Evaluating Different Explainability Methods for Coronary Artery Segmentation

Apostolos Stogiannis
Democritus University of Thrace
Xanthi, Greece
e-mail: aposstog@ee.duth.gr

Nikos Tsolakis
Information Technologies Institute
Centre for Research & Technology Hellas
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
e-mail: ntsolaka@csd.auth.gr

Miriam Gutierrez
Vicomtech, Basque Research and Technology Alliance
Spain
e-mail: mgutierrezf@vicomtech.org

Laura Valeria Pérez
Vicomtech, Basque Research and Technology Alliance
Spain
e-mail: lvperez@vicomtech.org

Karen López-Linares
Vicomtech, Basque Research and Technology Alliance
Spain
e-mail: klopez@vicomtech.org

Christoniki Maga-Nteve
Information Technologies Institute
Centre for Research & Technology Hellas
Thessaloniki, Greece
e-mail: chmaga@iti.gr

Georgios Meditskos
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
e-mail: gmeditsk@csd.auth.gr

Stefanos Vrochidis
Information Technologies Institute
Centre for Research & Technology Hellas
Thessaloniki, Greece
e-mail: stefanos@iti.gr

Abstract—**Deep learning achieves high accuracy in coronary artery segmentation but lacks interpretability for clinical deployment. We present a benchmark of five Explainable Artificial Intelligence (XAI) methods (Gradient-weighted Class Activation Mapping (Grad-CAM), Grad-CAM-plus-plus, Score-CAM, Integrated Gradients, Local Interpretable Model-agnostic Explanations (LIME)) on coronary angiography images from the ARCADE dataset. We introduce a vessel-aware evaluation framework with four metrics, Pointing Game, Average Precision, Intersection over Union, Energy Concentration and systematically optimize layer selection and scoring strategies for each method across three patients. Gradient-based CAM methods achieved an aggregate score of 0.420, with consistent layer preferences across patients, while perturbation methods fail. Our findings establish Grad-CAM as optimal for coronary vessel explanation and demonstrate that layer optimization is method-specific but patient-invariant.**

***Keywords*-Explainable Artificial Intelligence (XAI); Coronary Artery Segmentation; Medical Imaging; Healthcare.**

I. INTRODUCTION

Cardiovascular Diseases (CVDs) remain the leading cause of mortality worldwide [1]. A critical aspect in both diagnosis and treatment planning is the analysis of coronary arteries, where stenosis, occlusions, or anatomical abnormalities can have life-threatening consequences. X-ray Coronary Angiography (XCA) is the clinical gold standard for visualizing the coronary vasculature [2], yet manual interpretation is labor-intensive, subjective, and prone to inter-observer variability [3]. Automated vessel segmentation can accelerate clinical workflows and provide consistent anatomical delineation, directly supporting decision making in coronary interventions and surgical planning.

Deep learning has achieved state-of-the-art performance in medical image segmentation. Architectures such as U-Net [4] and its variants effectively capture fine vessel structures, enabling automated extraction of coronary artery trees from angiographic images [5][6][7]. However, despite these advances, most models are often treated as black boxes, offering limited insight into how predictions are made [8]. This opacity raises concerns in safety-critical clinical applications, where transparency and trustworthiness are as important as accuracy. In coronary imaging, a lack of explainability can hinder clinical adoption, as physicians must understand and validate the rationale behind automated vessel delineations before integrating them into diagnostic or interventional decisions. Misinterpretation of coronary anatomy, such as inaccurate boundary detection or missing lesions, could directly affect treatment planning, stent placement, or surgical strategy [9].

Explainable Artificial Intelligence (XAI) aims to bridge this gap by enhancing the interpretability of deep learning models. In coronary artery segmentation, XAI can highlight which image regions most influenced the model's output, providing a visual rationale for automated predictions. While several XAI methods have been proposed, such as gradient-based (Gradient-weighted Class Activation Mapping (Grad-CAM) [10], Grad-CAM++ [11]), perturbation-based (Score-CAM [12]), and surrogate-model approaches like Local Interpretable Model-agnostic Explanations (LIME) [14], their effectiveness for thin, branching, and sparse vascular structures has not been systematically evaluated.

In this work, we address this gap by implementing and benchmarking multiple XAI methods for coronary artery segmentation on the publicly available ARCADE dataset [6] using a U-Net model. Our goal is to provide a reproducible framework for evaluating explanation methods tailored to vascular imaging. We apply a set of vessel-aware quantitative metrics: Pointing Game [15], Average Precision [16], Intersection over Union [17], and Energy Concentration Ratio [18], that complement visual inspection and enable rigorous comparison. By analyzing results across multiple patients, we highlight which methods are most informative and practical for clinical use, ultimately aiming to improve the trustworthiness and adoption of AI in cardiology.

