Computers in Biology and Medicine 153 (2023) 106484

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A token-mixer architecture for CAD-RADS classification of coronary stenosis on multiplanar reconstruction CT images

Marco Penso [a,b,*], Sara Moccia [c], Enrico G. Caiani [b,d], Gloria Caredda [a], Maria Luisa Lampus [a], Maria Ludovica Carerj [a,e], Mario Babbaro [f], Mauro Pepi [a], Mattia Chiesa [a,b], Gianluca Pontone [a]

[a] Cardiovascular Imaging Department, Centro Cardiologico Monzino IRCCS, Milan, Italy
[b] Department of Electronics, Information and Biomedical Engineering, Politecnico di Milano, Milan, Italy
[c] The BioRobotics Institute and Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Pisa, Italy
[d] Istituto Auxologico Italiano IRCCS, Milan, Italy
[e] Department of Biomedical Sciences and Morphological and Functional Imaging, "G. Martino" University Hospital Messina, Messina, Italy
[f] Department of Cardiology, IRCCS Policlinico San Donato, San Donato Milanese, Milan, Italy

ARTICLE INFO

Keywords:
Deep learning
Coronary artery disease
Stenosis classification
CAD-RADS
Coronary CT angiography
ConvMixer
Token-Mixer architecture

ABSTRACT

Background and objective: In patients with suspected Coronary Artery Disease (CAD), the severity of stenosis needs to be assessed for precise clinical management. An automatic deep learning-based algorithm to classify coronary stenosis lesions according to the Coronary Artery Disease Reporting and Data System (CAD-RADS) in multiplanar reconstruction images acquired with Coronary Computed Tomography Angiography (CCTA) is proposed.
Methods: In this retrospective study, 288 patients with suspected CAD who underwent CCTA scans were included. To model long-range semantic information, which is needed to identify and classify stenosis with challenging appearance, we adopted a token-mixer architecture (ConvMixer), which can learn structural relationship over the whole coronary artery. ConvMixer consists of a patch embedding layer followed by repeated convolutional blocks to enable the algorithm to learn long-range dependences between pixels. To visually assess ConvMixer performance, Gradient-Weighted Class Activation Mapping (Grad-CAM) analysis was used.
Results: Experimental results using 5-fold cross-validation showed that our ConvMixer can classify significant coronary artery stenosis (i.e., stenosis with luminal narrowing ≥50%) with accuracy and sensitivity of 87% and 90%, respectively. For CAD-RADS 0 vs. 1–2 vs. 3–4 vs. 5 classification, ConvMixer achieved accuracy and sensitivity of 72% and 75%, respectively. Additional experiments showed that ConvMixer achieved a better trade-off between performance and complexity compared to pyramid-shaped convolutional neural networks.
Conclusions: Our algorithm might provide clinicians with decision support, potentially reducing the interobserver variability for coronary artery stenosis evaluation.

1. Introduction

According to the Global Burden of Diseases study [1], the prevalence of Coronary Artery Disease (CAD) was approximately 150 million globally in 2016. CAD represents one of the major causes of mortality and morbidity worldwide, and its progression can lead to important adverse cardiovascular complications like acute myocardial infarction, stroke, and death [2]. These complications are generally the result of cumulative progression of atherosclerotic plaques that limit blood flow locally, causing stenosis.

As recommended in the recent CAD Reporting and Data System (CAD-RADS) consensus document [3], the degree of stenosis can be categorized into no (0%), minimal (1–24%), mild (25–49%), moderate (50–69%), severe (70–99%) stenosis and total occlusion (100%) of the coronary tree. An early and accurate assessment of stenosis degree is crucial to design proper therapeutic intervention, especially in cases with obstructive CAD (≥50% stenosis). While patients with no stenosis do not need any additional investigation, patients with non-obstructive CAD should be regularly follow-up. In patients with moderate stenosis, additional functional assessment along with a medical pharmacotherapy is recommended, whereas patients with severe stenosis or total

* Corresponding author. Cardiovascular Imaging Department, Centro Cardiologico Monzino IRCCS, Via C. Parea 4, 20138, Milan, Italy.
E-mail addresses: marco1.penso@mail.polimi.it (M. Penso), sara.moccia@santannapisa.it (S. Moccia), enrico.caiani@polimi.it (E.G. Caiani), careddagloria@gmail.com (G. Caredda), mluisalampus@gmail.com (M.L. Lampus), m.ludovica.carerj@outlook.it (M.L. Carerj), babbaromario@gmail.com (M. Babbaro), mauro.pepi@cardiologicomonzino.it (M. Pepi), mattia.chiesa@cardiologicomonzino.it (M. Chiesa), gianluca.pontone@cardiologicomonzino.it (G. Pontone).

https://doi.org/10.1016/j.compbiomed.2022.106484
Received 26 September 2022; Received in revised form 1 December 2022; Accepted 25 December 2022
Available online 26 December 2022

Abbreviations and acronyms

Acc	accuracy
AUC	Area Under the Curve
CAD	Coronary Artery Disease
CAD-RADS	Coronary Artery Disease Reporting and Data System
CCTA	Coronary Computed Tomography Angiography
CI	confidence intervals
GELU	Gaussian Error Linear Units
Grad-CAM	Gradient class activation map
MLP	Multi-Layer Perceptron
MPR	Multiplanar Reconstruction
NLP	Natural Language Processing
NPV	negative predictive value
PPV	positive predictive value
ROC	Receiver Operating Characteristic
Sens	sensitivity
ViT	Vision Transformer

occlusion should undergo further invasive testing along with a preventive pharmacotherapy or surgical intervention [3] (Table 1).

Recently, Coronary Computed Tomography Angiography (CCTA) has emerged as a noninvasive technique in the diagnosis of patient with suspected CAD. CCTA has been proved to be an effective modality for coronary stenosis degree quantification and characterization of the morphology and composition of coronary artery plaques [4]. Evidence increasingly supports the clinical utility of CCTA for risk stratification and decision making relevant to CAD, resulting in high diagnostic performance [5,6]. Thus far, in clinical practice, the severity of coronary artery stenosis relies on visual assessment of the whole coronary tree. This procedure is experience-related, time-consuming, and cumbersome [7]. To alleviate these issues, an accurate and automatic method to support physicians in the identification of coronary artery stenosis may play a critical role.

Several approaches have been proposed in the literature to automatically detect obstructive stenosis using machine-learning models [8–10]. Recently, a number of approaches based on deep learning has emerged to evaluate the degree of coronary stenosis in CCTA. In Ref. [11], a 2D CNN was proposed for CAD-RADS classification of the whole CCTA volume arranging slices in a 2D mosaic to reduce computational complexity. However, salient lesion details could have been lost in the image resizing procedure, thus decreasing the model performance. In addition, authors may have overestimated results since train and test set are not described to be split based on patient identities. In Ref. [12], a CNN with a support vector machine classifier was used to identify patients with functionally significant stenosis (i.e., $\geq$50% luminal narrowing) from the segmented left ventricular myocardium in CCTA images, but without providing single vessel analysis, and therefore limiting stenosis localization.

Table 1
CAD-RADS classification recommendations.

	CAD Class	Range of Coronary Stenosis	Clinical investigation recommendation
No stenosis	0	0%	No treatment
Minimal stenosis	1	1–24%	No additional diagnostic investigation
Mild stenosis	2	25–49%	No additional diagnostic investigation
Moderate stenosis	3	50–69%	Functional assessment
Severe stenosis	4	70–99%	Functional assessment or invasive coronary angiography
Total occlusion	5	100%	Viability assessment and invasive coronary angiography

Several approaches were developed to process Multiplanar Reconstruction (MPR) images from CCTA, which allow to display the complete course of a vessel in 2D [13]. One approach included texture-based multi planar analysis to predict significant stenosis from several views of coronary arteries [14]. In Ref. [15], a recurrent CNN was employed for significant stenosis detection and coronary plaque characterization. In Ref. [16], a 2.5 CNN was designed to classify stenosis according to the CAR-RADS score. Another study [17] combined recurrent CNN with shape-based radiomic features to predict significant stenosis. Also, in Ref. [18] a deep learning method based on CNNs was adopted to achieve automatic stenosis assessment.

