Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

CAD-RADS scoring of coronary CT angiography with Multi-Axis Vision Transformer: A clinically-inspired deep learning pipeline

Alessia Gerbasi [a,*], Arianna Dagliati [a], Giuseppe Albi [a], Mattia Chiesa [b], Daniele Andreini [b,c], Andrea Baggiano [b,d], Saima Mushtaq [b], Gianluca Pontone [b,e], Riccardo Bellazzi [a,f,1], Gualtiero Colombo [b,1]

[a] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 5, Pavia, Italy
[b] Centro Cardiologico Monzino IRCCS, Milan, Italy
[c] Department of Biomedical and Clinical Sciences, University of Milan, Milan, Italy
[d] Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy
[e] Department of Biomedical, Surgical and Dental Sciences, University of Milan, Milan, Italy
[f] IRCCS Istituti Clinici Scientifici Maugeri, Pavia, Pavia, Italy

ARTICLE INFO

Keywords:
Coronary Artery Disease
CAD-RADS
Deep Learning
Max-ViT
Explainable AI

ABSTRACT

Background and objective: The standard non-invasive imaging technique used to assess the severity and extent of Coronary Artery Disease (CAD) is Coronary Computed Tomography Angiography (CCTA). However, manual grading of each patient's CCTA according to the CAD-Reporting and Data System (CAD-RADS) scoring is time-consuming and operator-dependent, especially in borderline cases. This work proposes a fully automated, and visually explainable, deep learning pipeline to be used as a decision support system for the CAD screening procedure. The pipeline performs two classification tasks: firstly, identifying patients who require further clinical investigations and secondly, classifying patients into subgroups based on the degree of stenosis, according to commonly used CAD-RADS thresholds.

Methods: The pipeline pre-processes multiplanar projections of the coronary arteries, extracted from the original CCTAs, and classifies them using a fine-tuned Multi-Axis Vision Transformer architecture. With the aim of emulating the current clinical practice, the model is trained to assign a per-patient score by stacking the bi-dimensional longitudinal cross-sections of the three main coronary arteries along channel dimension. Furthermore, it generates visually interpretable maps to assess the reliability of the predictions.

Results: When run on a database of 1873 three-channel images of 253 patients collected at the Monzino Cardiology Center in Milan, the pipeline obtained an AUC of 0.87 and 0.93 for the two classification tasks, respectively.

Conclusion: According to our knowledge, this is the first model trained to assign CAD-RADS scores learning solely from patient scores and not requiring finer imaging annotation steps that are not part of the clinical routine.

1. Introduction

Coronary artery disease (CAD) is the leading cause of cardiovascular mortality worldwide. It is caused by *atherosclerosis*, a phenomenon consisting in the formation of plaques that gradually narrow the diameter of the arteries reducing the oxygen-rich blood supply to the heart. The disruption of these plaques is the main cause of the acute coronary syndromes (i.e., unstable angina, myocardial infarction) [1]. Although the underlying causes for the development of atherosclerosis are not yet fully understood, there are numerous risk factors that can lead to accelerated plaque formation, such as high low-density lipoprotein cholesterol (LDL) level, high blood pressure, diabetes mellitus, smok-

[*] Corresponding author.
E-mail address: alessia.gerbasi01@universitadipavia.it (A. Gerbasi).

[1] Co-last authors.

https://doi.org/10.1016/j.cmpb.2023.107989
Received 26 August 2023; Received in revised form 10 November 2023; Accepted 17 December 2023

Available online 21 December 2023

ing, obesity, advancing age and family predisposition. For these reasons, identification and monitoring of patients at high risk of CAD through noninvasive procedures are of utmost importance.

Cardiac computed tomography angiography (CCTA) allows noninvasive identification of coronary stenosis and high-risk plaque features, which is useful for risk stratification. Recently, CCTA has been integrated into the routine clinical management of patients with suspected CAD because of its value as an effective rule-out tool. Imaging software included in the most commonly used CT scanners, usually includes multiplanar reconstruction (MPR) and straightened curved planar reformation (CPR) methods [2] able to trace the blood vessels and generate 2D longitudinal cross-sections from the CCTA scans. These representations are extremely useful to better visualize the vessel of interest without surrounding structures that could make it harder for the physician to identify the plaque from the original three-dimensional scan.

With the aim of creating a standardized method to communicate imaging findings, CAD-Reporting And Data System (CAD-RADS) scoring has been proposed and it is currently used in the clinical practice [3]. According to CAD-RADS classification system, each patient can be classified with a score ranging from 0 to 5. A score of 0 indicates absence of CAD; 1 corresponds to stenosis between 1-24%; 2 to stenosis between 25-49%; 3 to stenosis between 50-69%; 4 to stenosis between 70-99% or >50% left main or three vessels >70%; 5 to total occlusion. One of the main limitations associated with manual scoring of CCTA scans is its dependence on the physician expertise, which can be crucial in borderline cases. On the other hand, automating this process is challenging because the CAD-RADS is a per-patient score and it is assigned on the basis of a visual assessment of the degree of occlusion of the three major coronary arteries: left anterior descending artery (LAD), left circumflex artery (LCX) and right coronary artery (RCA).

In the last years, deep learning (DL) models have been widely explored in the medical field. If correctly designed and evaluated, these methods offer a chance to enhance healthcare accessibility, fairness, precision, and inclusivity [4]. In particular, convolutional neural networks (CNNs) have dominated the field of computer vision in the past years and are still the most widely used models for solving tasks ranging from segmentation to classification. By employing filters, these networks are able to learn feature maps that highlight the most relevant parts of the input images. Vision transformers have recently gained a great popularity in computer vision achieving state-of-the-art (SOTA) performance in many visual tasks [5]. The main advantage of transformer architectures is their attention mechanism. However, when compared to classical CNNs, their reduced inductive bias can easily lead to overfitting. This is the reason why their superiority to convolutional models is generally appreciable when large data sets are available. In the medical context, this is almost never the case since several factors ranging from data privacy to heterogeneity and lack of standard quality, make it difficult to create large and consistent datasets. Therefore, simpler models to mitigate the risk of overfitting and to facilitate output interpretation are usually preferred. This can often lead to sub-optimal results, given the well-known limitations of many of the most popular convolutional models compared to the most recently proposed ones. Many improvements of standard convolutions have been recently proposed to make them more efficient (e.g. [6,7]). On the other hand, many recent works have tried to improve scalability of attention mechanisms (e.g. [8]), or to propose hybrid methods such as [9,10]. In particular, Tu et al. [11] recently proposed MaxViT, a hybrid model combining the strengths of both approaches (efficient convolutions and sparse attention) in a new "base-block" able to significantly improve upon SOTA performance *under all data regimes* for many visual tasks, including image classification. This new base-block consists of a MB-Conv block [6] with a squeeze-and-excitation (SE) module [12] followed by a multi-axis attention block appositely designed to capture both local and global pixels interactions.

In the light of these considerations, this paper proposes a DL pipeline for the automatic classification of straightened MPR images obtained from CCTA scans, based on MaxVit architecture. The goal of the study is to automate the CAD-RADS scoring process with an explainable decision support system able to (1) rule-out patients needing for further investigations and (2) classify the patients into three main groups based on the degree of stenosis, according to commonly used CAD-RADS thresholds. Our aim is to build a fully automated DL pipeline able to guide the physician in the clinical practice, providing a tool which is at the same time accurate and easy to interpret by the final user.

The main contributions of this study are:

- The development of a novel fully automated pipeline based on MaxViT architecture specifically trained to assign a patient-based CAD-RADS score. As far as we know this is the first approach tackling the CAD-RADS scoring problem emulating the clinical procedure.
- The design of a flexible approach not requiring vessel, segment or lesion-wise annotations and considering the three main coronary arteries.
- An extensive experimentation on a curated dataset of 253 patients presenting quantitative evaluations and visually explainable results.

Code is available at https://github.com/ales-git/DeepCADRADS.

2. Related works

Several different approaches have been proposed with the aim of automating the identification and grading of coronary stenosis. We report the most recently proposed works that exploit DL methods.

Huang et al. [13] showed that there is no significant difference between the DL-based (convolutional models in this case) and the expert-based CAD-RADS grading of CCTAs (Kappa value of 0.77). This result is very interesting from a clinical perspective because it suggests the high potentiality of DL based decision support systems for this particular clinical task.

Li et al. [14] developed a coronary tree segmentation algorithm (Dice score 0.771) and proposed a binary classification algorithm (3DNet) taking as input the segmented tree and other relevant clinical features, with the aim of predicting patient-wise CAD-RADS score achieving a diagnostic performance in terms of area under the ROC curve of 0.737.

Denzinger et al. [15] proposed a DL strategy that reaches a ROC AUC of 0.923 on the task of identifying patients with a CAD-RADS score > 2 that was then improved in a more recently proposed version to 0.950 [16]. The proposed method is very promising and tested on a large cohort of patients, but requires segment-level annotations, a step that is not usually part of clinical routine.