The remainder of this paper is organized as follows. Section II reviews related work on XAI in medical imaging and coronary artery analysis. Section III describes the proposed methodology, including the dataset, model architecture, explainability methods, and evaluation metrics. Section IV presents the explanation evaluation framework. Section V reports the experimental results and quantitative comparisons. Section VI discusses the obtained findings and their implications. Finally, Section VII concludes the paper and outlines directions for future work.

II. Related Work

In recent years, considerable research has focused on integrating XAI into medical imaging to enhance interpretability and clinical trust. These approaches aim not only to make deep learning models transparent but also to create conditions that support their safe deployment in clinical workflows.

Do et al. [19] proposed a deep learning framework for diagnosing Coronary Artery Disease (CAD) using Single-Photon Emission Computed Tomography (SPECT) Myocardial Perfusion Imaging (MPI) polar maps. Their ResNet152V2 model incorporated LIME, Grad-CAM, and RISE for interpretability, evaluated via deletion and insertion metrics to overcome limitations of qualitative heatmap inspection. Similarly, Papandrianos et al. [20] developed an RGB-Convolutional Neural Network (CNN) model for automatic CAD classification from SPECT MPI images, achieving 93.3 % accuracy and 94.6 % AUC, and used a Grad-CAM-based color visualization to explain model decisions. Beyond imaging, Goettling et al. [21] introduced xECGArch, an interpretable deep learning architecture for ECG analysis that combines short- and long-term CNN branches. They compared 13 XAI methods using perturbation analysis, demonstrating the value of systematic XAI evaluation for physiological signal interpretation.

Bhandari et al. [22] applied Grad-CAM, LIME, and SHAP to classify chest X-ray images into COVID-19, pneumonia, and tuberculosis, showing that combining complementary XAI methods can improve visual interpretability. Their work highlighted the importance of multi-method comparison for clinical validation of deep learning models.

Anand et al. [23] employed a U-Net architecture for coronary vessel segmentation in X-ray angiography, achieving strong quantitative results (mean F1 = 0.921) but without integrating explainability. Gao et al. [24] later proposed an ensemble framework combining deep learning and filter-based features for coronary artery segmentation, reporting high precision and sensitivity across 130 angiographic images. These works demonstrate the maturity of segmentation pipelines but also the lack of explainability integration in vessel-specific contexts.

Bhati et al. [9] provided a broad survey of XAI techniques in medical imaging, categorizing gradient-based, perturbation-based, decomposition, and attention-driven methods, and discussing challenges to clinical adoption and validation.

Collectively, these studies show that while XAI is increasingly applied to medical imaging, quantitative evaluation of explanations, particularly for segmentation tasks involving thin, branching structures like coronary arteries, remains underexplored. To address this gap, our study focuses on systematically benchmarking multiple canonical XAI methods (Grad-CAM, Grad-CAM++, Score-CAM, Integrated Gradients [13], LIME) on coronary artery segmentation, introducing vessel-aware evaluation metrics (pointing game, average precision, Intersection Over Union (IoU), energy concentration ratio) and ensuring reproducibility through open, well-specified implementation details.

III. Methodology

This section describes the overall methodology adopted in this study, including the dataset selection, the preprocessing steps, the segmentation model architecture, the evaluation metrics, and the explainability methods used to interpret and evaluate the model predictions.

A. Dataset Acquisition and Preprocessing

This study employs data from the ARCADE dataset which contains 1,200 X-ray Coronary Angiography (XCA) images with pixel-level vessel annotations. The dataset is divided into 1,000 training and 200 validation images. Each image is accompanied by a binary mask labeling vessel vs. background. While ARCADE also provides region-specific annotations for 26 SYNTAX anatomical regions, this study focuses on binary segmentation to isolate the vessel tree as a whole. The dataset was converted into TFRecord format to enable efficient training and inference.

Besides ARCADE there are other public datasets for coronary artery analysis, such as CADICA [25], XCAD [26]. CADICA provides annotated invasive coronary angiography videos; however, it is limited by a relatively small cohort size. XCAD includes segmentation masks for coronary arteries but offers fewer annotated samples and a restricted testing set. In contrast, the ARCADE dataset provides a large-scale, standardized benchmark with high-quality pixel-level vessel annotations and a diverse clinical population. Its size, annotation quality, and task diversity make ARCADE particularly suitable for robust coronary artery segmentation and explainable artificial intelligence evaluation. Therefore, ARCADE was selected as the primary dataset for this study.