Although these methods generally report high accuracy in the assessment of coronary stenosis, they have been mainly developed to perform binary classification (i.e., non-obstructive vs. obstructive or non-significant vs. significant stenosis). More importantly, CNNs exhibit a major limitation in modeling long-range contextual information due to an inherent restricted receptive field, thus potentially leading to poor classification performance. CNN success can be explained considering the inductive biases (such as translation and scale invariance), obtained stacking convolutions and requiring consecutively down-sampling operations in a pyramidal structure. However, when available old CT scanners resulting in poor images quality, as the spatial resolution is gradually reduced in deep networks, the classification accuracy may dramatically be compromised. In addition, convolutional is usually applied on small image regions, naturally leading to local inductive bias. This inductive bias enables CNNs to learn even in condition of small amount of data (as in the field of medical image analysis [19]) but results in a lack of global understanding. A number of approaches have been developed to model long-range dependencies in CNN, including atrous (a.k.a. dilated) convolutions [20–22], image pyramid [23,24] and large kernel [25,26]. However, these approaches may bring several drawbacks: (1) training deep networks with very large receptive fields on small medical image datasets tends to easily lead unstable performance and overfitting; (2) not effectively capture interaction over long-range spatial regions that is crucial for global understanding. The ability of neural network to learn long-range dependencies between pixels might help in making efficient classification, thereby leading to capture most salient global features explaining the variability of coronary stenosis in shape and size. Particularly, modeling global context might help to differentiate foreground pixels to those of the background as in cases of total occlusion where there is a complete interruption of contrast-enhancement along the coronary artery (Fig. 1A) or to identify lesions with dramatic size changes. In addition, introducing structural global knowledge may contribute to prevent miss-classifying anatomical proximal to distal coronary lumen diameter variations as occlusion (Fig. 1B). These characteristics represent a challenge for automatic learning tools, often leading to inaccurate lesion classification.

To model global features, a mechanism for improving the overall understanding of the image is needed. Recently, there has been a growing interest in Transformers [27,28] due to their global self-attention mechanism used to model long-range dependencies in Natural Language Processing (NLP) and, a few studies have explored their applications in computer vision [29–32]. Indeed, recent evidence suggests as convolutions might not be strictly necessary to reach performing visual recognition tasks [33,34]. In this context, Vision Transformer (ViT) [35] represents an attempt to develop a convolutional-free model for image classification. ViT reached competitive performance on the ImageNet classification task, but is computationally expensive, thus hampering its diffusion in clinical scenarios. Moreover, without any inductive bias, ViT requires huge amount of data to obtain good generalization with compared to CNN. However, this is not always feasible in many contexts, especially for medical imaging tasks as the number of available images is relatively scarce.

With the goal to provide inductive priors to transformer, standard

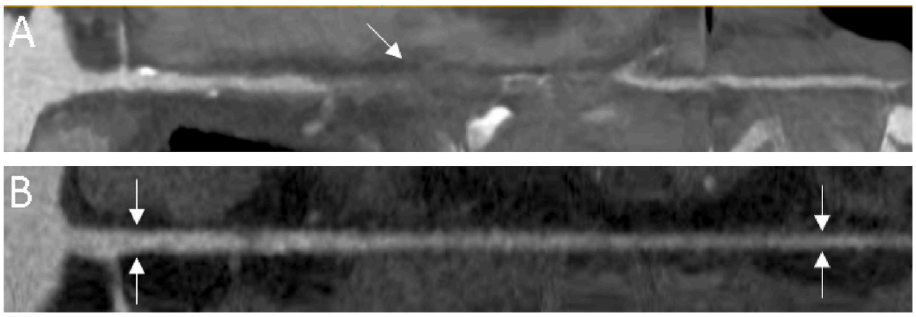

Fig. 1. Panel A: A segment of coronary artery with total occlusion (as highlighted by the white arrow) shows texture similar to that of the background. Panel B: Proximal and distal cross-sectional vessel lumen diameter may have highly different thickness.

convolutions were integrated with ViT improving its performance [36–38]. Recently, searching for computational efficiency alternative to transformer but with the objective of mixing information between patches (also called tokens) like self-attention, token-mixing architectures have been proposed. Tolstikhin et al. [39] presented MLP-mixer, a patch-base model with only Multi-Layer Perceptrons (MLPs) to simulate self-attention, achieving promising performance in image recognition. More recently, Trockman et al. [40] proposed ConvMixer, a pure-CNN backbone, that works like MLP-mixer in processing relationships between local patches in different spatial locations of the image but, with the advantage of image-specific inductive bias from convolutions.

Inspired by the literature, in this work we propose an algorithm to classify coronary stenosis according to the CAD-RADS score from MPR images. We used a token-mixer architecture to capture correlations between local tokens and learn their dependencies. Then, we highlighted the salient regions along coronary arteries for predicting the vessel-wise stenosis degree. Our main contributions are summarized as follows:

(1) We propose a classification algorithm to classify coronary artery vessels according to the CAD-RADS reporting system, as a support to diagnosis, by reducing the interobserver variability among physicians.
(2) We apply for the first time a token-mixer architecture to achieve large reception field in an efficient manner to specifically evaluate stenosis with challenging appearance, preserving inductive bias, thus leading to perform on small-scale data as in cardiac images analysis.
(3) We conducted extensive experiments, without limiting our attention to determination obstructive CAD (i.e., obstructive vs. non-obstructive stenosis), demonstrating as the proposed model achieved strong performance over different CNN-like models.

2. Methods

We here present the proposed algorithm (Sec. 2.1) for coronary artery stenosis classification in CCTA MPR images as well as our experiments (Sec. 2.2).

2.1. Token mixer architecture

The macro structure of the proposed token-mixer architecture (ConvMixer) is shown in Fig. 2. ConvMixer is based on a patch embedding layer (Sec. 2.1.1) to convert an input image into patches, like in vision transformers, and project them into a c-dimensional feature vector, followed by repeated convolutional blocks (Sec. 2.1.2) of equal size to update patch-wise representation, preserving the spatial resolution throughout all layers [40]. All these convolutional blocks rely on two main steps for patch communications: token mixing step and channel mixing step. We hypothesize that this isotropic architecture, combining the advantages of convolutions' locality biases with the advantages of processing long-range dependencies similarly to transformer, might be a better approach than conventional pyramid-CNN models for coronary artery stenosis characterization.

2.1.1. Patch embedding

Transformers like ViT [35] adopt an isotropic structure with a fixed number of non-overlap patches and unchanged embedding size (i.e., channel dimension), thus ideally leading to learn global interaction among different patch tokens. However, the self-attention module of ViT has a computational cost that is quadratically linear to the length of the input sequence (i.e., the number of patch tokens). This quadratic relationship between the number of patches and the image resolutions, makes it challenging to train the ViT model from limited data, undermining its application in many vision tasks. To address the computational limitation of the ViT, in this paper, we propose to linearly embed patches, thus improving performance in low data image classification. Specifically, given an input image of $W \, x \, H \, x \, C$ size, where W is the width, H the height and C the number of channels, it is first divided into N non-overlap patches of size $p \, x \, p$ that are linearly projected into a fixed D_i-dimensional embedded vector. This results in a patch features $X \in \mathbb{R}^{D \, x \, N}$ where $N = \frac{WH}{p^2}$ is the number of patches. The spatial structure of the embedded patches is maintained constant through the network. The

