Development of deep learning segmentation models for coronary X-ray angiography: quality assessment by a new global segmentation score and comparison with human performance

Miguel Nobre Menezes[1,2,*], João Lourenço Silva[3], Beatriz Silva[1,2],Tiago Rodrigues[1,2], Ana Rita Francisco[1,2], Pedro Carrilho Ferreira[1,2], Arlindo L. Oliveira[3], Fausto J. Pinto[1,2]

[1]Structural and Coronary Heart Disease Unit, Cardiovascular Center of the University of Lisbon, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
[2]Serviço de Cardiologia, Departamento de Coração e Vasos, CHULN Hospital de Santa Maria, Lisboa.
[3]INESC-ID / Instituto Superior Técnico, University of Lisbon

*Corresponding author:
E-mail address: mnmenezes.gm@gmail.com (M-Menezes)

Desenvolvimento de modelos de *Deep Learning* para segmentação de coronariografias: aferição de qualidade por um novo modelo de segmentação global e comparação com desempenho humano

Resumo

Introdução e objetivos: a segmentação automática de coronariografia (CRG) por inteligência artificial (IA) encontra-se pouco explorada na literatura médica. Os objetivos do presente estudo são (1) desenvolver modelos de IA para segmentação de CRG e (2) aferir os resultados por *scores* de similaridade e critérios definidos por peritos.

Métodos: doentes submetidos a CRG foram retrospectivamente selecionados aleatoriamente num centro. Por incidência, segmentou-se um *frame* ideal, formando uma segmentação humana basal (HB), usada para treinar um modelo de IA basal (IAB). Da combinação de ambos formou-se uma segmentação humana aperfeiçoada (HA), utilizada para treinar um modelo de IA aperfeiçoado (IAA). Os resultados foram aferidos com 11 critérios balanceados definidos por peritos, combinados num *Score de Segmentação Global* (SSC – 0 – 100 pontos). O *Score de Dice Generalizado (*SDG) e *Score de Dice de Similaridade* (SDS) aplicaram-se aos modelos de IA.

Resultados: geraram-se 1664 imagens processadas. Os SCC para a HB, HA, IAB e IAA foram 96,9 +/-5,7; 98,9 +/- 3,1; 86,1 +/- 10,1 e 90 +/- 7,6, respetivamente (IC 95%, p < 0,001 - diferenças globais e emparelhadas). O SDG para o IAB e IAA foi 0,9234 ± 0,0361 e 0,9348 ±

0,0284, respetivamente. O SDS foi 0,8904 ± 0,0464 e 0,9134 ± 0,0410 para o IAB e IAA, respetivamente. O IAA exibiu superior desempenho ao IAB para as todas tarefas de segmentação coronária, mas não para todas as de cateter.

Conclusões: desenvolvemos modelos de IA de segmentação automática de CRG, com bom desempenho de acordo com aferição por todos os *scores*.

Palavras-Chave

Aprendizagem profunda; Inteligência artifical; Aprendizagem de máquinas; Coronariografia; Doença coronária; Intervenção coronária percutânea.

Abstract

Introduction and Objectives: Although automatic artificial intelligence (AI) coronary angiography (CAG) segmentation is arguably the first step toward future clinical application, it is underexplored. We aimed to (1) develop AI models for CAG segmentation and (2) assess the results using similarity scores and a set of criteria defined by expert physicians.

Methods: Patients undergoing CAG were randomly selected in a retrospective study at a single center. Per incidence, an ideal frame was segmented, forming a baseline human dataset (BH), used for training a baseline AI model (BAI). Enhanced human segmentation (EH) was created by combining the best of both. An enhanced AI model (EAI) was trained using the EH. Results were assessed by experts using 11 weighted criteria, combined into a Global Segmentation Score (GSS: 0–100 points). Generalized Dice Score (GDS) and Dice Similarity Coefficient (DSC) were also used for AI models assessment.

Results: 1664 processed images were generated. GSS for BH, EH, BAI and EAI were 96,9 +/-5.7; 98.9 +/- 3.1; 86.1 +/- 10.1 and 90 +/- 7.6, respectively (95% confidence interval, p<0.001 for both paired and global differences). The GDS for the BAI and EAI was 0,9234 ± 0,0361 and 0,9348 ± 0,0284, respectively. The DSC for the coronary tree was 0,8904 ± 0,0464 and 0,9134 ± 0,0410 for the BAI and EAI, respectively. The EAI outperformed the BAI in all coronary segmentation tasks, but performed less well in some catheter segmentation tasks.

Conclusions: we successfully developed AI models capable of CAG segmentation, with good performance as assessed by all scores.

Keywords

Deep learning; Artificial Intelligence; Machine Learning; Coronary Angiography; Coronary Artery Disease; Percutaneous Coronary Intervention.