To enhance vessel visibility and standardize inputs, several preprocessing techniques were applied. First, a white top-hat filter was used to enhance the contrast of bright vessels against the dark background. Through normalization, the intensity values were scaled from [0-255] to [0-1]. Lastly, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to further enhance contrast. Collectively, these preprocessing steps enhance the visibility of thin vessel structures that are otherwise difficult to distinguish from background noise in XCA.

B. Model Architecture and training configuration

The segmentation architecture is based on the U-Net model, selected for its proven versatility and strong performance in medical image segmentation. Several segmentation architectures were evaluated during preliminary experiments, with U-Net providing the best trade-off between performance and interpretability. The network comprises five encoder and five decoder stages. Each encoder block applies two 3×3 convolutions with ReLU activation and HeNormal kernel initialization, followed by batch normalization, dropout, and max pooling for down sampling. Each decoder block performs 2× upsampling via a 3×3 transposed convolution (stride 2), concatenates the corresponding encoder skip connection, and applies two 3×3 convolutions. The final output layer uses a 1×1 convolution followed by a softmax activation to generate per-class probability maps. During training, a Centerline Cross-Entropy (clCE) loss was employed to emphasize accurate segmentation of thin and elongated coronary vessels [27]. This loss increases penalties for misclassification near vessel centerlines, helping preserve connectivity and reduce fragmentation in fine vascular branches, which is critical for clinical interpretability.

C. Model Performance Evaluation

To evaluate model performance, four metrics were selected, Precision, Recall, Dice Coefficient and IoU. Precision is the number of true positive results divided by the number of all positive results. It quantifies to avoid false detections.

$$\text{Precision} = \text{TP}/ (\text{TP} + \text{FP}) \tag{1}$$

TP, FP, and FN denote True Positives, False Positives, and False Negatives, respectively.

Recall is the number of true positive results divided by the number of all cases that should have been identified as positive; in this case it measures how many true vessel pixels are successfully identified, reflecting the ability to capture thin and peripheral branches.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \tag{2}$$

Dice Coefficient quantifies the overlap between the predicted and ground truth masks by balancing precision and recall. It is widely used in medical segmentation benchmarks.

$$\text{Dice} = 2 * \text{TP} / (2 * \text{TP} + \text{FP} + \text{FN}) \tag{3}$$

The IoU metric measures the ratio of the intersection area to the union area between prediction and ground truth. It is a stricter metric that penalizes both false positives and false negatives, providing a more conservative measure of segmentation quality.

$$\text{IoU} = \text{TP} / (\text{TP} + \text{FP} + \text{FN}) \tag{4}$$

D. Explainability Methods

To investigate how different explainability techniques perform in the context of coronary artery segmentation, five widely used XAI methods were evaluated. These methods represent different methodological families: gradient-based, perturbation-based, path integrated, and surrogate modeling approaches. Each method was selected to provide a diverse view of explainability approaches, as their underlying assumptions and outputs differ significantly.

Grad-CAM is one of the most established methods for visualizing convolutional neural networks. It computes gradients of the target class with respect to the feature maps of the selected convolutional layer to produce a coarse localization heatmap of discriminative regions. Grad-CAM was included as a baseline for explainability in medical imaging tasks.

Grad-CAM++ extends Grad-CAM by incorporating higher order derivatives into the gradient computation, enabling better capture of multiple relevant regions and small-scale structures. Given the thin and branching morphology of coronary vessels, Grad-CAM++ was

Copyright (c) IARIA, 2026. ISBN: 978-1-68558-362-0

included to determine whether its improved sensitivity could lead to more precise vessel localization compared with standard Grad-CAM.

Score-CAM generates heatmaps by weighting activation maps according to the model's forward prediction scores, thereby eliminating the dependency on gradients. For each feature map, Score-CAM upsamples the activation to input resolution, masks the original image with this map, and re-evaluates the model. The importance of the feature map is then derived from how much the masked input changes the model's confidence relative to the baseline prediction. In our implementation, four scoring strategies were assessed based on the model outputs:

Increase: Measures the confidence gain compared to the baseline prediction, emphasizing features that increase the likelihood of vessel detection.

Absolute: Uses the raw predicted probability for the vessel class, directly weighting features by their contribution to class confidence.

Entropy: Computes the negative entropy of the softmax distribution, rewarding feature maps that produce more confident predictions.