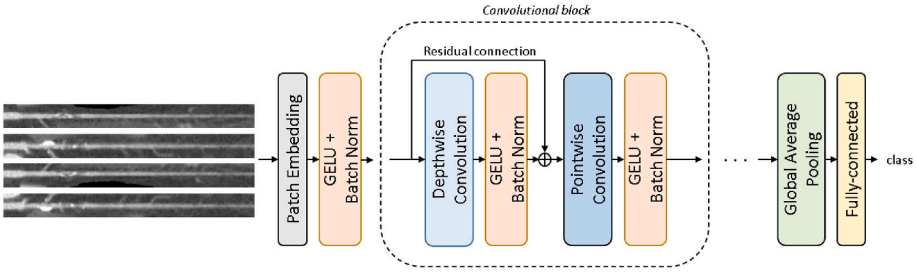

Fig. 2. Overview of the proposed architecture. From an artery MPR vessel, four views are defined and passed to the ConvMixer model which consists of a patch embedding, convolutional blocks and a classifier. Each convolutional block presents a depthwise convolution for token-mixing and a pointwise convolution for channel-mixing, each followed by a GELU nonlinearity and Batch Normalization.

lack of inductive prior limits the attention from exploiting the input image. Consequently, this step is implemented with a convolutional layer with kernel and stride size p and h filters. Indeed, replacing the initial linear embedding layer used in ViT by convolution layer led an inductive bias to improve model capability and generalization performance [41], and it also maintains memory efficiency. Using a convolutional layer to generate the embeddings provides both the inductive bias and spatial information to the subsequence layers, thus removing the need for additional positional embeddings as in ViT-based architectures.

2.1.2. Convolutional block

We introduce the convolutional block layer which updates patch features employing depthwise separable convolutions [42] that is a form of factorization of the convolution operation, with a depthwise convolution followed by a pointwise convolution operation. Specifically, while depthwise convolution performs a spatial convolution independently over each channel of the input data, pointwise convolution, that is a regular convolution with kernel 1×1, projects the output of the depthwise convolution onto a new channel space (more details are reported in Supplementary Material). The convolutional block is based on the idea of splitting the channel-mixing operations (i.e., channel combination) from the token-mixing operations (i.e., spatial feature learning). While the channel-mixing achieves the communication between different channels, the token-mixing enables communication between patches, simulating the self-attention block in transformer. Unlike [39], which use different MLP layers to replace self-attention, in this paper, we use convolutional for mixing information in spatial and channel dimensions: depthwise convolution to mix spatial locations, and pointwise convolution to mix channel locations. By having convolutions, the proposed model can take advantage of inductive biases, thus potentially leading to high performance on small-scale dataset. Indeed, while MLP-Mixer attains promising performance on large-scale scenarios, it is less effective when trained on small-scale data, even achieving lower performance than transformer [43]. In contrast to token-mixing MLP, the depthwise convolution provides different weights on different channels to enable information interaction among tokens, thus proving stronger encoding capacity.

Each convolution is followed by Gaussian Error Linear Units (GELU) [44] as activation function and batch normalization [45] to help prevent over-fitting of the model. GELU function is a smoother variant of Rectifier Linear Unit (ReLU) and is used instead to alleviate the problem of "leakage gradients", as shown in recent works including ViT [35] and MLP-mixer [39]. Furthermore, for each depthwise convolution, residual connection was introduced to promote the learning capability and reduce the overfitting problem [46].

Finally, the output patches features from the final convolutional block are flattened using an Average Global Pooling and fed to the fully connected layer, which serves as a classifier.

2.2. Experiments

2.2.1. Dataset

This study includes retrospectively collected CCTA acquisitions of 288 patients (age: 60.6 ± 12.4 years, 90 females) acquired between 2016 and 2018 at IRCCS Centro Cardiologico Monzino hospital (Milan, Italy). Institutional review board approval was obtained and patients provided informed consent. Patient characteristics as well as CCTA acquisition and analysis protocols have been described previously [11]. In briefly, CCTA scans were acquired using Discovery CT 750 HD or Revolution CT (GE Healthcare, Milwaukee, IL). CCTA acquisition and imaging protocols at each site were in adherence with the Society of Cardiovascular Computed Tomography guidelines [47]. Using MPR, coronary segments were evaluated for the presence of stenosis by a team of ten expert clinicians. To reduce observer variability especially in edge-cases, all examinations were analyzed by 2 expert readers (with ≥5 years of cardiac reading experience). For disagreements on data analysis between the 2 readers, a consensus agreement was achieved involving a third expert (with ten years of experience in cardiovascular imaging). According to Ref. [3], based on the degree of the stenosis, each coronary artery segment was label as no-stenosis (class 0, N = 248), minimal (class 1, N = 106), mild (class 2, N = 103), moderate (class 3, N = 122), severe (class 4, N = 124) or occluded (class 5, N = 76). As a result, the overall study cohort included a total of 779 coronary artery segments obtained from the 288 patients studied. For each segment, four MPR images (with size ranging from 120 × 800 to 170 to 850 pixels) were obtained rotating by 90° along the centerline of the vessel and then concatenated to define the input volume of $W \times H \times C$ size (C equal to 4 denoting the number of input channels). Note that, according to the clinical practice, when the exact required action needs to be identified, given a coronary artery tree with multiple stenosis (see Fig. 3) the associated CAD score is based on the maximum degree of lesion present.

2.2.2. Training

We set the patch size as 3×3. Input images were cropped along the segment centerlines, resized into 32×304 pixels size for computational efficiency, thus preserving their aspect ratio, and normalized into [0–1] range. The embedding patches were passed through 14 convolutional blocks, with setting the output channels to 68. Each depthwise convolution was set with 8×8 kernel size to minimize local receptive field constrain. The learning rate was set to 1e-3 and was used to minimize the cross-entropy loss function. Due to limited available training samples, an extensive on the fly data augmentation including random flipping, scaling, channel shuffling and gamma correction was used to improve the model's generalizability and robustness of the final model. All the training parameters were established with a trial-and-error procedure.

To validate our results, a 5-fold cross validation was performed to reduce performance bias. The splits were made on patient level with stratification by subgroup (i.e., stenosis degree) size. For each cross-validated fold, the dataset was partitioned into a training set (70%), validation set (10%) and test set (20%). The loss performance on the validation set is monitored during training and used for model selection.

2.2.3. Evaluation

For model evaluation, commonly used classification metrics were used, including accuracy (*Acc*), positive predictive value (*PPV*), negative predictive value (*NPV*), sensitivity (*Sens*) and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve:

$$Acc = \frac{TP + TN}{n} \tag{1}$$

$$PPV = \frac{TP}{TP + FP} \tag{2}$$

$$NPV = \frac{TN}{TN + FN} \tag{3}$$

$$Sens = \frac{TP}{TP + FN} \tag{4}$$

where *TP*, *TN*, *FP*, and *FN* are number of true positive, true negative, false positive and false negative, respectively. Moreover, for classification accuracy the Matthews Correlation Coefficient (MCC) was reported.

Gradient class activation map (Grad-CAM) [48] analysis was implemented to match clinical expectations of a visual explanation of model results allowing visualization of potentially salient regions of the image the model paid attention to.

From a clinical perspective, several experiments were conducted to assess the performance of the ConvMixer model for coronary artery stenosis classification. First, we evaluated the algorithm in predicting the clinically relevant stenosis (i.e., non-obstructive (0–2) vs.