Other works focused instead on single lesion or single vessel scoring automated systems. However, deriving patient scores based on individual lesions can lead to a significant number of potential errors, as it fails to take into account the overall context in making decisions. Paul et al. [17] for example, achieved a 96% accuracy in identifying significant stenosis from a huge dataset of curved multiplanar reformatted (cMPR) CCTA images originally classified by an expert radiologist. However, the single vessel grading strategy is time consuming, highly influenced by the radiologist expertise and not usually part of clinical routine.

Penso et al. [18] proposed a token-mixer architecture for CAD-RADS classification achieving 82% of accuracy in classifying significant stenosis and 72% in a multi-class experimental set-up predicting CAD-RADS 0 vs. 1–2 vs. 3–4 vs. 5. Even in this case, each coronary artery was individually labelled.

Tejero-de Pablos et al. [19] proposed a model leveraging multiple feature extractors for texture classification using multiple CPR views of the coronary arteries. The method shows good performance in predicting significant stenosis on a dataset of 57 patients. Although the limited dataset size and the need for manual ground-truth annotation of the stenosis, they achieved 80% of accuracy using a leave-one-out cross-validation strategy.

Candemir et al. [20] proposed a 3D-CNN obtaining good performance for coronary artery atherosclerosis detection on MPR volumes. The algorithm uses pre-processing techniques and a 3D-CNN to identify atherosclerotic plaques and provides visual clues for location. The method obtains an accuracy of 90.9% in identifying patients with atherosclerosis. The authors proposed it as a method for assisting physicians in excluding coronary atherosclerosis in patients with acute chest pain.

Muscogiuri et al. [21] demonstrated how DL methods for CAD-RADS scoring are significantly faster if compared to human on site reading of clinical scans. They proposed several small custom 2D-CNN models with the best one achieving an accuracy of 81% in classifying patients with CAD-RADS 0 vs. CAD-RADS >0. The models were trained using single 2D slices from original CCTA scans without extracting the coronary arteries.

A recurrent CNN was proposed by Zreik et al. [22] for automatic detection and classification of coronary artery plaque and stenosis, achieving an accuracy of 0.77 and 0.80, respectively, on a test set of 65 patients. In this last work the presence and the anatomical significance of coronary stenosis were manually annotated on the MPR images.

Most of the discussed works present interesting different approaches from both technical and clinical perspective, however they all require vessel, segment, lesion annotations or derive patient-scores considering single lesions. While these approaches include more information if compared to a single patient-wise score, they require an additional effort in the clinical practice since segment or lesion-wise annotation is not a routine operation. Moreover, they are not able to take into account the global patient status. Several methods have been proposed to tackle the challenge of limited medical imaging annotations based on semi-supervised or weakly-supervised learning (e.g. [23,24]). However, in our case the greatest challenge lies in creating a system that can comprehensively analyze multiple MPR views of coronary arteries and assign patient-level scores without relying on finer annotations. To achieve this objective, unlike the previously mentioned methods, we developed an innovative pipeline based on a recently introduced neural architecture. Our intention is to mirror the real clinical approach, where an expert physician visually examines multiple MPR views of the three primary coronary arteries and assigns a patient-level score, bypassing additional annotation steps.

3. Materials and methods

We set up two different experiments following the same analysis pipeline. In the first one (named *binary experiment*) we binarized the CAD-RADS score with a threshold of 2 (0-1-2 vs. 3-4-5) with the aim of simply distinguishing patients in need for further examinations or direct intervention (CAD-RADS > 2). In the second experiment (named *multi-class experiment*) we trained the model to predict 3 different classes: healthy subjects (CAD-RADS = 0), patients with minimal to moderate stenosis (CAD-RADS = 1-2-3) and patients with severe stenosis or complete occlusion (CAD-RADS = 4-5). This second approach would be useful to quickly identify completely healthy subjects as well as patients with very severe stenosis, grouping the intermediate or borderline cases that could need a more accurate inspection from the physicians. The complete pipeline is illustrated in Fig. 1, each step is described in the following subsections.

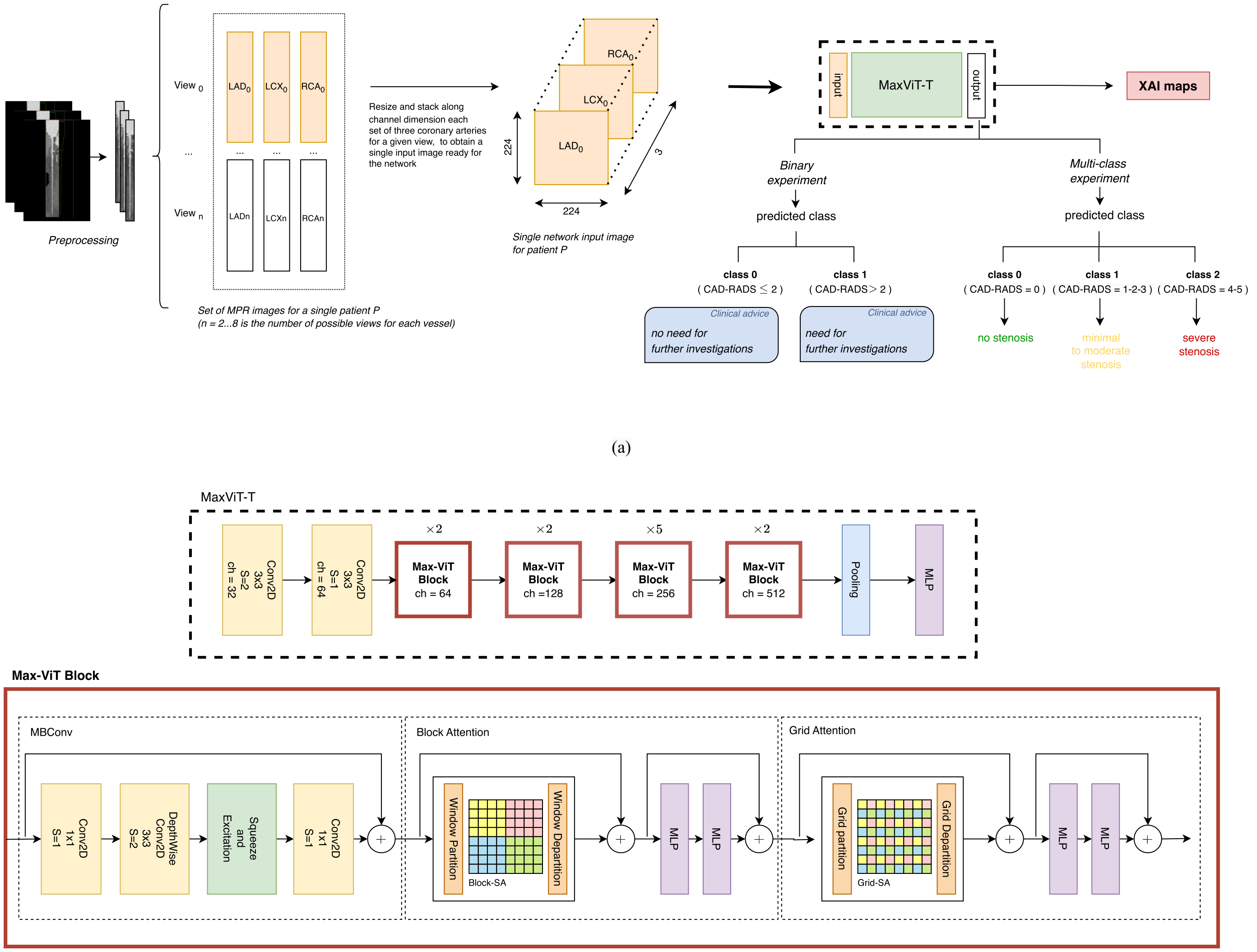
3.1. Dataset

We applied our novel pipeline to a dataset of 253 patients who underwent CCTA for clinical purposes from 2016 to 2018 in the Monzino Cardiology Center (Milan, Italy). The study protocol conformed to the principle of the Declaration of Helsinki and was approved by the “Ethics Committee of the IRCCS Istituto Europeo di Oncologia and Centro Cardiologico Monzino” (protocol code R329/15 - CCM 341, date of approval 23/09/2015). All recruited patients signed the written informed consent and participants did agree to share their de-identified information. The original population is fully described by Muscogiuri et al. [21]. Exclusion criteria for this study were heart rate ≥ 80 bpm despite intravenous administration of beta blockers, atrial fibrillation, BMI ≥ 35 kg/m^2 [25] and presence of stent. Sublingual nitrates were administered 5 minutes before the CCTA scan [26]. Two CT scanners, the Discovery CT 750 HD and Revolution CT (GE Healthcare, Milwaukee, IL), were used for CCTA acquisition. The CCTA protocol defines a 64×0.625 mm and a 256×0.625 mm slice configuration for the Discovery CT 750 HD and the Revolution CT, respectively. The tube current and voltage were adjusted based on the patient’s BMI [27]. In both protocols, 50–70 mL of contrast medium was given through the antecubital vein at an infusion rate of 5 mL/s, followed by 50 mL of saline solution at the same rate. The bolus tracking technique was used for CCTA acquisition, and images were reconstructed using filtered back projection and in 75% or 40–80% of the cardiac cycle, depending on the ECG-triggering acquisition used [28]. In cases of poor image quality, intracycle motion correction was performed [29,28]. A consensus of five different random couples between ten radiologists and cardiologists was formed to score the pool of CCTA examinations. The cardiac imagers had experience ranging from 5 to 10 years. A CAD-RADS score was attributed for each examination, and in cases of disagreement, a cardiac imager with 10 years of experience in cardiovascular imaging adjudicated the final CAD-RADS score.