Max-logit: Weighs maps according to the difference between the vessel and background logits, highlighting features that maximize class separation.

By testing multiple scoring functions, we aimed to determine whether vessel-specific signal characteristics could be better captured through alternative weighting schemes. This method has the potential to generate smoother and less noisy explanations for vessel structures, compared to the gradient-based methods analyzed previously.

Integrated Gradients (IG) explain model predictions by integrating gradients along a continuous path from a baseline input to the actual input. To reduce sensitivity to baseline selection, we averaged results over three baselines: a black image, Gaussian noise, and a mean-intensity image. IG provides theoretically grounded attributions that satisfy sensitivity and implementation invariance, and has been widely used in medical imaging. We applied IG to test how path-integrated attributions perform in coronary angiography, where vessel structures are thin and globally sparse. However, due to their gradient-based nature, IG explanations may emphasize vessel edges rather than the vessel interior, which is an important limitation explored in this study.

LIME approximates a model's decision boundary locally by fitting an interpretable surrogate model (e.g., linear regression) to perturbed versions of the input. We implemented LIME with Felzenszwalb superpixel segmentation, which groups neighboring pixels into locally homogeneous regions. This choice was motivated by the need to preserve vessel-like structures while reducing the dimensionality of the perturbation space. LIME was included because it is one of the most widely recognized model agnostic methods and is often used as a baseline in XAI research. In medical imaging, it has been applied to modalities such as chest radiographs and skin lesion images. In the context of coronary angiography, however, the superpixel-based approach introduces challenges: vessels are thin, elongated, and sparse, and a single superpixel may encompass both vessel and background regions. This can reduce the fidelity of the surrogate model and lead to fragmented or misleading explanations, which we explicitly evaluate in our experiments.

E. Explanations Performance Evaluation

To quantitatively evaluate the explanations produced by each XAI method, we implemented a set of vessel-aware metrics tailored for image segmentation. These metrics capture complementary aspects of explanation quality: localization accuracy, ranking ability under class imbalance, overlap with ground truth vessels, and concentration of attribution energy. All heatmaps were normalized to the [0, 1] range prior to evaluation, ensuring consistency across methods.

Pointing Game (PG) measures localization accuracy by testing whether the regions of highest attention overlap with true vessel structures. For each heatmap, we normalize intensities and select pixels above the 80th percentile as high-attention regions. These regions are compared against a dilated ground truth mask, which tolerates small misalignments in thin vessels. The PG score is the ratio of high-attention pixels inside vessels to the total number of high-attention pixels. A higher score indicates better vessel localization.

Average Precision (AP) evaluates the ranking ability of heatmap values under class imbalance, since vessels occupy only a small fraction of the image. We flatten the normalized heatmap into prediction scores and the binary vessel mask into labels, and compute average precision across recall levels:

$$AP = \Sigma_n (R_n - R_{n-1}) P_n \tag{5}$$

R_n and P_n are recall and precision at the *n*-th threshold. If the ground truth mask is empty, AP is set to 0.0. A higher AP score reflects the ability of an explanation to consistently rank vessel pixels above background pixels.

IoU quantifies the overlap between binarized attention maps and vessel structures. Each heatmap is thresholded at the 80th percentile to create a binary mask of salient regions. We then compute:

$$IoU = TP / (TP + FP + FN) \tag{6}$$

IoU penalizes both false positives and false negatives, providing a strict measure of spatial overlap.

Energy Concentration Ratio (ECR) measures how much of the explanation energy is concentrated within vessels compared to background regions. It is defined as:

$$ECR= \Sigma (H*M) / \Sigma H \qquad (7)$$

where H is the normalized heatmap and M is the binary vessel mask. A higher ECR indicates that the majority of attribution energy is focused on true vessel structures rather than distributed across irrelevant regions.

Together, these four metrics provide a comprehensive framework for evaluating XAI methods. PG emphasizes localization of the most salient activations, AP tests ranking performance under imbalance, IoU measures strict spatial overlap, and ECR captures global energy distribution.