Introduction

Artificial intelligence (AI) has shown great potential in medicine, in applications such as predictive data analysis[1], decision making support[2] or even medical education/awareness improvement,[3] and especially in image analysis.

Several publications have demonstrated impressive results with regards to electrocardiogram[4], echocardiography,[5,6] or magnetic resonance imaging.[7,8]

The use of AI in Interventional Cardiology (IC) is, however, still a vastly underexplored field. Its application to coronary angiography (CAG) has been explored in very few medical or biology publication.[9–12] There are, nonetheless, many possibilities,[13] ranging from automatic anatomical identification, stenosis analysis, lesion subset characterization and perhaps even physiological index derivation. Regardless of the task, arguably the first step in applying AI to CAG is separating and identifying relevant information – the coronary tree – from non-relevant information (bones, other structures). This task is called segmentation.[14]

In this paper, we explore the development of AI models capable of automatic coronary artery segmentation from CAG, and assess the results from a clinical perspective, using a new set of criteria and score clinically defined by a panel of Interventional Cardiologists.

Methods

Dataset selection

Inclusion criteria

We retrospectively and randomly included patients who had undergone CAG and invasive physiology assessment (fractional flow reserve and/or other indexes) during the procedure at a single center (tertiary university hospital). These patients have at least intermediate lesions in one or more vessels. Around one third usually undergo revascularization due to the severity of their disease.[15,16] Therefore, a dataset focusing on these patients comprises a wide spectrum of obstructive coronary artery disease in a relatively balanced way.

Exclusion criteria

We excluded cases where any of the following applied:

1) Major occluded vessels (acute or chronic)
2) Poor image quality
3) Less than two orthogonal views in the left coronary artery (LCA) - one caudal and one cranial - or absence of at least one left oblique (LAO) view - either cranial or simple - in the right coronary artery (RCA)
4) Patients with previous cardiac surgery, cardiac devices or other sources of potential artifact.

Image selection

A single best frame was selected for each diagnostic angulation incidence in each patient,

Dataset size

The dataset size was the result of a trade-off between two opposing criteria: dimension large enough for successful training of a deep convolutional neural network, estimated from published data[9,12,17,18] vs. expected time required to complete the annotation. We estimated the latter based on a short period of annotation testing prior to formal dataset creation. The trade-off pointed to a training set size of roughly 400.

We then randomly and consecutively selected patients until a total of at least 400 annotated images were obtained.

Baseline annotation process

Baseline human dataset images were annotated by two senior Cardiology Fellows (TR/BS) previously trained in CAG interpretation, under the supervision of an Interventional Cardiologist (MNM), who also annotated. Images were periodically reviewed and perfected by all three. This meant that any initial heterogeneity between annotators was corrected by consensus. The small size of the team was aimed at reducing heterogeneity, as we noticed during the preparatory phase that some operators tended to annotate too much (Supplementary figure 1), while others did the opposite Supplementary figure 2).

Both the catheter (labeled red) and the coronaries (labeled white) were to be segmented.

The coronary tree was to be fully segmented up to branches of approximately 2 mm in caliper at their origin (as the vessel became smaller, it was to be segmented until discernible), using the catheter as reference (without formal measurements – eyeball appreciation was used). There were several reasons for this: (1) when performing percutaneous coronary intervention, vessels <2 mm are usually approached conservatively, as the risk of target lesion failure increases significantly[19,20]; (2) human annotation is cumbersome – segmenting every single vessel would increase the risk of errors significantly; (3) including very small vessels might increase the chances of artifacts from bone or other structures when training and applying AI models.

Baseline artificial intelligence model training

We performed segmentation using an encoder-decoder fully convolutional neural network based on the U-Net,[21] commonly used in medical image segmentation. As the name suggests, these neural networks are composed of an encoder, responsible for extracting image features, and a decoder, which processes those features to produce segmentation masks. To derive the best approach for this task, we conducted a comparative study of encoder and decoder architectures, which resulted in the proposal of the EfficientUNet++, a computationally efficient and high-performing decoder architecture[22] that, in this work, we combine with an EfficientNet-B5 encoder[23] (Figure 1).

To ensure fair evaluation, it was necessary to guarantee that each model was tested on data that it had not seen during training. Therefore, the dataset was split, at the patient level, into 13 subsets of approximately 32 angiograms each. Each subset segmentation was performed using a neural network trained exclusively on the remaining data. This enabled the assessment of the segmentation results for the entire cohort, as the usual splitting into a training and testing dataset would have yielded a much smaller group of images for result assessment.

The training hyperparameters, including the number of training epochs and the learning rate decay schedule, were set on the first train-test split, using one of the 12 training data subsets for validation. The selected values were then used on every other train-test split, and to train the model on the whole training set of the first split.