For each patient and each coronary artery (RCA, LCX, LAD) up to eight straightened MPR views were extracted from the original CCTA, with a 45° angle offset. In particular, if the subject was classified with CAD-RADS = 0 (0% stenosis), we always had exactly 8 images for each main coronary artery in our dataset. On the other hand, for patients with CAD-RADS > 0, only images from non-healthy coronaries were collected (e.g. if a patient has a CAD-RADS score of 3 and only the LAD and LCX arteries present with stenosis, we have up to 8 views for each of these two vessels, but no images for the healthy RCA). Therefore, in our dataset, completely healthy coronary arteries, which do not influence the CAD-RADS score when this is greater than 0, were discarded a priori by the clinicians at the data acquisition stage. Our aim is to fully automate the process and train an algorithm to predict the CAD-RADS score from the three main coronary arteries without any prior knowledge. For this reason, the missing healthy vessels were imputed before the classification step as fully described in Section 3.3. In Fig. 2 we can see an example CCTA for a patient included in our dataset. We show two slices of the original 3D DICOM scan where it is possible to notice the 3 main coronary arteries in a bi-dimensional space and, on the right, we can see the LAD, LCX and RCA images resulting from the straightened curved planar reconstruction.

3.2. Preprocessing

The first step of our pipeline is image preprocessing. Our input data is composed by 2D images representing different views of the straightened MPR volume for each patient. In the first step we removed artifacts on digital scans by binarizing each image, sorting white objects by size and keeping just the largest one (representing the vessel). Therefore, we made sure to delete all annotations and small artifacts derived from image reconstruction. Finally, we applied Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance the local contrast of the image [30]. As a final step we automatically cropped the images to reduce background black pixels on the 4 sides. A sample image before and after the preprocessing steps is showed in Fig. 3.



(a)

(b)

Fig. 1. (a) Schematic representation of the main pipeline steps. All the MPR projections are first of all preprocessed to enhance the image contrast. Afterwards, for each patient, an imputing step is performed in order to always have three 2D images representing LAD, LCX and RCA for each of the n different views (where $n = 2 \ldots 8$ is the number of possible views for each vessel). For each view, the 3 images are then resized (224x224) and stack along channel dimension in order to obtain a single input image ready for the network. The images thus created are used to fine-tune a MaxViT-T architecture to solve two different tasks (binary and multi-class). Finally, SOTA eXplainable AI (XAI) models are used to create qualitative maps to visually inspect the reliability of network's predictions. (b) Max-ViT-Tiny architecture used in the proposed pipeline. It is composed by two convolutional blocks, followed by several MaxViT blocks and a final pooling layer that precedes the MLP head. The architecture of each MaxViT block is schematized in the figure: there is an initial MBCov block followed by a block-attention and a grid-attention block.

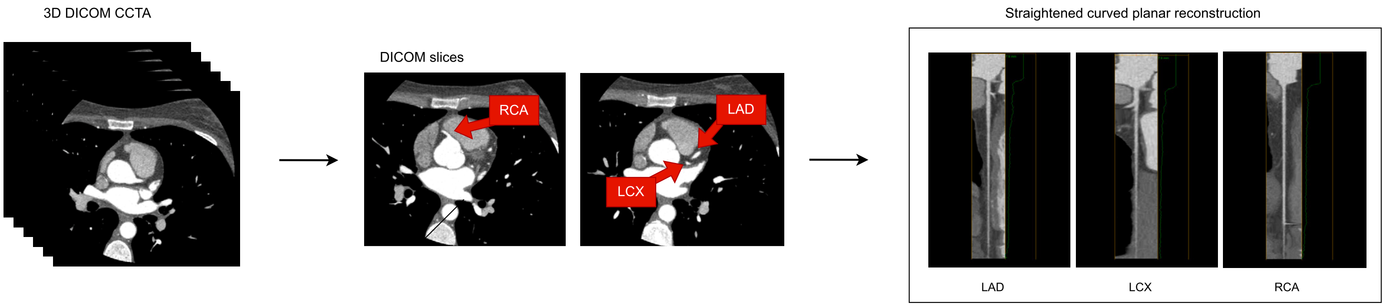


Fig. 2. Example of CCTA scan for a random patient included in the study population. From the 3D DICOM scan we extracted two slices where the three main coronary arteries are indicated in red. On the right we can see the 2D representation of LAD, LCX and RCA obtained through straightened curved planar reconstruction.

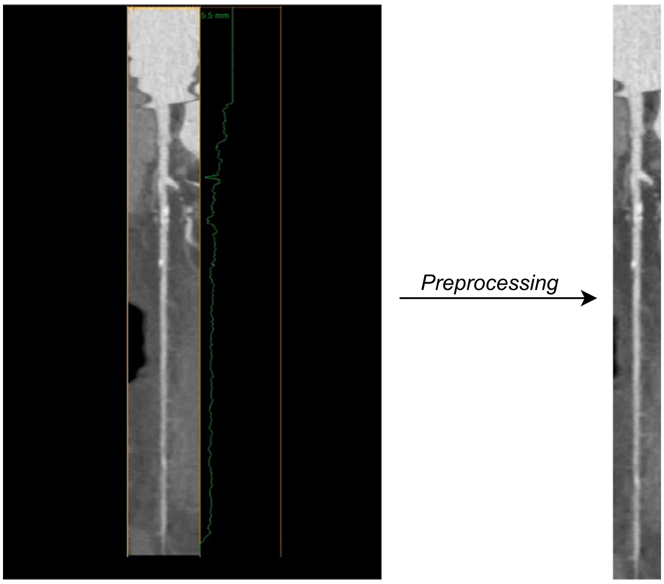


Fig. 3. Sample image before and after the preprocessing steps: (1) Annotation/artifacts removal, (2) Contrast enhancement (CLAHE), (3) Background crop.

Algorithm 1 Data preparation.

```
Input: patient_ids, dataset   ▷ Dataset represents the whole set of straightened MPR images.
Support functions:
remove_artifacts($x$): Binarize image and keep largest white object representing the vessel.
CLAHE($x$): Split the image into small tiles and apply Contrast Limited Adaptive Histogram Equalization.
crop_background($x$): Reduce black pixels on the four image sides.
imputing($v$): Impute missing image with average healthy coronary computed from control patients in the training set.

for all $x$ in dataset do
    $x$ = remove_artifacts($x$)
    $x$ = CLAHE($x$)
    $x$ = crop_background($x$)
end for

for all $patient\_id$ in patient_ids do
    $view\_pat = dataset[patient\_id]$
    for all $v$ in $view\_pat$ do ▷ v = [LAD,LCX,RCA] where one or two may be None if completely healthy
        if len($v$) < 3 then
            $v$ = imputing($v$)
        end if
        $v$ = resize(stack($v$))                ▷ $v \in \mathbb{R}^{3\times224\times224}$
    end for
end for
```

3.3. Dataset split and imputing

After the preprocessing, data was randomly split into training (80%) and test (20%) set. The training set was then further split into training and validation set using a 10-fold cross-validation strategy. All the splits were always done patient-wise and stratified by CAD-RADS score in order to avoid any selection bias. Images from completely healthy vessels for patients with CAD-RADS > 0 were missing in our dataset as previously described in Section 3.1. Therefore, we averaged (separately for each view) the images of RCA, LCX and LAD coronaries of the healthy subjects (CAD-RADS $= 0$) included in the training set and used them to impute the missing data in our dataset. This step allowed us to always have three coronary arteries to evaluate, to simulate as closely as possible the decision-making process actually followed in clinical practice.

3.4. Model architecture and training strategy

Given the limited size of our dataset we decided to fine-tune the MaxViT-T (where T stands for tiny) pretrained on ImageNet [31]. This recently proposed architecture combines the strengths of efficient convolutions and attention mechanism achieving SOTA classification performance under all data regimes. The architecture details are shown in Fig. 1b. It is composed by 2 convolutional blocks followed by multiple groups of MaxViT blocks. A single MaxViT block is always composed by an initial MBConv block, also called inverted residual block, that follows a narrow → wide → narrow structure approach which greatly reduces the number of parameters if compared to a standard residual block [7]. As it is shown in the figure, between the 3×3 depthwise convolution and the final 1×1 convolution, there is a SE module [12] able to model interdependencies between channels. Following the MBConv block, we always have a block-attention and a grid-attention block which are used to capture local and global patterns respectively. Each group of MaxViT blocks differs from the other by the number of spatial filters in the convolutional layers. Finally, an average pooling layer precedes the MLP head used to classify the input into 2 (binary experiment) or 3 classes (multi-class experiment).