IV. Explanations Evaluation Framework

The proposed explainability framework compares each method's heatmap with the ground truth vessel mask for the selected patients. For methods that require a layer choice (Grad-CAM, Grad-CAM++, Score-CAM), we systematically tested multiple candidate layers and retained the configuration that achieved the best metric performance. For Score-CAM, all four scoring strategies were applied on the tested layers, and the best-performing variant was reported. For IG, the attributions were accumulated over 100 steps, and only positive contributions were retained for the final explanation. For LIME, we implemented a vessel-focused setup. Images were segmented into superpixels using the Felzenszwalb algorithm (scale = 50, sigma = 0.3, min_size = 20). A fixed random seed was used to ensure reproducibility. The prediction function was adapted to coronary vessels by combining per-pixel probabilities with a connectivity-based adjustment, favoring explanations that highlight continuous vascular structures. LIME explanations were generated with 1500 perturbation samples and up to 150 interpretable features. To enable overall ranking of XAI methods, we computed an aggregate score defined as the unweighted average of the four evaluation metrics (PG, AP, IoU, ECR) for simplicity, clinical deployment could benefit from task-specific weighting (e.g., higher weight on PG for localization-critical applications). This evaluation framework ensures that each XAI method is tested fairly under well-specified conditions, and that reproducibility is guaranteed by reporting implementation details such as baseline choices, scoring functions, segmentation parameters, and random seeds.

V. Results

This section presents the evaluation results in two parts: first, the segmentation performance of the U-Net model under different preprocessing configurations; and second, a quantitative and qualitative comparison of five XAI methods applied to three patient cases.

A. Model Performance Results

We evaluated the U-Net model under four conditions: baseline (no modifications), post processing only, filtering only, and post processing combined with filtering. The quantitative results are summarized in Table I:

TABLE I. MODEL PERFORMANCE RESULTS FROM EACH TEST

Metrics	TEST 1	TEST 2	TEST 3	TEST 4
Dice	0.653	0.653	0.677	0.679
IoU	0.484	0.485	0.512	0.514
Precision	0.712	0.712	0.720	0.721
Recall	0.603	0.604	0.640	0.642

Post processing consisted of morphological closing (radius=3) followed by removal of small connected components (<100 px). This step produced only marginal gains (+0.001 IoU, +0.001 Recall), suggesting that U-Net predictions were already smooth and contained limited spurious noise.

Filtering, implemented via White Top-Hat vessel enhancement and CLAHE, had a stronger effect. Dice increased from 0.653 to 0.677 (+3.7%), IoU from 0.484 to 0.512 (+5.8%), and recall from 0.603 to 0.640 (+6.1%), while precision remained essentially stable. These improvements indicate that filtering improved vessel contrast, enabling the network to capture more thin and peripheral branches without introducing false positives.

The combination of filtering and post processing yielded the best overall performance (Dice=0.679, IoU=0.514), but only marginally above filtering alone. This suggests that filtering is the dominant factor driving improvements, while post processing provides small refinements.

B. Explanations Performance Results

We evaluated explanation performance on patients 15, 148 and 237. Results are reported per metric (PG, AP, IoU, ECR) and summarized into a total score. We also present heatmaps for qualitative comparison.

As explained in section IV we begin by testing Grad-CAM and Grad-CAM++ on the selected layers to find which one performs the best metrics wise. Across all three patients the best performing layer was "activation" for these methods. Then we tested Score-CAM, this time the best performing layer for all patients was "conv2d_15" and the best scoring method was "entropy". Lastly, we tested IG and LIME with the implementation that was explained in section IV. In Tables II, III, and IV, we can see how each explanation performed, across all metrics for each patient.

TABLE II. EXPLANATIONS PERFORMANCE RESULTS ACROSS EACH METRIC FOR PATIENT 15

Method	PG	AP	IoU	ECR	Score
Grad-CAM	0.286	0.580	0.188	0.465	0.380
Grad-CAM++	0.286	0.580	0.188	0.465	0.380
Score-CAM	0.200	0.416	0.143	0.283	0.260
IG	0.145	0.059	0.067	0.068	0.085
LIME	0.064	0.061	0.040	0.186	0.088

TABLE III. EXPLANATIONS PERFORMANCE RESULTS ACROSS EACH METRIC FOR PATIENT 148

Method	PG	AP	IoU	ECR	Score
Grad-CAM	0.223	0.714	0.129	0.491	0.389
Grad-CAM++	0.223	0.714	0.129	0.491	0.389
Score-CAM	0.191	0.572	0.116	0.231	0.278
IG	0.112	0.038	0.038	0.056	0.061
LIME	0.048	0.054	0.027	0.165	0.074