Enhanced human model

The results of the baseline AI training were reviewed by the annotating team, without any formal grading, which would be performed subsequently (see below). For each image, both human and AI segmentation were compared with the original. Each annotation was then perfected using a mixture of the best of baseline human segmentation and baseline AI, with additional de novo manual segmentation as needed.

Enhanced artificial intelligence model

The neural network architecture and training procedure were identical for both the baseline and enhanced AI model (figure 1). The sole difference was the dataset. The baseline AI model was trained using the baseline human annotations, whereas the enhanced AI model was trained using enhanced human annotations.

Figure 2 outlines the development stages.

Performance assessment

Non-medical metrics

AI models were assessed using the Dice Similarity Coefficient (DSC) and Generalized Dice Score (GDS), measures of the overlap between segmentations. Given two segmentations, the DSC has a value between 0: no overlap and 1: total overlap, corresponding to the ratio between the area of their intersection and the sum of their areas. GDS[24] is a weighted sum of each class's DSC that attributes the same importance to all classes, regardless of their frequency. While DSC and GDS alone do not reflect clinical usefulness, they are helpful and entirely objective metrics that enable a simple comparison between models.

Clinical performance criteria

The DSC objectively assesses model performance. However, it does not provide a medically meaningful impression of whether segmentation is appropriate. Also, because the DCS can only be calculated based on previously annotated images, it cannot be applied to new, unannotated datasets in the future. To overcome these limitations, we created a set of criteria to assess performance as interpreted by expert physicians.

The following 11 criteria are as objectively defined as possible and were analyzed for each image. Each was independently met or not. A “perfect” example is shown in Figure 3. Supplementary Figures 3 to 13 show error examples for each.

1) **Catheter segmentation**:
    a. **Main segmentation**: The distal part of the catheter (i.e. the closest discernible portion to the coronary artery in the ascending aorta) is correctly segmented and labeled (supplementary figure 3). If minor gaps are present, this criterion should be scored as met.
    b. **Gaps** (minor) are absent (supplementary figure 4).
    c. **Catheter thickness** is accurate, by visual appreciation (supplementary figure 5).
    d. **Location**: if parts of the catheter far from the coronary ostia (ascending and/or descending aorta) are segmented, there are no major gaps or artifacts (supplementary 6).
2) **Vessel segmentation**:
    a. **Main vessels** are correctly segmented and labeled. For the RCA, this includes the segments from the ostium to the crux (supplementary figure 7). For the LCA, this includes the segments

from the left main ostium to the visually discernible distal segments of the left anterior descending or the circumflex (or most important obtuse marginal branch), depending on incidence. Branches are excluded from this criterion. If minor gaps are present, this criterion should be scored as met.

   b. **Branch segmentation**: branches with a luminal diameter of at least approximately 2 mm (using the catheter size as reference) are correctly segmented and labeled (supplementary figure 8). Size is estimated by visual appreciation. If minor gaps are present, this criterion should be scored as met.
   c. **Main vessel gaps** (minor) are absent (supplementary figure 9).
   d. **Branch gaps** (minor) are absent (supplementary figure 10).
   e. Catheter to artery **transition**: correct labeling of the catheter tip vs. coronary artery origin (supplementary figure 11).

3) **Artifacts**
   a. **Coronary**: No non-coronary structures are labeled as part of the coronary tree (supplementary figure 12).
   b. **Catheter**: No non-catheter structures are incorrectly labeled as part of the catheter (supplementary figure 13).
   The criteria for these two artifacts are not applicable to the small catheter-artery transition area.

To provide an objective assessment, these criteria were scored by a panel of three Interventional Cardiologists (MNM, ARF, PCF), of whom two (ARF, PCF) took no part in any stage of the annotation/training process. Discrepancies were solved by agreement. All images were graded across all groups: baseline human segmentation, enhanced human segmentation, baseline AI and enhanced AI. During the grading process, the image group was blinded.

Lastly, because the abovementioned criteria are not equally important, a Global Segmentation Score (GSS – 1.5 to 100 points) was devised, taking into account the relevance of each criterion as defined by the three experts (Table 1). The panel was also asked to select which of the two AI models was preferred for each image, regardless of the final score.

Statistical analysis

Descriptive variables are shown in absolute and relative (percentage) numbers. To assess the association between qualitative (categorical) variables the Chi-Square test was used. To assess differences in quantitative variables we used the Mann-Whitney test (two independent groups) or the Kruskal-Wallis test (multiple independent groups). A $p<0.05$ was used for statistical significance, except for multiple groups comparisons, where we used a $p< 0.01$. IBM SPSS Statistics 27 was used for statistical analysis.

Ethical issues

This study complies with the Declaration of Helsinki and was approved by the local ethics committee.