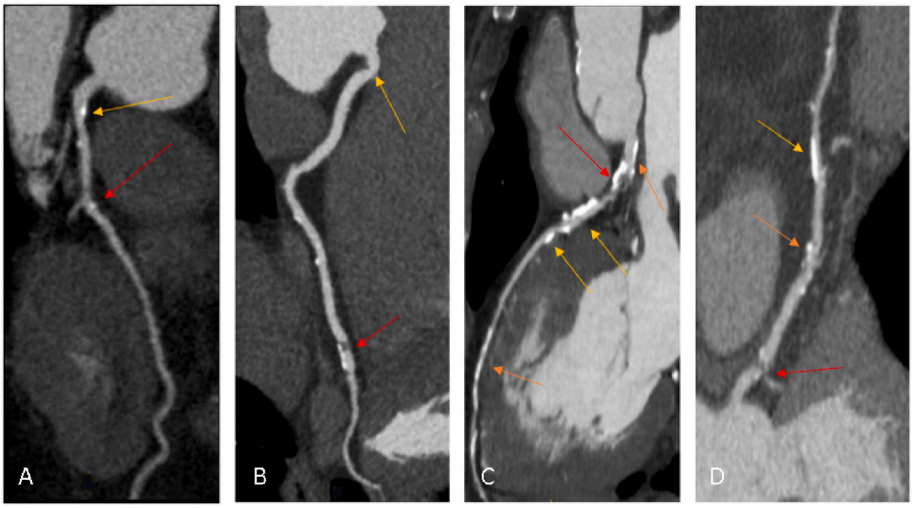

Fig. 3. Examples of coronary artery labeled as CAD-RADS 4. Mild stenosis (yellow arrow); Moderate stenosis (orange arrow); Severe stenosis (red arrow).

obstructive (3–5) stenosis). Second, we evaluated ConvMixer in the differentiation between non-obstructive coronary stenosis (1–2) and normal segments (0). Third, we reported the results of the token mixing algorithm in classifying severity of obstructive coronary stenosis, before detecting cases with total vessel occlusion (3–4 vs. 5) and then differentiating among all degrees (3 vs. 4 vs. 5). Lastly, we evaluated the effectiveness of the algorithm in the case 0 vs. 1–2 vs. 3–4 vs. 5. In all experiments, minimal and mild stenosis were grouped, according to the patient clinical management recommendation (Table 1).

For each experiment, we compared ConvMixer with conventional classification CNN-based models, including ResNet-50 [49], VGG16 [50], and DenseNet121 [51] pre-trained on the Imagenet dataset. To adapt the input image to a pre-trained model, at the beginning of the encoder of ResNet-50, VGG16 and DenseNet121, one convolutional layer with 3 x 3 kernel size and three filters was added for changing the number of channels of the input image. Further, to verify the hypothesis of effectiveness in modeling long-range dependencies, as the majority of CNNs adopt pyramid structure to compute multi-scale feature efficiently, we compared the proposed model (ConvMixer) with CNN-like pyramid architectures. Inspired by recent work in processing multi-scale information [52,53], in the tested CNNs conventional convolutions were replaced with either atrous convolutions (AtrousCNN) or Inception module [54] (InceptionV2CNN, InceptionACNN, InceptionBCNN). Specifically, an Inception-v2 version [26] was adopted for InceptionV2CNN, while an inspired Inception-ResNet version [46] was used in both InceptionACNN and InceptionBCNN. More details about network configurations can be found in Supplementary Material. For a fair comparison, we conducted a hyperparameter sweep for every different model and report the best results we were able to achieve. In all experiments the learning rate was reduced on plateau of 6 epochs by a factor of 0.8 and early stopping was applied after validation loss had no longer decreased for 30 epochs. All models were trained using in an end-to-end fashion using Stochastic Gradient Descent optimizer with momentum (0.9) and batch size set to 4.

A qualitative comparison with previous deep learning methods [11, 14–18] is also reported. These works perform obstructive stenosis classification using different procedure and evaluated on different datasets and so, for a fair comparison the performance of competing state-of-the-art methods was adopted from the original publications.

To evaluate whether our automatic algorithm represents a helpful decision-making support system for coronary artery stenosis reducing the interobserver variability, two expert readers, blinded to patient clinical history and data, independently evaluated each coronary stenosis in two separate ways: before without any support, and at two weeks with the support of our automatic algorithm. This was performed for one-hundred coronary vessels randomly selected from the study cohort and classified according to the presence of obstructive stenosis (i.e., 0–2 vs. 3–5), and also differentiating between CAD-RADS class 0 vs. 1–2 vs. 3–4 vs. 5.

3. Results

Table 2 lists the results of the averaged performance obtained for classification of the obstructive CAD (obstructive vs. non-obstructive stenosis) comparing different models. Values are expressed as mean and confidence intervals (CI) were set at 95%. We compared the ConvMixer network with conventional CNN-based classification models. As shown in the table, ConvMixer achieved better performance than fine-tuning ResNet50, VGG16 and DenseNet121 in term of both *Acc* (0.87) and AUC (0.93; 95% CI: 0.87–0.98) (Fig. 4a). This corresponds to a mean *PPV*, *NPV* and *Sens* of 0.82, 0.93 and 0.90, respectively.

The second group of comparison is based on CNN architectures, specifically AtrousCNN, InceptionV2CNN, InceptionACNN, and InceptionBCNN, that capture context in an image by modeling long-range dependencies. All these networks have achieved similar performance but lower recognition accuracy compared with ConvMixer (Table 2). Overall, ConvMixer achieved a better trade-off between performances and complexity, with considerably fewer parameters than the other networks, making the proposed algorithm attractive in term of computational efficiency.

After that, we compared the ConvMixer to the previous relevant state-of-the-art methods that exploited deep learning for obstructive stenosis prediction. The results reported in Table 2 (bottom) showed that the proposed method seems to outperform previous ones. Specifically, ConvMixer achieved overall better performances, with an absolutely superior *Acc* of 16%, 1%, 6% and 1% than [11,14,16,18], respectively.

After that, we compared the ConvMixer to the previous relevant state-of-the-art methods that exploited deep learning for obstructive stenosis prediction. The results reported in Table 2 (bottom) showed that the proposed method seems to outperform previous ones. Specifically, ConvMixer achieved overall better performances, with an absolutely superior *Acc* of 16%, 1%, 6% and 1% than [11,14,16,18], respectively. Compared to Ref. [15], our algorithm reported an inferior *Acc* (0.87 vs. 0.93) but a higher *Sens* (0.90 vs. 0.61) and *PPV* (0.82 vs. 0.65) to differentiate vessel with/without obstructive CAD. Also, we observed [17] achieved a more accurate diagnosis of significant coronary artery stenosis than our algorithm (0.92 vs. 0.87). This can be explained considering that their dataset was highly unbalanced, with only 25% of

Table 2
Diagnostic accuracy of the significant coronary stenosis (non-obstructive vs. obstructive stenosis) for the token mixer architecture (ConvMixer) and other models/methods.

Method	AUC (95%CI)	Acc	PPV	NPV	Sens	Params
Stenosis 0-2 vs. 3–5						
ConvMixer	**0.93 (0.87–0.98)**	**0.87**	0.82	**0.93**	0.90	**174.7 K**
ResNet50	0.60 (0.42–0.79)	0.54	0.49	0.88	**0.93**	23.5 M
VGG16	0.90 (0.83–0.96)	0.85	**0.83**	0.85	0.78	14.7 M
DenseNet121	0.69 (0.62–0.77)	0.60	0.53	0.82	0.85	7.0 M
AtrousCNN	0.90 (0.85–0.95)	0.82	0.77	0.88	0.84	895.6 K
InceptionV2CNN	0.88 (0.80–0.95)	0.82	**0.83**	0.84	0.75	813.4 K
InceptionACNN	0.89 (0.81–0.97)	0.83	0.79	0.87	0.83	524.0 K
InceptionBCNN	0.88 (0.82–0.95)	0.82	0.77	0.86	0.81	371.9 K
Muscogiuri et al. [11]	0.78 (−15%)	0.71 (−16%)	0.69 (−13%)	0.74 (−19%)	0.82 (−8%)	–
Zreik et al. [15]	–	0.93 (+6%)	0.65 (−17%)	–	0.61 (−29%)	340 K
Denzinger et al. [17]	0.96 (+3%)	0.92 (+5%)	0.94 (+12%)	0.82 (−11%)	0.96 (+6%)	–
Denzinger et al. [16]	0.92 (−1%)	0.86 (−1%)	–	–	0.89 (−1%)	–
Tejero-de-Pablos et al. [14]	–	0.81 (−6%)	–	–	0.90 (+0%)	–
Han et al. [18]	0.87 (−6%)	0.86 (−1%)	0.73 (−9%)	0.94 (+1%)	0.88 (−2%)	–

AUC, area under the curve; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value. (Note: means the results are not reported by that methods.)

