification in X-ray Angiography Using Deep Neural Network. In International Conference on Bioinformatics and Biomedicine (BIBM), 2019; 1301-1308. San Diego, CA, USA, IEEE. https://doi.org/10.1109/BIBM47256.2019.8983033.

[17] Wu W, Zhang J, Xie H, Zhao Y, Zhang S, Gu L. Automatic detection of coronary artery stenosis by convolutional neural network with temporal constraint. Computers in Biology and Medicine, 2020; 118: 103657. https://doi.org/10.1016/j.compbiomed.2020.103657.

[18] Pang K, Ai D, Fang H, Fan J, Song H, Yang J. Stenosis-DetNet: Sequence consistency-based stenosis detection for X-ray coronary angiography. Computerized Medical Imaging and Graphics, 2021; 89: 101900. https://doi.org/10.1016/j.compmedimag.2021.101900.

[19] Antczak K, Liberadzki L. Stenosis Detection with Deep Convolutional Neural Networks. In MATEC Web of Conferences, 2018; 210: 04001. EDP Sciences, https://doi.org/10.1051/matecconf/201821004001.

[20] Ovalle-Magallanes EI, Avina-Cervantes JG, Cruz-Aceves I, Ruiz-Pinales J. Transfer Learning for Stenosis Detection in X-ray Coronary Angiography. Mathematics, 2020; 8(9): 1510. https://doi.org/10.3390/math8091510.

[21] Sameh S, AbdelAzim M, AbdelRaouf A. Narrowed coronary artery detection and classification using angiographic scans. In 2017 12th International Conference on Computer Engineering and Systems (ICCES), 2017; 73-79. Cairo, Egypt, IEEE. https://doi.org/10.1109/ICCES.2017.8275280.

[22] Kishore AHN, Jayanthi VE. Automatic stenosis grading system for diagnosing coronary artery disease using coronary angiogram. International Journal of Biomedical Engineering and Technology, 2019; 31(3): 260-277. https://doi.org/10.1504/IJBET.2019.102974.

[23] Wan T, Feng H, Tong C, Li D, Qin Z. Automated Identification and Grading of Coronary Artery Stenoses with X-ray Angiography. Computer Methods and Programs in Biomedicine, 2018; 167: 13-22. https://doi.org/10.1016/j.cmpb.2018.10.013.

[24] Szegedy C, Liu W, Jia Y, Sermanet P, Reed SE, Anguelov D, *et al*. Going deeper with convolutions. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015; 1-9. IEEE Computer Society, 2015. https://doi.org/10.1109/CVPR.2015.7298594.

[25] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 2012; 25: 1097-1105. https://doi.org/10.1145/3065386.

[26] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 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; 1-14, URL http://arxiv.org/abs/1409.1556.

[27] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016; 770-778. IEEE Computer Society, 2016. https://doi.org/10.1109/CVPR.2016.90.

[28] Zandigohar M, Erdogmus D, Schirner G. NetCut: Real-Time DNN Inference Using Layer Removal. In 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2021; 1845-1850, Grenoble, France, IEEE. https://doi.org/10.23919/DATE51398.2021.9474052.

[29] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014; https://arxiv.org/ abs/1412.6980.

[30] Loshchilov I, Hutter F. SGDR: Stochastic Gradient Descent with Warm Restarts. arXiv preprint arXiv:1608.03983, 2016; https://arxiv.org/ abs/1608.03983.

[31] Wightman R. PyTorch Image Models. https://github.com/rwightman/ pytorch-image-models, 2019.

[32] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In IEEE International Conference on Computer Vision (ICCV), 2017; 618-626, Venecia, Italia, oct 2017. IEEE Computer Society. https://doi.org/10.1109/ICCV.2017.74