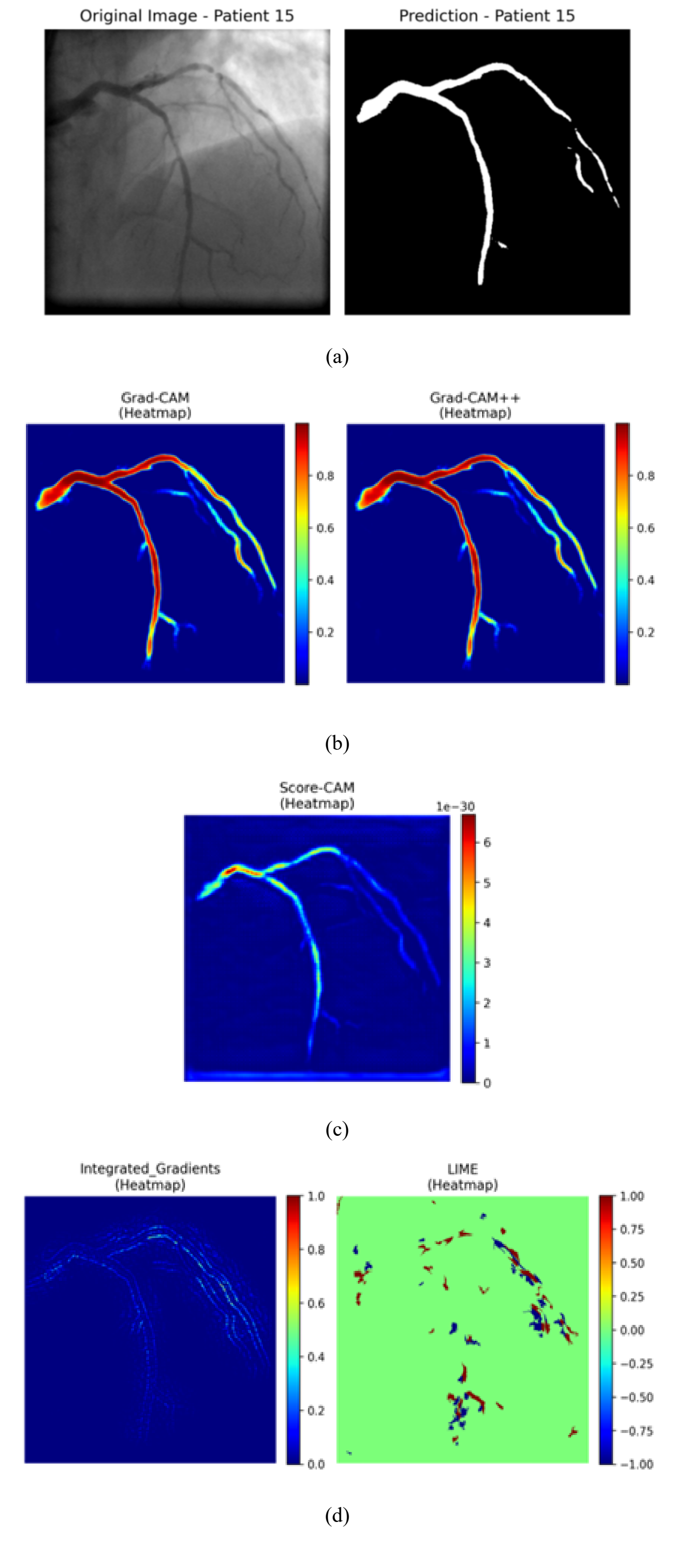
TABLE IV. EXPLANATIONS PERFORMANCE RESULTS ACROSS EACH METRIC FOR PATIENT 237

Method	PG	AP	IoU	ECR	Score
Grad-CAM	0.268	0.691	0.179	0.541	0.420
Grad-CAM++	0.268	0.691	0.179	0.541	0.420
Score-CAM	0.229	0.597	0.160	0.299	0.321
IG	0.128	0.062	0.057	0.097	0.086
LIME	0.055	0.083	0.036	0.268	0.111

Overall, Grad-CAM and Grad-CAM++ performed the best across each metric for all three patients, while having the same values. Score-CAM achieved middling results, IG and LIME performed the worst.

Now, for each patient, we present the heatmap from each explanation method. For added context we present the original X-ray image for each patient from the ARCADE dataset and the Ground Truth (GT) mask that was used to evaluate the explanations.