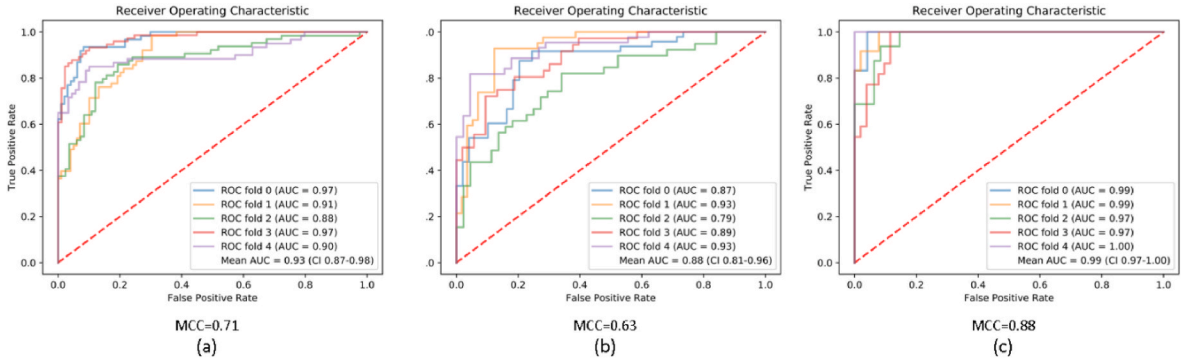

Fig. 4. Receiver operating characteristic curves for predicting coronary stenosis according to CAD-RADS classification and the overall Matthews Correlation Coefficient (MCC): (a) class 0–2 vs. 3–5, (b) class 0 vs. 1–2, (c) class 3–4 vs. 5.

lesions labeled as having luminal narrowing above 50%. In contrast, in our dataset the number of significant stenosis was 41%, thus leading to a potentially higher number of false predictions.

Table 3 summarized the results on the other experiments. Overall, ConvMixer performed better than the other models, with a higher recognition accuracy, suggesting that the proposed algorithm can efficiently model long-range dependencies in contrast to conventional methods. Specifically, moving on the task of discriminate between non-obstructive lesion and no stenosis (class 0 vs. 1–2), using ConvMixer, an average AUC of 0.88 (95% CI: 0.81–0.96) was achieved (MCC = 0.63, see Fig. 4b), while for *Acc* set to 0.84, corresponding *PPV*, *NPV* and *Sens* were 0.82, 0.86 and 0.84, respectively (Table 3). The relatively large variations between ROC curves might be explained considering the high inter-class similarity, especially between the CAD-RADS class 0 and 1, where relatively small lesions can be difficult to identify.

Given the importance of accurate distinction of coronary severe stenosis from occlusion (i.e., class 3–4 vs. 5), we observed a high recognition accuracy (0.96) for patients with a complete occlusion of the coronary vessel (Table 3). The average AUC over the five folds was 0.99 (95% CI: 0.97–1.00) and the MCC was 0.88, as shown in Fig. 4c. Assuming a differentiation between moderate and severe lesions, the averaged *Acc* of our automatic method on obstructive stenosis (class 3 vs. 4 vs. 5) was 0.67 (Table 3). As shown in Table 3, compared with ConvMixer, the VGG16 model achieved slightly higher performance, with *Acc* of 0.69. The final confusion matrix resulting from ConvMixer is reported in Fig. 5a. Our method achieved better accuracy for lesions of grade-5 compared with lesions of grade-3 or -4. As shown in the figure, inaccuracies generated by our system are mainly within one class distance. Indeed, correct assessment of moderate to severe lesions might be difficult even for clinicians, as experience-based. This is particularly true in borderline lesions since a distinction between moderate (i.e., $\leq$ 69%) to severe (i.e., $\geq$ 70%) stenosis remains a challenging decision, potentially leading to false positive and false negative detections.

When applying the algorithm to the multi-class classification (i.e., no stenosis vs. CAD 0-1 vs. CAD 3-4 vs. total occlusion), a mean *Acc* = 0.72 was obtained (Table 3). This led to *PPV*, *NPV* and *Sens* of 0.73, 0.90 and 0.75, respectively. The relative confusion matrix is shown in Fig. 5b. From the results of confusion matrix, we observed that 11 healthy coronary vessels (4.4%) were misclassified with obstructive CAD, and 19 vessels with obstructive stenosis (5.9%) were predicted as having no stenosis. This might be explained considering that, despite the image quality was generally good for all coronary segment, some images could have a low signal-to-noise ratio, thus leading to misclassifications.

Fig. 6 displays the results of the interobserver variability analysis. For the task of classifying obstructive stenosis (i.e., 0–2 vs. 3–5), no benefits were visible from the support of the algorithm (11% vs. 11%); on the contrary, while differentiating between classes of CAD-RADS, the variability in the visual assessment interpretation between expert readers was reduced from 26% to 14% when the clinical decision was supported by the automatic classification.

In Fig. 7 are visualized some examples of the output of Grad-CAM analysis to highlight salient localizations of lesions predicted by our

Table 3
Results of the ConvMixer architecture and other models on coronary stenosis classification.

Model	AUC (95%CI)	Acc	PPV	NPV	Sens
Stenosis 0 vs. 1–2					
ConvMixer	**0.88** (0.81–0.96)	**0.84**	**0.82**	0.86	0.84
ResNet50	0.53 (0.34–0.71)	0.56	0.50	**0.92**	**0.98**
VGG16	0.73 (0.55–0.91)	0.70	0.61	0.88	0.92
DenseNet121	0.54 (0.36–0.71)	0.56	0.50	**0.92**	0.96
AtrousCNN	0.81 (0.70–0.92)	0.78	0.71	0.87	0.86
InceptionV2CNN	0.81 (0.72–0.91)	0.74	0.67	0.88	0.89
InceptionACNN	0.87 (0.80–0.94)	0.80	0.73	0.90	0.91
InceptionBCNN	0.86 (0.81–0.92)	0.79	0.71	0.91	0.92
Stenosis 3-4 vs. 5					
ConvMixer	**0.99** (0.97–1.00)	**0.96**	**0.87**	**0.99**	**0.97**
ResNet50	0.65 (0.38–0.92)	0.70	0.45	0.93	0.79
VGG16	0.84 (0.73–0.96)	0.82	0.68	0.93	0.77
DenseNet121	0.59 (0.34–0.83)	0.65	0.40	0.86	0.71
AtrousCNN	0.91 (0.82–0.99)	0.87	0.64	0.97	0.90
InceptionV2CNN	0.88 (0.78–1.00)	0.88	0.72	0.97	0.93
InceptionACNN	0.91 (0.82–1.00)	0.91	0.81	0.96	0.85
InceptionBCNN	0.91 (0.82–0.99)	0.89	0.79	0.95	0.84
Stenosis 3 vs. 4 vs. 5					
ConvMixer	–	0.67	0.70	0.83	0.68
ResNet50	–	0.64	0.70	0.81	0.66
VGG16	–	**0.69**	**0.73**	**0.85**	**0.69**
DenseNet121	–	0.67	0.70	0.81	0.66
AtrousCNN	–	0.62	0.68	0.81	0.61
InceptionV2CNN	–	0.55	0.62	0.76	0.57
InceptionACNN	–	0.51	0.56	0.74	0.53
InceptionBCNN	–	0.55	0.58	0.77	0.55
Stenosis 0 vs. 1–2 vs. 3–4 vs. 5					
ConvMixer	–	**0.72**	**0.73**	**0.90**	**0.75**
ResNet50	–	0.27	0.10	0.80	0.25
VGG16	–	0.52	0.48	0.82	0.43
DenseNet121	–	0.39	0.34	0.79	0.34
AtrousCNN	–	0.63	0.70	0.85	0.63
InceptionV2CNN	–	0.61	0.64	0.84	0.64
InceptionACNN	–	0.65	0.65	0.85	0.67
InceptionBCNN	–	0.65	0.68	0.85	0.67

AUC, area under the curve; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value.

algorithm. Blue regions in the heatmap correspond to normal predicted regions in the coronary vessel, whereas hot colors highlight detected abnormal regions corresponding to stenosis. For most test cases, it was observed our algorithm was able to provide generally accurate qualitative localization of potential stenosis, although there is not always full agreement with the reference annotations (see Fig. 7 bottom).