Results

Baseline dataset

We included 416 images from 69 patients (Table 2). With two human and two AI datasets, 1664 processed images were generated.

Performance assessment

Non-medical metrics

Results are outlined in Table 3. These scores indicate that enhanced AI was generally superior to baseline AI. Segmentation performance was good and consistent across arteries, as indicated by the high mean and low standard deviation of the DSC. For the catheter, performance was lower and much less consistent.

Clinical performance

Overall performance – individual criteria assessment (Supplementary table 1)

Coronary segmentation

The main vessels were correctly segmented in almost all cases across groups. Minor gaps occurred rarely in the baseline human segmentation and both AI models, although there was a small but non-significant improvement with the enhanced AI vs. baseline AI.

Branch segmentation was also correct almost always in all groups, albeit less so than main vessel segmentation. There was a small, yet significant, improvement with the enhanced AI vs. baseline AI.

Minor branch gaps were quite common, revealing very significant differences between AI and human models. While enhanced AI performed numerically better than baseline AI, it still produced small gaps in nearly two thirds of cases.

Coronary artifacts were very uncommon in human annotations and were usually minor imperfections in catheter/coronary crossovers. They were common and usually minor in both AI models, although there was a very significant improvement with the enhanced AI vs. baseline AI (14.4% vs. 25.7%).

Catheter/artery transition

Baseline human segmentation failed in 12% of cases and enhanced human segmentation missed 3.8%. Baseline AI produced a higher error rate (19.7%), but enhanced AI was numerically more often correct than baseline human segmentation, sometimes correctly identifying the transition where humans failed (Figure 4).

Catheter segmentation

Baseline human segmentation produced thickness imperfections (usually mildly engorged catheter) in 13.9% of cases, but otherwise, segmentation was almost always correct regarding other criteria. Baseline AI produced low error rates in main body segmentation. However, artifacts, usually quite minor and in the vicinity of coronary segments, occurred very frequently (41.1%). Another common error was catheter thickness (36.3%), often resulting in an overestimation of catheter size.

Enhanced human segmentation significantly improved on thickness issues, although imperfections persisted in 6.2% of cases.

Enhanced AI produced better results than the baseline AI model for catheter thickness (correct in 96.4%), also surpassing both human models (although the difference was not statistically significant when compared to the enhanced human segmentation). However, the performance of the enhanced AI otherwise decreased in all other criteria, especially regarding minor gaps, which became much more common (3.1% in the baseline AI model to 23.3%). Even main body segmentation was significantly affected, although successful in the vast majority of cases (86.5%). Despite this, in most failures catheter identification was still possible, as major gaps often occurred distally in areas of contrast backflow. There was a slight numerical worsening in artifact and location issues in enhanced AI vs. baseline AI.

Overall performance – Global Segmentation Score assessment and expert preference (table 4)

Human models outperformed AI models. Enhanced models surpassed baseline models. The difference was statistically significant for all comparisons. GSS was very high for both AI models; the enhanced AI reached an average of 90 points.

With regards to expert preference, the enhanced AI model was preferred in 300 (72%) cases, the baseline AI model in 100 (24%) and in 16 (4%) cases no AI model was preferred.

Performance according to coronary artery – individual criteria assessment (Supplementary table 2)

There was a trend toward better performance in the RCA, both regarding human and AI groups. The most notable and statistically significant differences occurred in catheter transition (regarding both AI models and the baseline human segmentation) and catheter segmentation (both AI models performed better in the RCA). Branch gaps were quite less frequent in the RCA with the enhanced AI model. Other differences, even if statistically significant, were very small.

Performance by coronary artery – Global Segmentation Score assessment (Supplementary table 3)

All models scored very high for both arteries. There were very minor statistically significant differences for the baseline AI model only, favoring RCA segmentation.

Considering expert preference:

- RCA: Enhanced AI was preferred in 109 (68.6%) cases, the baseline AI was preferred in 43 (27%) and in 7 (4.4%) cases no AI model was preferred.
- LCA: Enhanced AI was preferred in 191 (74.3%) cases, the baseline AI was preferred in 57 (22.2%) and in 9 (3.5%) cases no AI was preferred.

Performance according to angulation incidence – individual criteria assessment (Supplementary tables 4 and 5)

Given the large amount of data, there being no significant differences in the vast majority of cases and for the sake of readability, only statistically significant differences are shown in the tables. Overall, the impact of incidences on model performance was limited, and affected almost exclusively the AI models.

Performance according to angulation incidence – Global Segmentation Score assessment (Supplementary tables 6 and 7)

Differences were minor and only statistically significant for human performance in less common incidences (PA views for the LCA and PA cranial for the RCA).