The relative attention function is defined as:

$$\sqrt{\text{RelAttention}(Q,K,V)} = \text{softmax}\left(\frac{QK^T}{d} + B\right)V, \tag{1}$$

where $Q, K, V \in \mathbb{R}^{(H\times W)\times C}$ are the query, key, and value matrices, and d is the hidden dimension. Here, attention weights are determined by a learned static location-aware matrix B and the scaled input-adaptive attention.

The Multi-Axis Attention includes operations such as Block, Unblock, Grid, and Ungrid. The Block operation partitions the input image into non-overlapping blocks of a specified size. After this partitioning, the block dimensions are rearranged onto the spatial dimension. Unblock is the inverse operation of the Block procedure. It reconstructs the original input from the partitioned blocks, essentially reversing the process of the Block operation. The Grid operation divides the input feature into a uniform grid structure. It arranges the input into a grid format, allowing for the processing of smaller grid cells or units within the larger feature space. Ungrid is the reverse operation of the Grid procedure. It reverts the gridded input back to the original feature space, enabling the continuation of computations in the standard feature format. Therefore given an input tensor $x \in \mathbb{R}^{H\times W\times C}$ local Block Attention is formulated as:

$$x \leftarrow x + \text{Unblock}\left(\text{RelAttention}(\text{Block}(LN(x)))\right) \tag{2}$$

$$x \leftarrow x + \text{MLP}(LN(x))$$

and global, dilated Grid Attention as:

$$x \leftarrow x + \text{Ungrid}(\text{RelAttention}(\text{Grid}(LN(x)))) \tag{3}$$

$$x \leftarrow x + \text{MLP}(LN(x))$$

where, LN represents Layer Normalization and MLP is a standard Multi-Layer Perceptron network [11].

To further reduce the risk of overfitting, we implemented online data augmentation on the training set with random rotation and horizontal/vertical flip, learning rate and weight decay, and label smoothing. The optimal parameters were tuned with a grid search strategy based on average validation set accuracy. A complete list of the parameters used is available in Section 3.5. As showed in Fig. 1, after the preprocessing and imputing steps previously described, for each view, the 3 images (representing LAD, LCX and RCA respectively) are then resized (224x224 pixels) and stack along channel dimension in order to obtain a single input tensor ready for the network. This approach allows us to train the network to classify a sequence of three coronary arteries as belonging to one of the classes representing patient-wise CAD-RADS

Table 1
Set of model's parameters leading to the best average validation accuracy.

Experiment	Lr	Lr decay epoch	Drop-out	L2	Label smoothing	Epochs	Batch size	Optimizer	Loss
Binary	$1e^{-4} \rightarrow 1e^{-5}$	30	0.5	0.1	0.1	50	8	AdamW	BCE
Multi-class	$1e^{-4} \rightarrow 1e^{-5}$	30	0.3	0.1	0.2	50	8	AdamW	wCE

Table 2
Results of the *binary* and *multi-class experiment* computed on the test set. For all the metrics 95% confidence interval is provided.

Experiment	Class	Type of metric	AUC [95% CI]	Accuracy [95% CI]	Precison [95% CI]	Recall [95% CI]	F1-score [95% CI]	n
Binary	1	per image	**0.89 [0.86, 0.93]**	0.82 [0.78, 0.86]	0.88 [0.83, 0.93]	0.75 [0.68, 0.80]	0.81 [0.76, 0.84]	374
		per patient	**0.87 [0.76, 0.95]**	0.82 [0.72, 0.93]	0.89 [0.75, 1.00]	0.71 [0.55, 0.89]	0.79 [0.69, 0.90]	51
Multi-class	0	per image	0.93 [0.84, 0.96]	0.94 [0.92, 0.96]	0.90 [0.83, 0.97]	0.82 [0.75, 0.90]	0.86 [0.82, 0.90]	80
		per patient	0.94 [0.81, 0.99]	0.96 [0.91, 1.00]	1.00 [1.00, 1.00]	0.80 [0.60, 1.00]	0.89 [0.80, 0.98]	10
	1	per image	0.87 [0.84, 0.95]	0.82 [0.78, 0.86]	0.79 [0.73, 0.85]	0.86 [0.81, 0.91]	0.82 [0.78, 0.86]	182
		per patient	0.91 [0.83, 0.98]	0.84 [0.74, 0.94]	0.81 [0.67, 0.95]	0.93 [0.83, 1.00]	0.87 [0.78, 0.96]	27
	2	per image	0.91 [0.84, 0.96]	0.87 [0.84, 0.90]	0.82 [0.75, 0.89]	0.75 [0.67, 0.83]	0.78 [0.74, 0.82]	112
		per patient	0.93 [0.84, 0.99]	0.88 [0.79, 0.97]	0.83 [0.62, 1.00]	0.72 [0.51, 0.95]	0.77 [0.68, 0.97]	14
	Weighted avg	per image	**0.90 [0.86, 0.92]**	0.86 [0.83, 0.89]	0.82 [0.78, 0.86]	0.82 [0.78, 0.86]	0.82 [0.78, 0.86]	374
		per patient	**0.93 [0.84, 0.99]**	0.88 [0.79, 0.95]	0.85 [0.76, 0.93]	0.84 [0.74, 0.92]	0.84 [0.73, 0.92]	51

scores and thus simulating the classification process followed in clinical practice.

All the operations described thus far, performed on the data before the model training, are summarized in Algorithm 1.

3.5. Experimental setup

After the preprocessing steps we obtain a total number of 5619 one-channel images, and consequently 1873 three-channel images for 253 patients. Models were trained for 50 epochs with a batch size of 8 using AdamW optimizer [32] for both the experiments, while binary cross-entropy (BCE) and weighted cross-entropy (wCE) loss were used for the binary and multi-class experiment, respectively.

Table 1 summarizes the aforementioned settings along with the best hyperparameters resulting from the grid-search according to the average validation accuracy during the cross-validation procedure. The tuned hyperparameters are: learning rate (LR) $\in \{1e^{-3}, 1e^{-4}, 1e^{-5}\}$, Drop-out $\in \{0.1, 0.3, 0.5\}$, weight decay (L2) $\in \{1e^{-1}, 1e^{-2}\}$, LR decay epoch $\in \{20, 30\}$, label smoothing $\in \{0.1, 0.2\}$.

The best hyperparameters are then used to train the models on the whole training set and evaluate the performance on the test set.

3.6. Statistical comparisons

The performance of the selected architecture was compared against several other fully convolutional (ResNet18, ResNet50 [33]; Vgg16, Vgg19 [34]), attention-based (ViT-T [5]) or hybrid models (ConvNeXt-T [35], CoAtNet-T [9]). To rigorously assess these models, pairwise comparisons were conducted using the De Long test for ROC AUC with a significance threshold set at 0.05. This method aids in evaluating whether the proposed model statistically outperforms the others. All the results are detailed in Section 4, providing valuable insights into the relative performance of the models in handling the specific task at hand.

3.7. Hardware and software

All the models were trained using an NVIDIA 3070 Ti GPU. The whole pipeline is developed in Python 3.10. Pytorch [36] is the framework used for implementing the models. Open-CV implementation of CLAHE algorithm was used for enhancing the contrast in the preprocessing step [37] and DeepSHAP [38] library for the final explainability maps.