To better understand the effectiveness of the proposed algorithm to model long-range dependencies to obtain global context information for CAD classification with comparted to atrous convolution or inception module, as shown in Fig. 8, we visualized the attention map generated by different models. It is possible to appreciate how the ConvMixer's ability to model long-range dependencies results in a more accurate attention on stenosis regions and thus, potentially leading to better CAD classifications. Without an explicit capacity of covering interdependencies between spatial regions, the model may fail to capture coronary artery stenosis, shifting its attention on normal regions.

4. Discussion and conclusions

This study presented a new framework for automatic coronary stenosis classification. It was employed a token-mixer architecture that analyzes MPR view of a coronary artery segment to diagnose CAD according to the recent CAD-RADS score [3]. The algorithm is based on an isotropic structure that preserves the feature maps' size throughout the network and enables the communications between spatial and channel locations using depthwise separable convolutions. Usually, in pyramid-shaped architectures, the number of filters increase gradually as the network goes deeper. The greater is the number of parameters, the larger would be the amount of training data to achieve high

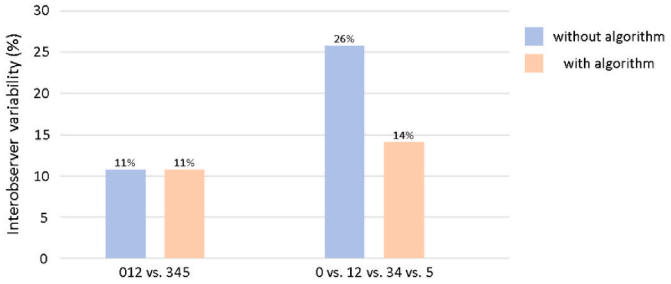

Fig. 6. Decision-making support system analysis: interobserver variability in stenosis assessment with and without the support of the automatic classification for CAD-RADS (see text for more details) to highlight the presence of obstructive stenosis (i.e., 0–2 vs. 3–5), and to differentiate between classes (0 vs. 1–2 vs. 3–4 vs. 5).

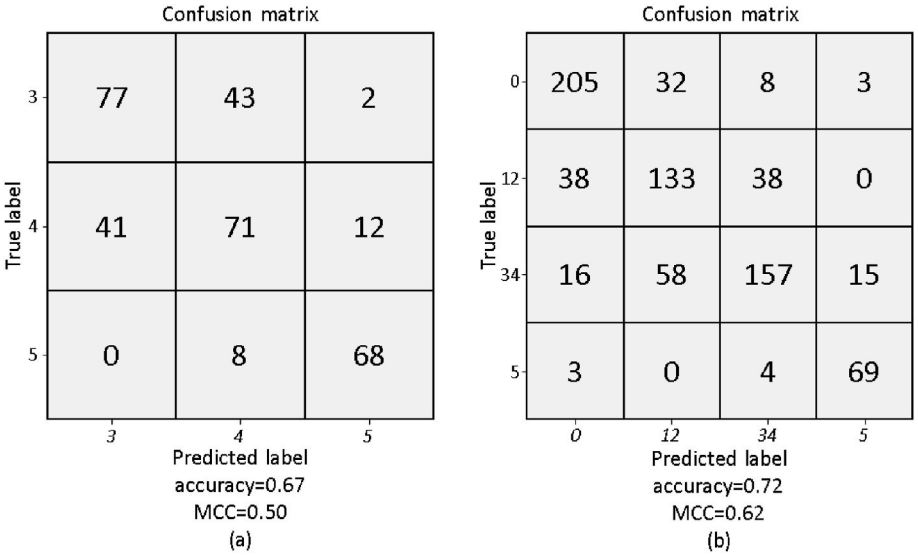

Fig. 5. Confusion matrix of the automatic stenosis classification and the Matthews Correlation Coefficient (MCC) for: (a) class 3 vs. 4 vs. 5, and (b) class 0 vs. 1–2 vs. 3–4 vs. 5.

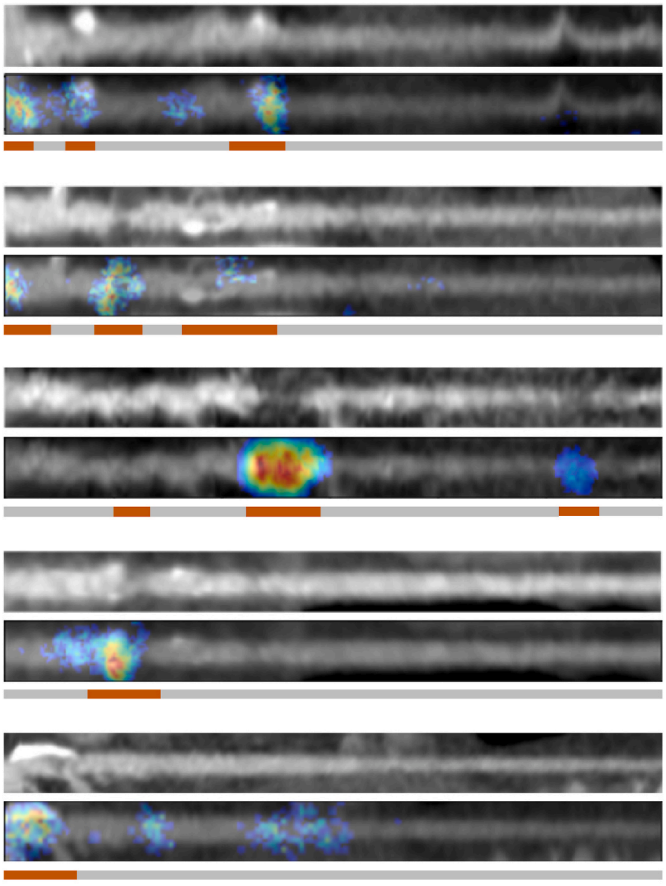

Fig. 7. Visual samples of Multiplanar Reconstruction (MPR) images (gray) and their corresponding visual attention map (color) generated by Grad-CAM [48].

performance. This might be especially difficult for clinical tasks, as labeling medical data is costly, and thus, hardly available. Moreover, reducing the number of parameters might control overfitting and thus greatly improved the testing accuracy. Inspired by the recent successful in computer vision of ViT, and by the more recently token mixer architectures, ConvMixer is able to encode long-range spatial information using exclusively convolutions with the effect of adding inductive biases, and thus leading for high data efficiency. As shown in Fig. 8, in contrast to conventional approaches that enable to increase the respective fields, as atrous convolution or inception module, ConvMixer can manipulate long-range dependencies more efficiently and with less number of parameters, thus leading to a better accuracy/cost trade-off. Incorporating knowledge of larger region rather than just locally might help the algorithm to capture the whole vessel structure in order to detect and classify lesions from multiple sizes, thus contributing to the overall performance improvement.