We present images for patients 15, 148 and 237 in Figures 1, 2, and 3:



(a)

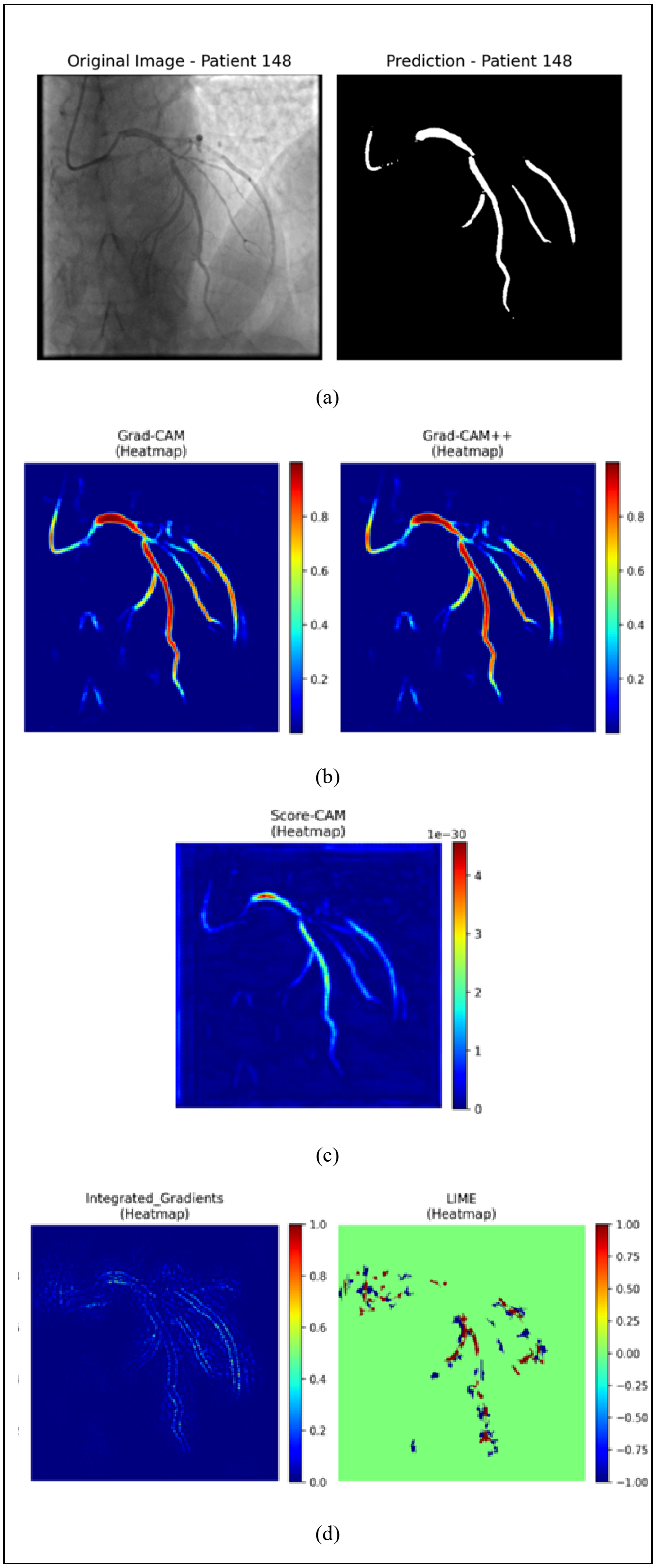
(b)

(c)

(d)

Figure 1. Comparison of heatmaps for patient 15: (a) Original image and GT mask, (b) Grad-CAM and Grad-CAM++, (c) Score-CAM, (d) IG and LIME.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org