Discussion

Overall considerations

Baseline human segmentation was generally correct. Catheter/coronary transition and catheter thickness errors were the most common. Poor individualization due to contrast backflow, catheter curves and human fatigue all likely contributed.

Enhanced human segmentation was nearly perfect. Mild transition issues remained, highlighting the difficulty of the task. As this model was actually a combination of the best of baseline human segmentation and baseline AI, it also demonstrates how AI can help improve human performance. Even these slight human imperfections highlight the need for rigorous quality control during and after the final results, rather than assuming human annotation is a "perfect" ground truth. This an inherent limitation to the annotation of medical images, as the sheer amount of cumbersome work is error prone.

Baseline AI performed CAG segmentation successfully yet was affected by the same two issues of the baseline human segmentation – transition and catheter thickness. The effort to correct these when developing the enhanced AI was fruitful in the case of transition but produced mixed results for catheter thickness. Impact on transition performance was impressive, as, at times, the enhanced AI even achieved correct assessments where humans failed (Figure 4). However, it seems the gain in catheter thickness accuracy was offset by losses in other catheter segmentation tasks. Lastly, every aspect of coronary segmentation improved in the enhanced AI, which performed better than baseline AI. The differences between the two AI models also highlight how relatively small differences in the ground truth can impact relevantly on AI training.

It may seem surprising that catheter segmentation was less successful than coronary segmentation. However, while intuitively one may think that catheter segmentation is an easier task and therefore the results would have been better for this task, from a machine learning perspective that is not the case. In particular, segmentation performance is highly dependent on the frequency of each class. Rarer classes, or ones that occupy smaller areas, are interpreted by the model as being less likely to appear. Furthermore, during training, the lower the number of pixels belonging to a particular class, the lower the penalty for segmenting that class incorrectly. Even though we used a loss function designed to mitigate this phenomenon, the poorer segmentation of less common classes (the catheter, in this case) is still evident in the results.

Right coronary artery segmentation was easier than LCA, however the differences were quite small and there were fewer than expected, considering its greater anatomical simplicity. Angulations also had a relatively small impact both on human and AI performance and small observed differences may be attributed to specific issues that are more common in certain incidences: contrast backflow (less problematic in PA or RAO caudal); coronary/catheter crossovers (such as spider or extreme RAO cranial – Figure 5); proximity of bone (such as RCA LAO views); smaller samples of some incidences, such as PA cranial; uncommon catheter pathways, such as the femoral approach, which sometimes produces a central vertical outline.

Globally, both AI models achieved a very high DSC, with higher performance in artery segmentation than in catheter segmentation, supporting the results of qualitative clinical assessment. When factors are weighed up based on their perceived relevance – as assessed by GSS – both performed very well. The enhanced AI scored an average of 90 points, meaning it provided 90% of what experts deemed most relevant when viewing a CAG. By all measures, the enhanced AI was the better model. However, the fact that differences between the two AI models were not large and that the enhanced AI was preferred in most, but not all cases, highlights the difficulty in improving an already good performance.

Other studies with artificial intelligence applied to coronary angiography segmentation/interpretation

Few studies regarding coronary artery segmentation based on AI technologies have been published in medical/biology journals to date. Yang et al.[12] successfully developed AI models capable of segmenting CAG. Their dataset was larger (3302 images/2042 patients) and was also annotated by two expert physicians. Different incidences were also used. They also focused exclusively on segmenting specific segments of major vessels with at least mild (>30%) stenotic lesions. Neither the branches nor the catheter were segmented, leading to a much simpler problem than the one addressed in this article.

Two other work,s[9,10] from the same baseline dataset, also developed AI-based CAG segmentation. Their dataset was also larger (4904 images from 170 videos). However, the annotations were performed by medical students and no details are provided regarding patient subset, target vessel or incidence.

Very recently, Du et al. [11] published the results of a broad study. They focused on two tasks: CAG segmentation and special lesion morphology identification (calcium, thrombus, among others). For the former task, which overlaps with ours, they used a very large dataset of 13373 images distributed across ten incidences (six LCA and four RCA), annotated by ten qualified analysts. This was an all-comers study, rather than focusing on patient subsets. They too annotated catheter/arteries and additionally marked different coronary segments. Their model is impressive as judged by the presented images, as they even distinguished between contrast backflow, catheter and coronary. However, they did not specify the exact criteria for segmenting the coronary tree and their exact metrics make it difficult to assess exactly how their models performed in detail regarding segmentation.

While all the abovementioned groups have worked with datasets larger than ours, our study has several unique features: (1) there was medical rationale for vessel size segmentation; (2) results were assessed from a set of criteria defined by experts, capturing the quality of the segmentation from an Interventional Cardiologist's eyes; (3) human annotations were also graded, rather than assuming a perfect human ground truth; (4) specific segmentation tasks were appraised individually, enabling insights into strengths and weaknesses of AI and human models alike; (5) results were also considered globally with the GSS, by factoring the relevance of each criterion, enabling a broad, simple appreciation of the results. Furthermore, the ability to perform high-quality segmentation in a system trained using less data provides relevant evidence that more advanced AI systems can be effectively applied even in situations where the available data are limited.