The results showed that our neural network was able to identify obstructive stenosis accurately (AUC = 0.93). From a clinical perspective, the diagnosis of obstructive CAD is relevant considering that severe lesions may lead to adverse acute events as myocardial ischemia. Indeed, several studies have demonstrated that patients with moderate to severe stenosis had an increased risk of adverse cardiovascular events compared to non-obstructive CAD and, therefore, need further functional assessment and intensive treatment [55,56]. In contrast, patients with non-obstructive CAD are generally not related to myocardial ischemia and may not need further diagnostic work-up or extensive follow-up. CCTA has increasingly been used to exclude obstructive stenosis in suspected CAD due to its high accuracy and NPV [57]. If CCTA fails to detect obstructive CAD, optimal medical care is uncertain. Indeed, from a clinical perspective, *FN* are more dangerous than *FP* as subject to decreasing prevalence. While the cost of *FP* is generally limited to the cost of additional assessments and/or therapies, the cost of *FN* is, in this scenario, the risk of future acute coronary syndrome, as myocardial ischemia. Recently, Chang et al. [58] demonstrated that about 75% of the lesions that will develop acute coronary syndrome were non-obstructive stenosis. Our algorithm demonstrated to be good in discriminating segments with obstructive CAD from those without non-obstructive CAD or no CAD, reporting a NPV of 0.93. This may help the clinician in the diagnosis reducing the interobserver variability, to guide subsequent management and, thus making patient management more efficient. Further analysis revealed that even in the hardest scenario to differentiate between CAD-RADS 0 vs. 12 vs. 34 vs. 5, the NPV was still clinically relevant (0.90). Furthermore, when ≥50% luminal stenosis was diagnosed as CAD, we investigated whether the algorithm was able to identify coronary artery segments with a total occlusion, especially considering that patients in the extreme case (i.e., CAD-RADS 5) always need invasive coronary angiography. Results validated the robust performance of the presented DL algorithm in detect total occlusion. However, the differentiation between moderate and severe stenosis remains challenging. This is not surpassing, considering that even for expert clinicians, image interpretability often results in non-negligible interobserver variability.

The application of deep learning systems based on token mixer architectures in cardiac radiology might represent an important step in the detection and management of patients with suspected CAD. An automatic algorithm for accurate CAD classification can be helpful in supporting clinicians in their health-care delivery, reducing the time of analysis and facilitating the diagnosis of patients with significant stenosis that may benefit from medical therapy and further investigation. Furthermore, as the majority of patients who undergo CCTA for suspected CAD results in no evidence of obstructive CAD, automatic deep learning analysis might represent a helpful tool for reducing the number

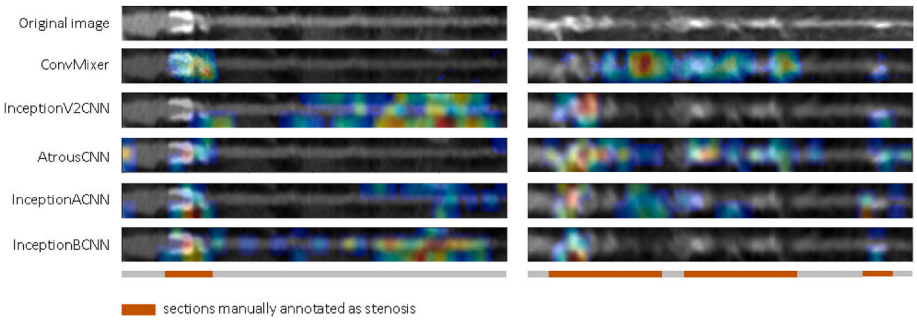

Fig. 8. Visual comparison of attention maps generated by Grad-CAM [48] between ours (ConvMixer) and CNNs with either atrous convolutions (AtrousCNN) or Inception module (InceptionV2CNN, InceptionACNN, InceptionBCNN).

of patients who undergo further unnecessary invasive investigation. Results in Fig. 6 highlight the potential role that the deep learning might represent in improving the current clinical workflow, reducing the interobserver variability, and promoting for consistency diagnosis. Moreover, especially in small hospitals, due to the lack of expert radiologists or cardiologists, these algorithms might be a valid tool for assisting and training unexpert clinicians.

Besides the reported results, this study has several limitations. First, the dataset was sourced by a single hospital adopting the same imaging protocol, thus limiting reproducibility of the deep learning algorithm. Moreover, although CCTA were acquired using two different scanners, it should be mentioned that these scanners are of recent generation and therefore, the algorithm may not generalize well on images obtained with other versions of scanners characterized by a lower level of signal-to-noise-ratio. Second, despite our sample size was comparable with those of previous works [14–16,18], a larger dataset would allow enhancing the conclusion about the effective role of the deep learning algorithm in the decision-making process. Despite the computationally efficient approach, a large, annotated dataset is still required due to the high number of classes to be classified. Indeed, in the most general evaluation only 4 classes (0 vs. 1–2 vs. 3–4 vs. 5) were considered, but still representing a step forward compared to many of the previous works. Third, image noise and potentially annotations errors of the CAD-RADS class may affect the algorithm accuracy. Indeed, using invasive coronary angiography as reference instead of visual estimation of CCTA may reduce CAD-RADS classification errors of expert readers. However, in clinical practice, this procedure is limited to severe lesions. Fourth, our algorithm was designed as a support in the evaluation of coronary stenosis, but the inherent limited explicability of the algorithm results might limit its applicability in the clinical practice. Indeed, the classification of the algorithm cannot provide intuitive explanations and reasoning of the diagnosis like clinical experts. To overcome this limitation, the visual attention maps generated by Grad-CAM were provided. By visualizing the attention maps generated by the Grad-CAM, clinicians can visualize potentially salient areas in the images focused by the algorithm to predict stenosis degree.

As future extension of this work, instead of focusing only on the assessment of the most severe lesion of a given coronary tree, an object detection approach could be investigated to deal with the need of improving the quality of diagnostic accuracy.

In conclusion, alongside the increased number of patients with suspected CAD and the need to reduce patients that undergoing further unnecessary invasive analysis, a deep learning algorithm that automatically evaluates the coronary stenosis degree according to the recent CAD-RADS classification score is proposed. The proposed algorithm is based on a token mixer architecture that by adopting depthwise separable convolutions, can combine both inductive biases of convolutions, with the capacity of encoding long-range dependencies of Transformers. With several experiments, we have shown that compared to traditional CNN-based architectures our algorithm achieved better performances with a smaller number of parameters, representing a support system to reduce the interobserver variability in coronary artery stenosis assessment.

Author Contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

Funding

This research was supported by the Italian Ministry of Health-Ricerca Corrente to Centro Cardiologico Monzino IRCCS.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.106484.