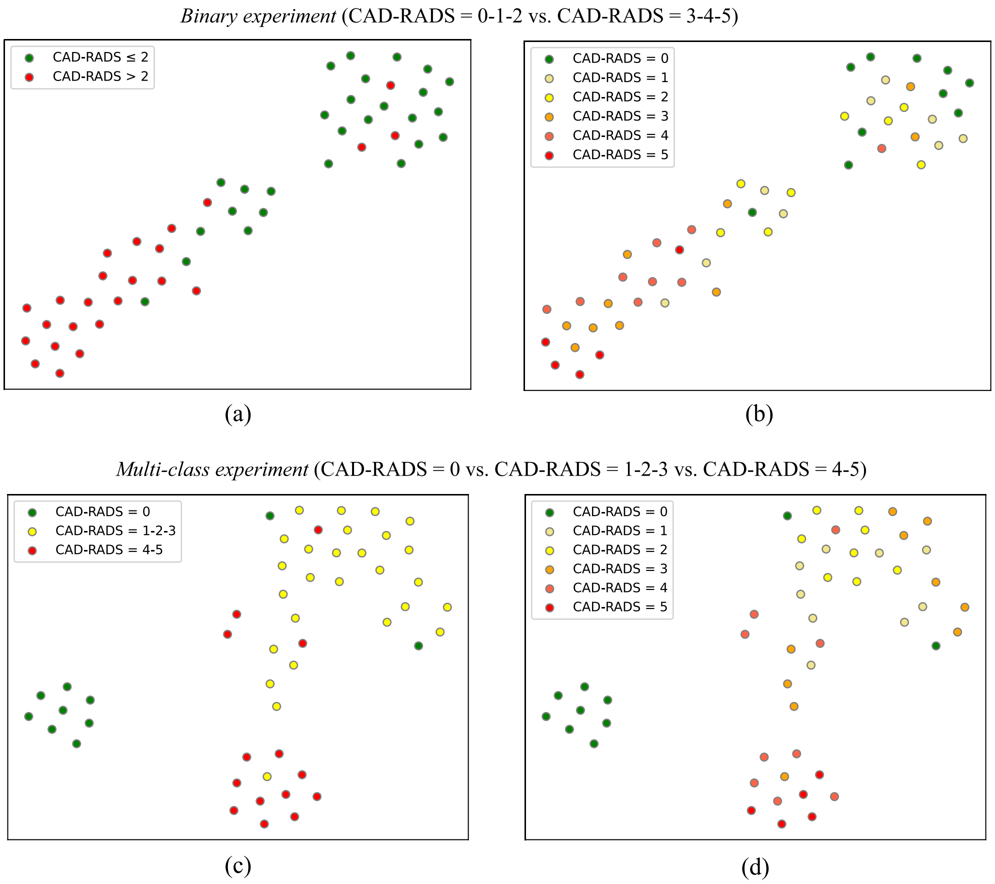
3.8. Visual interpretation

One of the main drawbacks of DL models is their difficult interpretability, which has been tackled with explainability approaches [39,40]. Furthermore, it is of foremost importance to evaluate models reliability and assure systems trust [41], especially in the medical field, where final users might not be aware of the methodological implementation of the supporting system. The reliability of a decision support system strongly increases when it becomes easily interpretable by the final user, who, based on his experience, could simply detect edge cases in which the algorithm results may not be trustable. With this aim in mind, we added three levels of visual explainability of the results:

- **t-SNE** [42] plot to project the last layer features down to a 2D space in order to evaluate the semantic understanding of the network.
- **maximally activated patches** [43] representing the most responsive areas of an image, which are identified through a forward pass with that image that is partially "occluded". This is accomplished by masking portions of the image (patches) and evaluating the impact on the predicted scores for the top class. The patches that cause the greatest change in the scores are prioritized and the top-k of these patches are then visualized.
- **Deep SHAP** [38] explainability maps. This algorithm is a fast approximation method for computing SHAP (SHapley Additive exPlanations) values in DL models. It is a game theoretic approach that generates visual maps for each image where pink pixels indicate the image values that contributed to the model's prediction of a specific output class, while blue pixels represent the values that pushed the prediction towards the alternative class. This visual representation allows for inspection of pixels that were most significant in determining the final classification, according to the model. This approach is particularly interesting since it has been shown to align better with human intuition compared to other explainability methods.

4. Results

Final results on the test set for both experiments are provided in Table 2. Algorithms performance was measured in terms of area under the ROC curve (AUC), accuracy, precision, recall, F1 score. All the results are presented both image-level and patient-level. In the second case, we considered all the images from the same patient (representing different



Binary experiment (CAD-RADS = 0-1-2 vs. CAD-RADS = 3-4-5)

(a) (b)

Multi-class experiment (CAD-RADS = 0 vs. CAD-RADS = 1-2-3 vs. CAD-RADS = 4-5)

(c) (d)

Fig. 4. t-SNE plots showing a 2D representation of the 256 features representing the average patient embeddings. Results for the binary and multi-class experiment are reported in the first (a, b) and second (c, d) row, respectively. Each dot represents a patient of the test set. On the left side we colored the dots based on the labels used for training the network (a, c), on the right side instead, they are colored according to the original CAD-RADS classification (b, d).

views of the same vessels) and we assigned as final class the one with higher average predicted probability.

In Fig. 4 we reported per patient t-SNE plots of the binary and the multi-class experiment in the first (a, b) and second row (c, d), respectively. The dots are colored by the labels used by the algorithm on the left (a, c) and the original CAD-RADS scores on the right (b, d), to visually inspect the extent of the errors.

Finally, Fig. 5 shows an example of maximally activated patches and Deep SHAP maps generated for a random patient from the test set of the multi-class experiment. The selected patient has an original CAD-RADS score of 5, therefore it belongs to class 2 in our multi-class experiment. In the figure we reported LAD, LCX and RCA for one of the 8 views available for this patient. These three images actually compose a single input image for our network that was correctly classified as representing a patient of class 2 with a probability of 0.88. For each vessel, we can see on the left the top-3 maximally activated patches in order of importance. The yellow dashed areas highlight the regions of the vessel represented by the patches that have a 80x80 pixels dimension and are extracted from the resized image given as input to the network during the forward pass. On the right of each vessel instead, we can see the map produced by Deep SHAP algorithm for the correctly predicted class (class 2 in this case). The output of the model is influenced by pink pixels in a positive way and by blue pixels in a negative way. The input images are shown as nearly transparent grayscale backings behind each of the explanations. On the right of each map, we can see a zoom of the most relevant pixels (red dashed areas).

We also reported in Table 3 the resulting per-image AUC and accuracy for both the experiments comparing the performance of the most common convolutional, attention-based and hybrid architectures. DeLong's test p-values for pairwise AUC comparisons ($\alpha = 0.05$), with the highest AUC as a reference, are provided. For each model the number of parameters and multiply-accumulate operations (MACs) is reported. ROC curves for each experiment and each model are provided in Fig. 6.

4.1. Ablation studies

Two ablation studies were conducted to assess the impact of (I) using multiple straightened CPR from different angles and (II) the number of MaxViT blocks at each stage on the overall performance. Both studies were specifically performed in the multi-class setting.

In the first experiment, we aimed to significantly reduce the maximum number of per-patient projections included, exploring configurations with 8 down to 4 and then 3 projections. These different configurations are summarized in Fig. 7. This experiment was designed to explore the model's capacity to learn general characteristics even when exposed to a limited number of view angles. To ensure comparable sample sizes, the per-patient AUC and accuracy on the test set (n = 51) are reported (Table 4).

The second experiment focused on evaluating how the final performance is influenced by the specific configuration of MaxViT blocks used. Initially, we reduced the number of blocks in the central stages (stages 1 and 2), while maintaining a constant number of blocks in the two external stages (stages 0 and 3). Subsequently, we explored the opposite approach. Finally, a configuration with reduced blocks in all stages was also examined. The results, expressed in terms of AUC and accuracy with 95% CI, are presented in Table 5, alongside the number of parameters for each novel architectural configuration. A detailed discussion of these results is provided in Section 5.