Limitations

This is a single center retrospective dataset, involving a single image per projection and a smaller sample size than some previously published manuscripts. The images come from the same angiography devices (Siemens Artis) and thus we have not yet tested our models on images obtained from other equipment or image settings.

We have not yet conducted formal assessment on how well the models perform in segmenting specific degrees of stenosis severity. Our models are also yet to be tested for specific vessel disease types (calcium, thrombus), clinical settings (chronic total occlusion, ST-elevation myocardial infarction).

We have not yet assessed the performance of AI models on an external validation cohort. There are several reasons for this. We aimed to compare AI and human results in detail first and assess the exact performance of AI models for each segmentation task. A validation dataset would comprise a new set of images, which would not undergo human segmentation, thus impeding comparison with human performance. Also, validation implies that a metric be available for comparing results. Because the Dice methods require a ground truth human annotation for comparison, and the GSS was developed and applied for the first time for this paper, we felt a suitable metric was not yet available for performing validation prior to the current analysis. In addition, AI models are continuously and dynamically improving. As we are currently working on further testing and enhancing current AI models (view Future direction and implications section below), we felt performing external validation at this stage was premature.

The exclusion of cardiac devices/cardiac surgery and other foreign objects renders our models not yet applicable to such cases. We did not, however, exclude cases with previously implanted stents.

Lastly, focusing specifically on patients undergoing invasive physiology assessment may have created bias, limiting a broader application of the models to other patient subsets.

We are currently working to address all these issues in future research.

Future direction and implications

Coronary angiography segmentation in itself is not a end objective but rather an essential milestone for developing AI systems capable of CAG analysis and interpretation. These results should, therefore, be regarded as a first step, rather than a final deployment tool. While not yet mature for immediate clinical application, the results of both AI models are already relevant, providing a framework that can be built upon in the future.

Further steps include testing the models for stenosed segments, which will be critical for clinical application. In the future, we aim to test our models with a validation cohort using new angiograms. Sub-segmentation, automatic anatomical identification and physiology are also areas for future research.

We will also strengthen the capabilities of our models further by broadening our training base to other patient and lesion subsets, focusing on particular issues where there is still room for improvement, as identified by our uniquely detailed analysis.

Our results also provide insight into which human tasks are most challenging, which may be of use to other researchers.

Global Segmentation Score is the first of its kind for assessing the quality of segmentations in CAG. By providing a reasonably objective and quantitative clinical measurement, it can be used as a benchmark for comparing and validating results across research groups.

Lastly, while conventional segmentation software does exist, it is not without limitations, and only by developing AI systems can we compare and improve both in the future. The potential implications of AI for Interventional Cardiology are immense, and we envisage a catherization lab of the future where all of these insights render the human eye more objective, thus improving patient care.

Conclusions

We successfully developed two AI models capable of good quality automatic CAG segmentation, as assessed by GDS, DSC and the GSS. From an expert's perspective, the latter and its individual criteria provided a feasible, reasonably objective and quantifiable way of assessing the results.

The enhanced AI model outperformed the baseline AI model in coronary segmentation tasks as well as globally. With regards to catheter segmentation tasks, the enhanced AI model improved on the task of catheter thickness, but performed less well in other catheter segmentation tasks. Both human segmentations were superior to both AI models, but only the enhanced human segmentation, built by combining the best of baseline human segmentation and baseline AI, achieved a near perfect GSS.

These results provide a relevant framework for building upon, potentially leading to future clinical application.

Acknowledgments

MNM was responsible for conceptualization, data gathering, processing and analysis, interpretating the results and paper drafting. JLS was responsible for technical and AI tasks, data and image processing, model implementation and training. TR and BS were responsible for data gathering, processing and analysis. ARF and PCF were responsible for data analysis and results interpretation. ALO was responsible for supervising the work of JLS, and participated directly in the same tasks. FJP was responsible for supervising the work of MNM, and participated directly in the same tasks.

All authors revised the paper critically for important intellectual content, gave final approval for its publication and agreed to be accountable for all respects of its accuracy and integrity.