References

[1] GBD 2016 Disease and Injury Incidence and Prevalence Collaborators, Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016, Lancet 390 (10100) (2017) 1211–1259.
[2] Task Force Members, G. Montalescot, U. Sechtem, et al., ESC guidelines on the management of stable coronary artery disease: the task force on the management of stable coronary artery disease of the European Society of Cardiology, Eur. Heart J. 34 (38) (2013) 2949–3003, 2013.
[3] R.C. Cury, S. Abbara, S. Achenbach, et al., Coronary artery disease–reporting and data system (CAD-RADS): an expert consensus document of SCCT, ACR and NASCI: endorsed by the ACC, JACC Cardiovasc. Imag. 9 (2016) 1099–1113.
[4] K.M. Abdelrahman, M.Y. Chen, A.K. Dey, et al., Coronary computed tomography angiography from clinical uses to emerging technologies: JACC state-of-the-art review, J. Am. Coll. Cardiol. 76 (10) (2020) 1226–1243.
[5] G. Pontone, D. Andreini, C. Quaglia, et al., Accuracy of multidetector spiral computed tomography in detecting significant coronary stenosis in patient populations with differing pre-test probabilities of disease, Clin. Radiol. 62 (10) (2007) 978–985.
[6] X. Yin, J. Wang, W. Zheng, et al., Diagnostic performance of coronary computed tomography angiography versus exercise electrocardiography for coronary artery disease: a systematic review and meta-analysis, J. Thorac. Dis. 8 (7) (2016) 1688–1696.
[7] A. Arbab-Zadeh, J. Hoe, Quantification of coronary arterial stenoses by multidetector CT angiography in comparison with conventional angiography methods, caveats, and implications, JACC Cardiovasc. Imag. 4 (2) (2011) 191–202.
[8] S. Sankaran, M. Schaap, S.C. Hunley, et al., Hale: healthy area of lumen estimation for vessel stenosis quantification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 380–387.
[9] S. Cetin, G. Unal, Automatic detection of coronary artery stenosis in CTA based on vessel intensity and geometric features, in: Proc. Of MICCAI Workshop" 3D Cardiovascular Imaging: a MICCAI Segmentation Challenge, 2012.
[10] M. Tessmann, F. Vega-Higuera, D. Fritz, et al., Multi-scale feature extraction for learning-based classification of coronary artery stenosis, in: Medical Imaging 2009: Computer-Aided Diagnosis, vol. 7260, 2009, pp. 21–28.
[11] G. Muscogiuri, M. Chiesa, M. Trotta, et al., Performance of a deep learning algorithm for the evaluation of CAD-RADS classification with CCTA, Atherosclerosis 294 (2020) 25–32.
[12] M. Zreik, N. Lessmann, R.W. van Hamersvelt, et al., Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis, Med. Image Anal. 44 (2018) 72–85.
[13] Z. Sun, G.H. Choo, K.H. Ng, Coronary CT angiography: current status and continuing challenges, Br. J. Radiol. 85 (1013) (2012) 495–510.
[14] A. Tejero-de-Pablos, K. Huang, H. Yamane, et al., Texture-based classification of significant stenosis in CCTA multi-view images of coronary arteries, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 732–740.
[15] M. Zreik, R.W. van Hamersvelt, J.M. Wolterink, et al., A recurrent CNN for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography, IEEE Trans. Med. Imag. 38 (7) (2019) 1588–1598.
[16] F. Denzinger, M. Wels, K. Breininger, et al., Automatic CAD-RADS scoring using deep learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020, pp. 45–54.
[17] F. Denzinger, M. Wels, N. Ravikumar, et al., Coronary artery plaque characterization from CCTA scans using deep learning and radiomics, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 593–601.
[18] D. Han, J. Liu, Z. Sun, et al., Deep learning analysis in coronary computed tomographic angiography imaging for the assessment of patients with coronary artery stenosis, Comput. Methods Progr. Biomed. 196 (2020), 105651.
[19] L. Cai, J. Gao, D. Zhao, A review of the application of deep learning in medical image classification and segmentation, Ann. Transl. Med. 8 (11) (2020) 713.
[20] S. Wang, S.Y. Hu, E. Cheah, et al., U-Net Using Stacked Dilated Convolutions for Medical Image Segmentation, 2020 arXiv preprint arXiv: 2004.03466.

[21] G. Papandreou, I. Kokkinos, P.A. Savalle, Modeling local and global deformations in deep learning: epitomic convolution, multiple instance learning, and sliding window detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 390–399.
[22] X. Lei, H. Pan, X. Huang, A dilated CNN model for image classification, IEEE Access 7 (2019) 124087–124095.
[23] M. Gridach, PyDiNet: pyramid dilated network for medical image segmentation, Neural Network. 140 (2021) 274–281.
[24] H. Zhao, J. Shi, X. Qi, et al., Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.
[25] C. Peng, X. Zhang, G. Yu, et al., Large kernel matters–improve semantic segmentation by global convolutional network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4353–4361.
[26] C. Szegedy, V. Vanhoucke, S. Ioffe, et al., Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[27] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[28] J. Devlin, M.W. Chang, K. Lee, et al., Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018 arXiv preprint arXiv:1810.04805.
[29] E. Xie, W. Wang, W. Wang, et al., Segmenting transparent object inthe wild with transformer, in: Proc Int Joint Conf Artificial Intell, 2021.
[30] N. Liu, N. Zhang, K. Wan, et al., Visual saliency transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4722–4732.
[31] X. Zhu, W. Su, L. Lu, et al., Deformable DETR: deformable transformers for end-to-end object detection, in: Proc Int Conf Learn Representations, 2021.
[32] S. Khan, M. Naseer, M. Hayat, et al., Transformers in Vision: A Survey. ACM Computing Surveys, CSUR), 2021.
[33] P. Ramachandran, N. Parmar, A. Vaswani, et al., Stand-alone self-attention in vision models, Adv. Neural Inf. Process. Syst. 32 (2019).
[34] H. Zhao, J. Jia, V. Koltun, Exploring self-attention for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10076–10085.
[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020 arXiv preprint arXiv: 2010.11929.
[36] A. Hassani, S. Walton, N. Shah, et al., Escaping the Big Data Paradigm with Compact Transformers, 2021 arXiv preprint arXiv:2104.05704.
[37] B. Graham, A. El-Nouby, H. Touvron, et al., Levit: a vision transformer in convnet's clothing for faster inference, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12259–12269.
[38] H. Wu, B. Xiao, N. Codella, et al., Cvt: introducing convolutions to vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 22–31.
[39] I.O. Tolstikhin, N. Houlsby, A. Kolesnikov, et al., Mlp-mixer: an all-mlp architecture for vision, Adv. Neural Inf. Process. Syst. 34 (2021) 24261–24272.
[40] A. Trockman, J.Z. Kolter, Patches Are All You Need?, 2022 arXiv preprint arXiv: 2201.09792.
[41] P. Jeevan, A. Sethi, Vision Xformers: Efficient Attention for Image Classification, 2021 arXiv preprint arXiv:2107.02239.
[42] L. Sifre, S. Mallat, Rigid-motion Scattering for Texture Classification, 2014 arXiv preprint arXiv:14031687.
[43] T. Yu, X. Li, Y. Cai, et al., S2-mlp: spatial-shift mlp architecture for vision, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 297–306.
[44] B. Heo, S. Yun, D. Han, et al., Rethinking spatial dimensions of vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11936–11945.
[45] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456.
[46] C. Szegedy, S. Ioffe, V. Vanhoucke, et al., Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI Conference on Artificial Intelligence, 2017.
[47] S. Abbara, P. Blanke, C.D. Maroules, et al., SCCT guidelines for the performance and acquisition of coronary computed tomographic angiography: a report of the society of Cardiovascular Computed Tomography Guidelines Committee: endorsed by the North American Society for Cardiovascular Imaging (NASCI), J. Cardiovasc Comput. Tomogr. 10 (2016) 435–449.
[48] R.R. Selvaraju, M. Cogswell, A. Das, et al., Grad-cam: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[49] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[50] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014 arXiv preprint arXiv:1409.1556.
[51] G. Huang, Z. Liu, L. Van Der Maaten, et al., Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[52] M. Kim, N. Ilyas, K. Kim, AMSASeg: an attention-based multi-scale atrous convolutional neural network for real-time object segmentation from 3D point cloud, IEEE Access 9 (2021) 70789–70796.
[53] W. Ma, Y. Wu, Z. Wang, et al., Mdcn: multi-scale, deep inception convolutional neural networks for efficient object detection, in: 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 2510–2515.
[54] C. Szegedy, W. Liu, Y. Jia, et al., Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[55] J. Schulman-Marcus, B.Ó. Hartaigh, H. Gransar, et al., Sex-specific associations between coronary artery plaque extent and risk of major adverse cardiovascular events: the CONFIRM long-term registry, JACC Cardiovasc. Imag. 9 (4) (2016) 364–372.
[56] I. Cho, H.J. Chang, B. Ó Hartaigh, et al., Incremental prognostic utility of coronary CT angiography for asymptomatic patients based upon extent and severity of coronary artery calcium: results from the COronary CT Angiography EvaluatioN for Clinical Outcomes InteRnational Multicenter (CONFIRM) study, Eur. Heart J. 36 (8) (2015) 501–508.
[57] M.J. Budoff, D. Dowe, J.G. Jollis, et al., Diagnostic performance of 64-multidetector row coronary computed tomographic angiography for evaluation of coronary artery stenosis in individuals without known coronary artery disease: results from the prospective multicenter ACCURACY (Assessment by Coronary Computed Tomographic Angiography of Individuals Undergoing Invasive Coronary Angiography) trial, J. Am. Coll. Cardiol. 52 (21) (2008) 1724–1732.
[58] H.J. Chang, F.Y. Lin, S.E. Lee, et al., Coronary atherosclerotic precursors of acute coronary syndromes, J. Am. Coll. Cardiol. 71 (2018) 2511–2522.