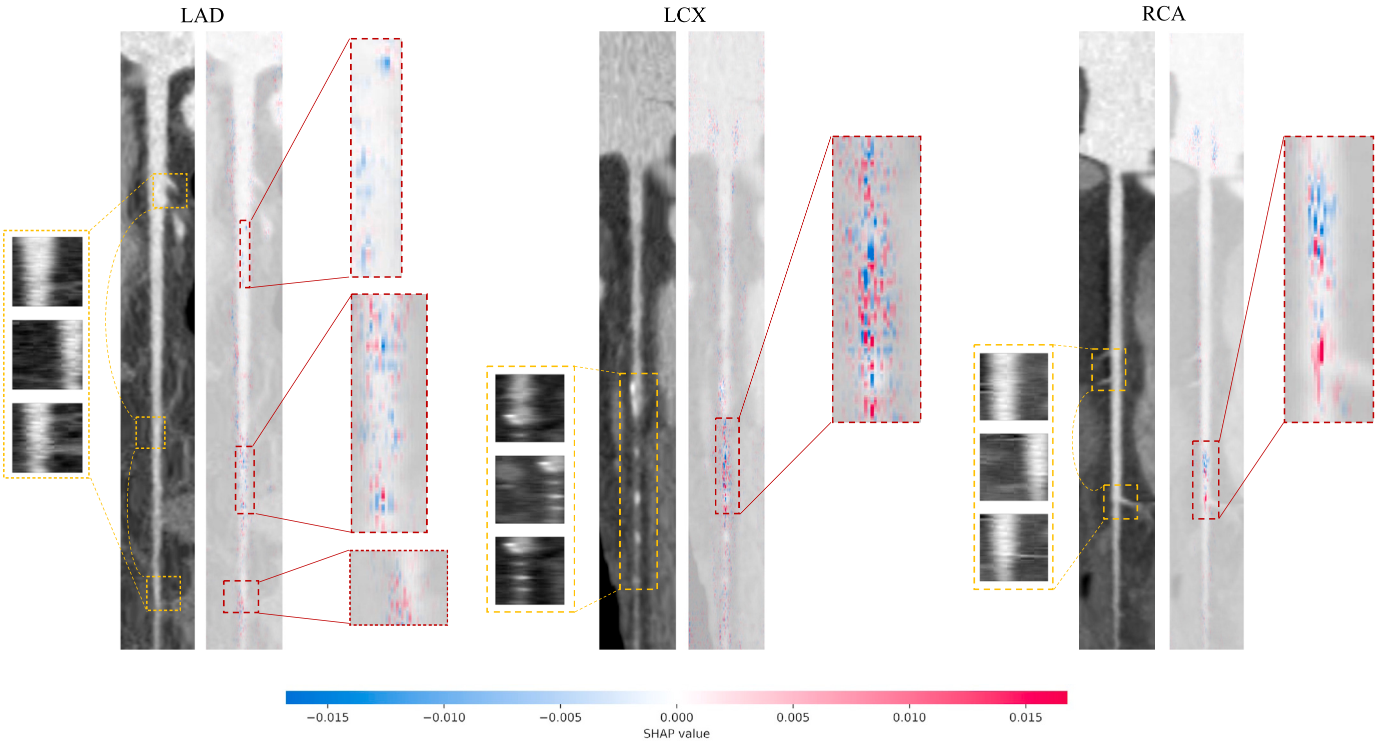
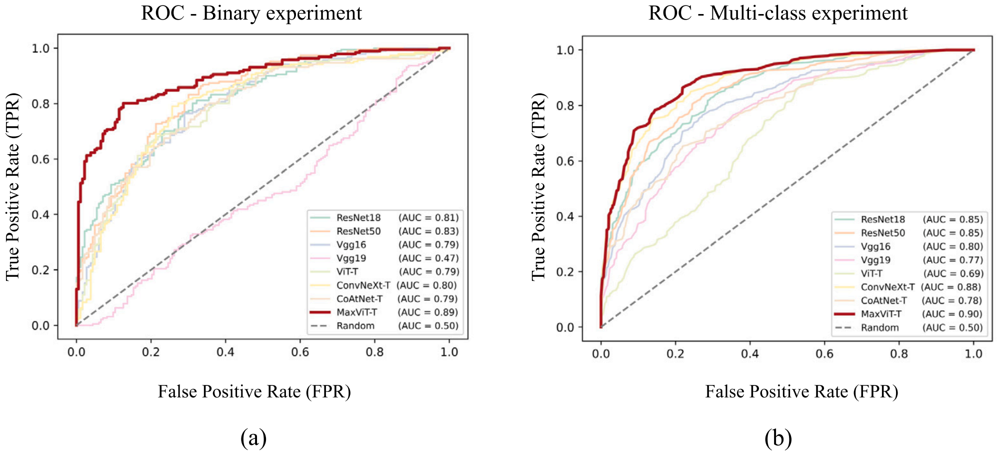


Fig. 5. The figure shows the three main coronary arteries of a patient from the test set with a CAD-RADS score of 5 and correctly classified as belonging to class 2 (CAD-RADS = 4-5) by the multi-class algorithm. The three vessels (LAD, LCX, RCA) represent the three channels of the input tensor pre-processed by the network as a single input image. For each vessel, we can see the original pre-processed scan with the top-3 maximally activated patches in order of importance on the left and the Deep SHAP map on the right, with a zoom on the most relevant part of the images. Deep SHAP maps are reported for the correctly predicted output class (class 2). The input images are shown as nearly transparent grayscale backings behind each of the explanations. Pink pixels increase the model's output while blue pixels decrease it.



(a)

(b)

Fig. 6. ROC curves of the compared models for the binary (a) and multi-class experiment (b).

5. Discussion

The proposed pipeline achieved high performance in all the reported metrics for both the binary and multi-class experiment, showing a per-patient AUC [95% CI] of 0.87 [0.76, 0.95] and 0.93 [0.84, 0.99], respectively. To the best of our knowledge, this is the first work classifying the sequence of the three main coronary arteries relying only on patient-wise CAD-RADS scores during the training procedure. None of the previous works are directly comparable to our method as most of them rely on vessel, segment, lesion-wise annotations. We were able to achieve highly accurate results without the need of any additional annotation step. This demonstrates that it is possible to train a model to assign an overall score to a patient's three main coronary arteries without requiring additional clinical annotations, which are not part of the screening routine. The work conceptually most similar to ours is that presented by Li et al. [14]. Although they used a completely different methodological approach, their aim was similar to ours since they trained a model to assign patient-wise CAD-RADS scores based on the whole coronary tree obtaining an AUC of 0.737 for a binary classification task. Compared with works that used finer annotation steps, we might expect our performance to be poorer due to the smaller amount of information used during the learning procedure. Nonetheless, we achieved very high performance in both tasks, often outperforming previous works that used finer annotations. It must be noted, however, that

Table 3
Comparison between different architectures: convolutional (ResNet18, ResNet50, Vgg16, Vgg19), transformer (ViT-T) and hybrid (ConvNeXt-T, CAtNet-T, MaxViT-T) models. For each architecture, per-image AUC and Accuracy with 95% CI are reported. DeLong's test p-values for pairwise AUC comparisons, with the highest AUC as a reference, are also provided. Finally, in the last two columns we reported the number of parameters and MACs of the compared models.

Model	Experiment	AUC [95% CI]	Accuracy [95% CI]	DeLong's p-value	Params(M)	MACs(G)
ResNet18	binary multi-class	0.81 [0.77, 0.85] 0.85 [0.82, 0.88]	0.74 [0.69, 0.79] 0.82 [0.78, 0.85]	< 0.001 < 0.001	11.18	1.82
ResNet50	binary multi-class	0.83 [0.79, 0.86] 0.85 [0.82, 0.89]	0.76 [0.71, 0.81] 0.81 [0.77, 0.84]	< 0.001 0.01	23.51	4.13
Vgg16	binary multi-class	0.79 [0.75, 0.84] 0.80 [0.76, 0.84]	0.72 [0.68, 0.77] 0.78 [0.75, 0.82]	< 0.001 < 0.001	134.27	15.47
Vgg19	binary multi-class	0.47 [0.42, 0.54] 0.77 [0.73, 0.81]	0.51 [0.46, 0.56] 0.76 [0.73, 0.80]	< 0.005 < 0.001	139.58	19.63
ViT-T	binary multi-class	0.79 [0.74, 0.83] 0.69 [0.65, 0.73]	0.72 [0.68, 0.76] 0.71 [0.67, 0.74]	< 0.001 < 0.001	5.49	1.08
ConvNeXt-T	binary multi-class	0.80 [0.75, 0.84] 0.88 [0.85, 0.91]	0.76 [0.72, 0.81] 0.83 [0.80, 0.86]	< 0.001 0.44	27.80	4.45
CoAtNet-T	binary multi-class	0.79 [0.75, 0.83] 0.78 [0.74, 0.82]	0.72 [0.75, 0.83] 0.77 [0.74, 0.81]	< 0.001 < 0.001	5.48	7.64
MaxViT-T	**binary** **multi-class**	**0.89 [0.86, 0.93]** **0.90 [0.86, 0.92]**	**0.82 [0.78, 0.86]** **0.86 [0.83, 0.89]**	-	28.45	4.89

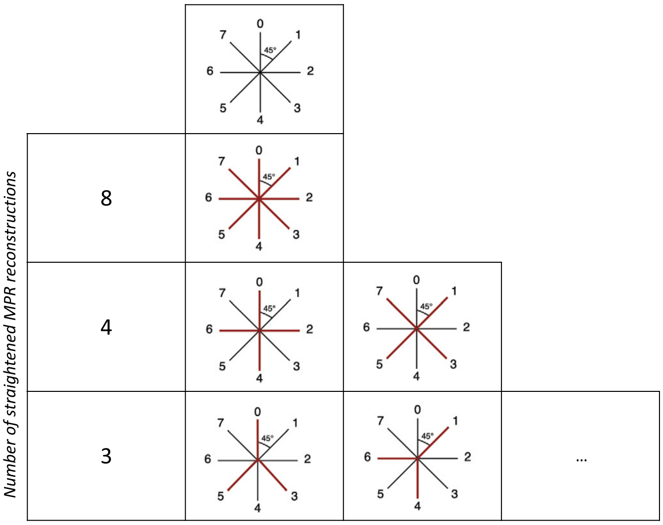


Fig. 7. Projections included in the ablation study. We compared the network trained using maximum 8 available views for each patient, with different versions trained on 4 and 3 views respectively. The projections included in each experiment are highlighted in red.

direct comparison is challenging as the datasets used are not publicly available. The strength of our work primarily relies on the network's ability to autonomously learn complex features, while taking the entire context into account, rather than considering single vessels, segments, or lesions.

From the results shown in Table 3, it is evident that the chosen architecture clearly outperformed all the other methods included in the comparison. In particular, Vgg19, which is the deepest fully convolutional architecture, and ViT-T which is the only vision transformer model, were the worst performing algorithms, showing poor generalization capabilities in both the experiments. The MaxViT architecture excels by leveraging a fusion of global and local receptive fields spanning across the network's entire depth. This integration at every stage of the model's architecture enhances its overall ability to generalize, resulting in superior performance.