References

1. Shah SJ, Katz DH, Selvaraj S, Burke MA et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation. Lippincott Williams and Wilkins; 2015;131(3):269–79.
2. Somashekhar SP, Sepúlveda M-J, Norden AD et al. Early experience with IBM Watson for Oncology (WFO) cognitive computing system for lung and colorectal cancer treatment. J Clin Oncol. American Society of Clinical Oncology (ASCO); 2017 May 20;35(15_suppl):8527–8527.
3. Fialho I, Beringuilho M, Madeira D et al. Acute myocardial infarction on YouTube – is it all fake news? Rev Port Cardiol (English Ed. Elsevier; 2021 Nov 1;40(11):815–25.
4. Hannun AY, Rajpurkar P, Haghpanahi M et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine. Nature Publishing Group; 2019. p. 65–9.
5. Narula S, Shameer K, Salem Omar AM et al. Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography. J Am Coll Cardiol. Elsevier USA; 2016 Nov 29;68(21):2287–95.
6. Asch FM, Poilvert N, Abraham T et al. Automated Echocardiographic Quantification of Left Ventricular Ejection Fraction Without Volume Measurements Using a Machine Learning Algorithm Mimicking a Human Expert. Circ Cardiovasc Imaging. NLM (Medline); 2019 Sep 1;12(9):e009303.
7. Ngo TA, Lu Z, Carneiro G. Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. Med Image Anal. Elsevier B.V.; 2017 Jan 1;35:159–71.
8. Bai W, Sinclair M, Tarroni G et al. Automated cardiovascular magnetic resonance image analysis with

fully convolutional networks 08 Information and Computing Sciences 0801 Artificial Intelligence and Image Processing. J Cardiovasc Magn Reson. BioMed Central Ltd.; 2018 Sep 14;20(1):65.

9. Wang L, Liang D, Yin X et al. Coronary artery segmentation in angiographic videos utilizing spatial-temporal information. BMC Med Imaging 2020 201. BioMed Central; 2020 Sep 24;20(1):1–10.

10. Liang D, Qiu J, Wang L et al. Coronary angiography video segmentation method for assisting cardiovascular disease interventional treatment. BMC Med Imaging 2020 201. BioMed Central; 2020 Jun 16;20(1):1–8.

11. Du T, Xie L, Zhang H et al. Training and validation of a deep learning architecture for the automatic analysis of coronary angiography. EuroIntervention. Europa Group; 2021 May 1;17(1):32–40.

12. Yang S, Kweon J, Roh J-H et al. Deep learning segmentation of major vessels in X-ray coronary angiography. Sci Reports 2019 91. Nature Publishing Group; 2019 Nov 15;9(1):1–11.

13. Ben Ali W, Pesaranghader A, Avram R et al. Implementing Machine Learning in Interventional Cardiology: The Benefits Are Worth the Trouble. Front Cardiovasc Med. Frontiers; 2021 Dec 8;0:1775.

14. Gonzalez & Woods, Digital Image Processing, 4th Edition | Pearson [Internet]. [cited 2021 Sep 2]. Available from: https://www.pearson.com/us/higher-education/program/Gonzalez-Digital-Image-Processing-4th-Edition/PGM241219.html

15. Davies JE, Sen S, Dehbi H-M et al. Use of the Instantaneous Wave-free Ratio or Fractional Flow Reserve in PCI. N Engl J Med. Massachusetts Medical Society; 2017 May 11;376(19):1824–34.

16. Götberg M, Christiansen EH, Gudmundsdottir IJ et al. Instantaneous Wave-free Ratio versus Fractional Flow Reserve to Guide PCI. N Engl J Med. Massachusetts Medical Society; 2017 May 11;376(19):1813–23.

17. Zhu X, Cheng Z, Wang S et al. Coronary angiography image segmentation based on PSPNet. Comput Methods Programs Biomed. Elsevier; 2021 Mar 1;200:105897.

18. Jun TJ, Kweon J, Kim YH, Kim D. T-Net: Nested encoder–decoder architecture for the main vessel segmentation in coronary angiography. Neural Networks. Pergamon; 2020 Aug 1;128:216–33.

19. Sim HW, Ananthakrishna R, Chan SP et al. Treatment of Very Small De Novo Coronary Artery Disease With 2.0 mm Drug-Coated Balloons Showed 1-Year Clinical Outcome Comparable With 2.0 mm Drug-Eluting Stents [Internet]. J Invasive Cardiol. 2018 Jul;30(7):256-261. 2018 [cited 2021 Aug 29]. Available from: https://www.hmpgloballearningnetwork.com/site/jic/articles/treatment-very-small-de-novo-coronary-artery-disease-20-mm-drug-coated-balloons-showed-1

20. LC van der H, MM K, PW D et al. Small-vessel treatment with contemporary newer-generation drug-eluting coronary stents in all-comers: Insights from 2-year DUTCH PEERS (TWENTE II) randomized trial. Am Heart J. Am Heart J; 2016 Jun 1;176:28–35.

21. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2015 May 18;9351:234–41.

22. Silva JL, Nobre Menezes M, Rodrigues T et al. Encoder-Decoder Architectures for Clinically Relevant Coronary Artery Segmentation. arXiv:210611447 [eessIV]. 2021 Jun 21;

23. Tan M, Le Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 36th Int Conf Mach Learn ICML 2019. International Machine Learning Society (IMLS); 2019 May 28;2019-June:10691–700.

24. Sudre CH, Li W, Vercauteren T et al. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. Deep Learn Med image Anal multimodal Learn Clin Decis Support Third Int Work DLMIA 2017, 7th Int Work ML-CDS 2017, held conjunction with MICCAI 2017 Quebec City, QC,. Europe PMC Funders; 2017;2017:240.

Figure Legends

Figure 1: Segmentation model composed of an EfficientNet-B5 encoder and an EfficientUNet++ decoder.

Figure 2: Annotation and training process.

Figure 3: A segmentation case fulfilling all 11 criteria.

Figure 4 (left to right): The first human segmentation incorrectly labels contrast backflow as coronary. The baseline AI model improves on the human segmentation but is still not perfect. The enhanced human model segments the transition perfectly. The enhanced AI model is hampered in catheter segmentation but identifies the transition correctly.

Figure 5: Crossovers in spider (above) and extreme RAO cranial (below) views generating artifacts.

Supplementary Figure 1: A test case with too many annotated vessels (not used for training).

Supplementary Figure 2: A test case with too few annotations (not used for training).

Supplementary Figure 3: Contrast backflow leads the artificial intelligence model to disregard most of the catheter.

Supplementary Figure 4: Minor catheter gaps, possibly facilitated by contrast backflow.

Supplementary Figure 5: Catheter thickness is overestimated.

Supplementary Figure 6: A small part of the catheter in the descending aorta was segmented in a femoral access case.

Supplementary Figure 7: A major part of the left main was not segmented.

Supplementary Figure 8: The posterolateral branch and much of the posterior descending artery were missed. The enhanced artificial intelligence model did not miss these vessels.

Supplementary Figure 9: A small gap is visible in the left anterior descending artery.

Supplementary Figure 10: Small gaps are visible in branches.

Supplementary Figure 11: Contrast backflow renders the transition less discernible, leading the model to miss the transition zone.

Supplementary Figure 12: A part of the intervertebral disk and vertebra are mislabeled as coronary.

Supplementary Figure 13: Contrast backflow is mislabeled as catheter.

Tables

Table 1: scoring metrics for application of the Global Segmentation Score.

Criteria	Catheter vs. Coronary Relative Weight	Individual Criteria Relative Weight	Points
Main vessel segmentation	70%	40%	28.0
Main vessel gaps		10%	7.0
Catheter to artery transition		15%	10.5
Branch segmentation		20%	14.0
BranchGaps		5%	3.5
Coronary artifacts		10%	7.0
Catheter segmentation	30%	40%	12.0
Catheter gaps		10%	3.0
Catheter artifacts		15%	4.5
Catheter location		5%	1.5
Catheter thickness		30%	9.0
Total			100

Table 2: Baseline clinical characteristics of patients from whom images were analyzed.

Factor	N +/- SD or N(%)
Age	67 +/- 11
Sex (male)	54 (78%)
Hypertension	56 (81.2%)
Diabetes mellitus	27 (39.1%)
Dyslipidemia	39 (56.5%)
Smoker (past or present)	26 (37.7%)
Chronic coronary syndromes	50 (72.5%)
Acute coronary syndrome	19 (27.5%)
Revascularization during/after CAG	21 (30.4%)

Table 3: Generalized Dice Score and class-wise Dice Similarity Coefficient obtained by the baseline and enhanced AI models. Results presented as mean ± standard deviation.

	BAI	**EAI**
GDS	0.9234±0.0361	0.9348±0.0284
Artery DSC	0.8904±0.0464	0.9134±0.0410
Catheter DSC	0.7526±0.1998	0.7975±0.1836

BAI: baseline AI model; EAI; enhanced AI model

Table 4: performance by group according to Global Segmentation Score (significance at $p<0.05$ for paired differences and $p<0.01$ for multiple comparisons)

GSS	Group				p-value						
	BH	EH	BAI	EAI	Between all*	BH vs EH**	BAI vs EAI**	BH vs BAI**	EH vs EAI**	BH vs EAI**	EH vs BAI**
Mean +/- SD	96,9 +/-5.7	98.9 +/- 3.1	86.1 +/- 10.1	90 +/- 7.6	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Median (IQR)	100 (9)	100(0)	87.5 (9)	92 (9.5)							

BH: baseline human model; EH: enhanced human model; BAI: baseline AI model; EAI; enhanced AI model; GSS: Global Segmentation Score; SD: standard deviation; IQR: interquartile range

* Kruskal-Wallis Test

** Mann-Whitney Test