Table 4
Ablation study on the number of projections included in the pipeline. We repeated the multi-class experiment several times, including up to 4 or 3 projections for each patient. In this table we provide per-patient multi classification results on the test set (n = 51).

Projections	AUC [95% CI]	Accuracy [95% CI]
0 - 7	**0.93 [0.84, 0.99]**	**0.88 [0.79, 0.95]**
0, 2, 4, 6	0.91 [0.83, 0.98]	0.88 [0.79, 0.94]
1, 3, 5, 7	0.92 [0.88, 0.99]	0.88 [0.80, 0.96]
0, 3, 5	0.89 [0.81, 0.97]	0.85 [0.77, 0.92]
1, 4, 6	0.91 [0.83, 0.97]	0.86 [0.77, 0.94]
2, 5, 7	0.93 [0.86, 0.98]	0.87 [0.79, 0.95]
3, 0, 6	0.91 [0.82, 0.98]	0.88 [0.79, 0.95]
4, 1, 7	0.91 [0.83, 0.97]	0.86 [0.77, 0.94]
5, 2, 0	0.93 [0.89, 0.99]	0.87 [0.79, 0.95]
6, 1, 3	0.93 [0.86, 0.98]	0.84 [0.75, 0.92]
7, 4, 2	0.92 [0.86, 0.98]	0.88 [0.81, 0.95]

For the binary experiment, we can recognize two areas from the t-SNE plots in Fig. 4 (a, b) with the patients in the lower left corner representing subjects with higher CAD-RADS score. For the multi-class experiment (c, d), instead, we can recognize three main groups basically representing patients with zero, mild and severe stenosis. It is interesting to notice that the separation between healthy subjects (green) and patients with high CAD-RADS scores (red) is clear-cut. Interestingly, most of the errors occurred in the intermediate group. This finding reflects the fact that misclassifications in the test set were always off by only one class (e.g., a patient classified as belonging to class 1 when he was actually in class 0, but never misclassified as belonging to class 2). Furthermore, when we examine the plot colored by the original CAD-RADS scores (d), we can observe that misclassifications in class 2 only occurred in patients with CAD-RADS scores lower than 5. This finding is particularly interesting because t-SNE is an unsupervised technique, yet the distribution of patients in the 2D space reflects the expected clusters based on domain knowledge.

In Fig. 5, we can observe both maximally activated patches and higher absolute SHAP values in the most critical regions of the vessels. Regarding the LAD vessel, we can see that most of the pixels are blue in the central part, indicating that, based on those pixels, the pre-

Table 5
Ablation study assessing the impact of individual MaxViT blocks within the model architecture. The results for the original MaxViT-T configuration are presented in the first row. Reduced blocks are highlighted in bold within the first column for all other configurations.

MaxViT blocks (S1, S2, S3, S4)	Params(M)	Type of metric	AUC [95% CI]	Accuracy [95% CI]	n
2, 2, 5, 2	**28.45**	**per image**	**0.90 [0.86, 0.92]**	**0.86 [0.83, 0.89]**	**374**
		per patient	**0.93 [0.84, 0.99]**	**0.88 [0.79, 0.95]**	**51**
2, **1**, **4**, 2	25.72	per image	0.88 [0.85, 0.92]	0.83 [0.80, 0.86]	374
		per patient	0.92 [0.84, 0.98]	0.89 [0.81, 0.96]	51
2, **1**, **3**, 2	23.47	per image	0.89 [0.86, 0.92]	0.83 [0.80, 0.87]	374
		per patient	0.92 [0.83, 0.97]	0.86 [0.78, 0.93]	51
2, **1**, **2**, 2	21.22	per image	0.86 [0.82, 0.89]	0.81 [0.77, 0.84]	374
		per patient	0.90 [0.83, 0.97]	0.83 [0.73, 0.91]	51
1, 2, 5, **1**	19.44	per image	0.89 [0.85, 0.92]	0.85 [0.82, 0.88]	374
		per patient	0.92 [0.84, 0.98]	0.88 [0.79, 0.95]	51
1, 1, 2, 1	12.11	per image	0.88 [0.85, 0.91]	0.83 [0.79, 0.86]	374
		per patient	0.92 [0.85, 0.98]	0.86 [0.78, 0.94]	51

dicted class could have been lower than 2. In the case of the LCX vessel, there are several clearly visible occlusions in the final part of the artery, and the SHAP values are higher in absolute value in this area, with pink pixels pushing the prediction towards the worst CAD class, while blue pixels attenuate this effect due to partially normal areas of the vessel. For the RCA, we can see two separate areas where the pixels are mostly blue and mostly pink, respectively. In the region occupied by the pink pixels, we can observe a slightly brighter area in the original image, which undoubtedly influenced the model output positively in predicting the higher CAD class. However, the final class is assigned based on all three coronary vessels, that represent 3 channels of a single input image and, in this case, it led to the correct prediction of the patient's outcome phenotype (class 2, i.e. CAD-RADS = 4-5).

This approach could be helpful in visually inspecting which part of the image the model gives more importance to, and to check for potential artifacts that may have erroneously influenced the final classification. This tool could be beneficial in better examining suspicious cases to quickly detect possible biases that could indicate that the algorithm may not be trustworthy in those cases. These user-friendly maps offer to non-expert clinical users a transparent way to assess the reliability of the prediction, avoiding a completely black-box approach.

Finally, the ablation studies showed that model's performance is minimally affected by reductions in both per-patient projections and MaxViT block configurations in the multi-class setting. The results of the first experiment, reported in Table 4, indicate that the model's performance in terms of per-patient efficiency remains relatively consistent despite a reduction in the number of projections. Notably, the minimal loss in performance suggests that the model can maintain a competitive level of accuracy and AUC, even when operating under reduced input information. This finding underscores the model's adaptability and resilience in learning essential information even when exposed to a limited number of view angles. The adaptability to fewer input views not only maintains robust performance but also potentially accelerates the process of extracting different projections, contributing to enhanced operational efficiency. The second experiment instead, focused on evaluating the impact of varying MaxViT block configurations on model performance. The results, detailed in Table 5, illustrate different combinations of MaxViT blocks and their respective performance metrics. Notably, among the tested configurations, the model exhibited relatively minimal performance loss in the configuration 1-2-4-1. Remarkably, this configuration has a significantly lower number of parameters while retaining robust performance. This adaptability could be crucial in resource-constrained environments where computational power is limited.

The strength of this work relies on the pipeline ability to achieve highly accurate CAD-RADS predictions without the need for any additional effort in the clinical practice, such as annotating single vessels, segments, or lesions. This could be advantageous in a real clinical setting where any additional step, not included in the standard clinical routine, would be limited to specific research studies. Furthermore, in our pipeline, expert clinical user manual intervention is minimized, thereby avoiding time-consuming and operator-dependent steps. Another crucial aspect is the way we composed the input data for the network. By always using the three main coronary arteries together as a single input, we trained a network to associate the CAD-RADS score to a sequence of arteries rather than each single vessel, exactly as a radiologist would do when assigning a patient-wise CAD-RADS score.

This work has some limitations that need to be addressed. First, the overall dataset, although significant for a clinical task, is relatively small. To mitigate this issue, we used a SOTA architecture specifically designed to work under all data regimes, and employed several technical strategies, including pre-training, cross-validation, data augmentation, weight and LR decay, and label-smoothing, to avoid overfitting. However, the performance of the proposed algorithms could be possibly further improved by training on a larger dataset. Additionally, using data from a single clinical center limits the generalization ability of our results. It is essential to test the models on independent populations to assess their generalization capabilities before application in clinical practice. Although during screening procedures the goal is more to quickly identify macro-categories of patients needing further assessment or to rule-out healthy subjects, a larger dataset could allow testing a multi-class classification approach that takes into account all CAD-RADS scores separately. Moreover, to restrict our analysis, in this study we did not include modifiers to describe patients with stents (modifier S), vulnerable plaque features (modifier V), or grafts (modifier G). From a clinical point of view, it would be interesting to evaluate the generalizability of the models to these sub-categories as well. From a technical perspective, potential future extensions of this work could involve replacing the suggested imputing strategy with a deep generative algorithm, specifically trained to generate synthetic samples representing healthy vessels. Training generative models can be challenging due to the extensive need for a vast amount of high-quality data, complex algorithm architectures, and significant computational resources necessary for their training. Although this approach would add a layer of complexity, requiring substantial computational resources, it would be interesting to assess whether there would be a substantial increase in the models' performance.

6. Conclusions

We proposed a fully automated pipeline able to classify images representing 2D longitudinal cross-sections extracted from CCTA scans of the three main coronary arteries, based on a fine-tuned Multi-Axis Vision Transformer model. The highly accurate results obtained in the two explored tasks (identify patients in need for further investigations and classify the patients according to the severity of the occlusion) suggest the great potential of the proposed approach. This is the first work that does not require any additional annotation step, which is not part of the clinical routine, and uses instead, a learning procedure that perfectly emulates the clinical screening process. Such a tool would be a useful decision support system in the clinical practice, able to help the radiologists in quickly identify severe patients as well as completely healthy subjects and better inspect borderline cases with the help of intuitive and visually explainable methods.

Declaration of competing interest

Riccardo Bellazzi is co-founder and shareholder of two spin-offs of the University of Pavia (Engenome s.r.l. and Biomeris s.r.l.) that operate in the field of data management, artificial intelligence and bioinformatics.

All the other authors have nothing to declare.

Acknowledgements

This work was supported by Fondazione Regionale per la Ricerca Biomedica (FRRB), Research Grant no. CP2 14/2018, Project “INTESTRAT-CAD” (PI: Gualtiero I. Colombo). The study protocol conformed to the principle of the Declaration of Helsinki and was approved by the “Ethics Committee of the IRCCS Istituto Europeo di Oncologia and Centro Cardiologico Monzino” (protocol code R329/15 - CCM 341, date of approval 23/09/2015). All recruited patients signed the written informed consent and participants did agree to share their de-identified information.

References

[1] E. Falk, P.K. Shah, V. Fuster, Coronary plaque disruption, Circulation 92 (3) (1995) 657–671.

[2] A. Kanitsar, D. Fleischmann, R. Wegenkittl, P. Felkel, E. Groller, CPR-Curved Planar Reformation, IEEE, 2002.

[3] R.C. Cury, S. Abbara, S. Achenbach, A. Agatston, D.S. Berman, M.J. Budoff, K.E. Dill, J.E. Jacobs, C.D. Maroules, G.D. Rubin, et al., CAD-RADSTM coronary artery disease–reporting and data system. An expert consensus document of the society of cardiovascular computed tomography (SCCT), the American college of radiology (ACR) and the North American society for cardiovascular imaging (NASCI). Endorsed by the American college of cardiology, J. Cardiovasc. Comput. Tomogr. 10 (4) (2016) 269–281.

[4] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, R. Socher, Deep learning-enabled medical computer vision, npj Digit. Med. 4 (1) (2021) 5.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint, arXiv:2010.11929, 2020.

[6] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: efficient convolutional neural networks for mobile vision applications, arXiv preprint, arXiv:1704.04861, 2017.

[7] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[9] Z. Dai, H. Liu, Q.V. Le, M. Tan, CoAtNet: marrying convolution and attention for all data sizes, Adv. Neural Inf. Process. Syst. 34 (2021) 3965–3977.

[10] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, R. Girshick, Early convolutions help transformers see better, Adv. Neural Inf. Process. Syst. 34 (2021) 30392–30400.

[11] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, Y. Li, MaxViT: multi-axis vision transformer, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV, Springer, 2022, pp. 459–479.

[12] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[13] Z. Huang, J. Xiao, X. Wang, Z. Li, N. Guo, Y. Hu, X. Li, X. Wang, Clinical evaluation of the automatic coronary artery disease reporting and data system (CAD-RADS) in coronary computed tomography angiography using convolutional neural networks, Acad. Radiol. (2022).

[14] Y. Li, Y. Wu, J. He, W. Jiang, J. Wang, Y. Peng, Y. Jia, T. Xiong, K. Jia, Z. Yi, et al., Automatic coronary artery segmentation and diagnosis of stenosis by deep learning based on computed tomographic coronary angiography, Eur. Radiol. 32 (9) (2022) 6037–6045.

[15] F. Denzinger, M. Wels, K. Breininger, M.A. Gülsün, M. Schöbinger, F. André, S. Buß, J. Görich, M. Sühling, A. Maier, Automatic CAD-RADS scoring using deep learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 45–54.

[16] F. Denzinger, M. Wels, O. Taubmann, M.A. Gülsün, M. Schöbinger, F. André, S.J. Buss, J. Görich, M. Sühling, A. Maier, CAD-RADS scoring using deep learning and task-specific centerline labeling, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2022, pp. 315–324.

[17] J.-F. Paul, A. Rohnean, H. Giroussens, T. Pressat-Laffouilhere, T. Wong, Evaluation of a deep learning model on coronary CT angiography for automatic stenosis detection, Diagn. Interv. Imaging 103 (6) (2022) 316–323.

[18] M. Penso, S. Moccia, E.G. Caiani, G. Caredda, M.L. Lampus, M.L. Carerj, M. Babbaro, M. Pepi, M. Chiesa, G. Pontone, A token-mixer architecture for CAD-RADS classification of coronary stenosis on multiplanar reconstruction CT images, Comput. Biol. Med. 153 (2023) 106484.

[19] A. Tejero-de Pablos, K. Huang, H. Yamane, Y. Kurose, Y. Mukuta, J. Iho, Y. Tokunaga, M. Horie, K. Nishizawa, Y. Hayashi, 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, Springer, 2019, pp. 732–740.

[20] S. Candemir, R.D. White, M. Demirer, V. Gupta, M.T. Bigelow, L.M. Prevedello, B.S. Erdal, Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary CT angiography with a deep 3-dimensional convolutional neural network, Comput. Med. Imaging Graph. 83 (2020) 101721.

[21] G. Muscogiuri, M. Chiesa, M. Trotta, M. Gatti, V. Palmisano, S. Dell'Aversana, F. Baessato, A. Cavaliere, G. Cicala, A. Loffreno, et al., Performance of a deep learning algorithm for the evaluation of CAD-RADS classification with CCTA, Atherosclerosis 294 (2020) 25–32.

[22] M. Zreik, R.W. Van Hamersvelt, J.M. Wolterink, T. Leiner, M.A. Viergever, I. Išgum, A recurrent CNN for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography, IEEE Trans. Med. Imaging 38 (7) (2018) 1588–1598.

[23] Z. Ren, X. Kong, Y. Zhang, S. Wang, UKSSL: underlying knowledge based semi-supervised learning for medical image classification, IEEE Open J. Eng. Med. Biol. (2023).

[24] Z. Ren, S. Wang, Y. Zhang, Weakly supervised machine learning, CAAI Trans. Intell. Technol. (2023).

[25] G. Pontone, G. Muscogiuri, D. Andreini, A.I. Guaricci, M. Guglielmo, A. Baggiano, F. Fazzari, S. Mushtaq, E. Conte, A. Annoni, et al., Impact of a new adaptive statistical iterative reconstruction (ASIR)-V algorithm on image quality in coronary computed tomography angiography, Acad. Radiol. 25 (10) (2018) 1305–1313.

[26] R.A. Takx, D. Suchá, J. Park, T. Leiner, U. Hoffmann, Sublingual nitroglycerin administration in coronary computed tomography angiography: a systematic review, Eur. Radiol. 25 (2015) 3536–3542.

[27] G. Pontone, D. Andreini, A. Bartorelli, E. Bertella, S. Mushtaq, C. Foti, A. Formenti, L. Chiappa, A. Annoni, S. Cortinovis, et al., Feasibility and diagnostic accuracy of a low radiation exposure protocol for prospective ECG-triggering coronary MDCT angiography, Clin. Radiol. 67 (3) (2012) 207–215.

[28] G. Pontone, G. Muscogiuri, A. Baggiano, D. Andreini, A.I. Guaricci, M. Guglielmo, F. Fazzari, S. Mushtaq, E. Conte, A. Annoni, et al., Image quality, overall evaluability, and effective radiation dose of coronary computed tomography angiography with prospective electrocardiographic triggering plus intracycle motion correction algorithm in patients with a heart rate over 65 beats per minute, J. Thorac. Imaging 33 (4) (2018) 225–231.

[29] G. Pontone, D. Andreini, E. Bertella, A. Baggiano, S. Mushtaq, M. Loguercio, C. Segurini, E. Conte, V. Beltrama, A. Annoni, et al., Impact of an intra-cycle motion correction algorithm on overall evaluability and diagnostic accuracy of computed tomography coronary angiography, Eur. Radiol. 26 (2016) 147–156.

[30] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J.B. Zimmerman, K. Zuiderveld, Adaptive histogram equalization and its variations, Comput. Vis. Graph. Image Process. 39 (3) (1987) 355–368.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[32] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, 2019.

[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[34] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint, arXiv:1409.1556, 2014.

[35] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., PyTorch: an imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).

[37] Itseez, Open source computer vision library, https://github.com/itseez/opencv, 2015.

[38] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, pp. 4765–4774.

[39] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI—explainable artificial intelligence, Sci. Robot. 4 (37) (2019) eaay7120.

[40] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, Explainable AI: a brief survey on history, research areas, approaches and challenges, in: Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8, Springer, 2019, pp. 563–574.

[41] C. Jones, J. Thornton, J.C. Wyatt, Enhancing trust in clinical decision support systems: a framework for developers, BMJ Health Care Inform. 28 (1) (2021).

[42] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

[43] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer, 2014, pp. 818–833.