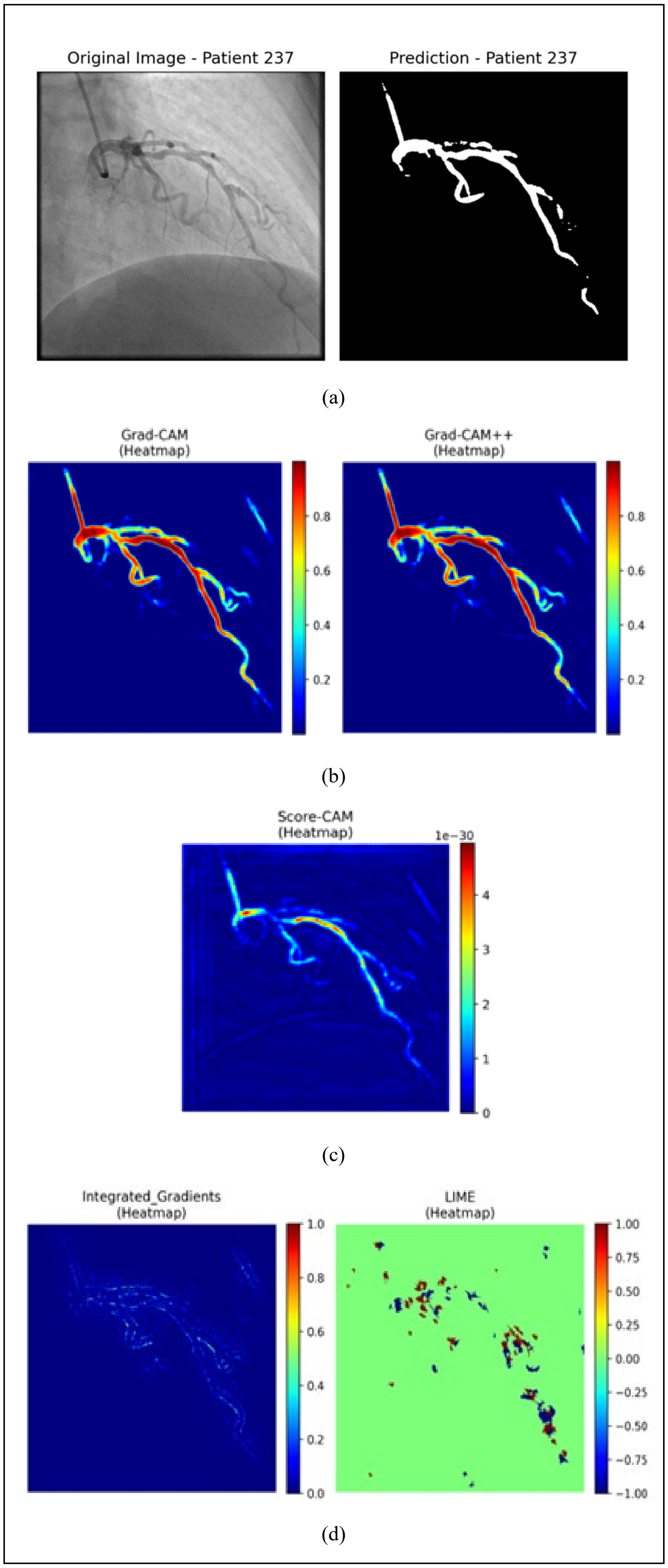
(a)

(b)

(c)

(d)

Figure 2. Comparison of heatmaps for patient 148: (a) Original image and GT mask, (b) Grad-CAM and Grad-CAM++, (c) Score-CAM, (d) IG and LIME.



(a)

(b)

(c)

(d)

Figure 3. Comparison of heatmaps for patient 237: (a) Original image and GT mask, (b) Grad-CAM and Grad-CAM++, (c) Score-CAM, (d) IG and LIME.

We observe that each ground truth mask is able to depict major coronary arteries with good continuity, making most of the vessel tree structure visible. However, some terminal branches don't appear in them and there seem to be some discontinuities in some parts of each vessel structure.

The heatmaps generated by Grad-CAM and Grad-CAM++ are able to highlight the main coronary arteries with high intensity values (0.8–1.0) and assigning lower activations (0.4–0.6) to peripheral branches. For each patient these two methods generate almost identical heatmaps. Score-CAM is able to some extent, portray the vessel tree, but from the range of values on each heatmap, it has problems giving more attention to bigger vessels than the smaller branches. It also has a lot of background noise. IG primarily emphasizes vessel edges, while LIME highlights isolated regions rather than reconstructing the full vascular structure.

VI. DISCUSSION

The results indicate that Grad-CAM and Grad-CAM++ produced the most clinically meaningful explanations, achieving the highest aggregate scores and generating continuous activations along coronary centerlines. Grad-CAM++ did not outperform Grad-CAM, suggesting that higher-order derivatives add limited value for thin vessel structures. Score-CAM localized vessels moderately well, but struggled to distinguish main vessels from smaller branches. Its superior performance of the entropy-based scoring suggests that uncertainty reduction is a better proxy for vessel importance. The optimal layer selection proved consistent across all three patients, suggesting that layer preferences are method-specific rather than image-specific, which simplifies clinical deployment. Moreover, Grad-CAM benefits from final high-level features, while Score-CAM's perturbation approach works better at intermediate feature levels where spatial resolution is higher. IG and LIME failed to capture vessel interiors or continuity, instead producing edge- or spot-like activations, reflected in very low composite scores.

The four evaluation metrics offered complementary insights. PG assessed the localization of peak activations, AP measured ranking under class imbalance, IoU captured strict overlap after binarization, and ECR quantified the distribution of energy attribution. Using both continuous and thresholded metrics revealed distinct failure modes, such as diffuse attention yielding high ECR but low IoU. These findings highlight the need for multi-metric evaluation in medical XAI.

Finally, qualitative inspection showed that CAM-based methods consistently highlighted large vessel trunks, while smaller branches were less emphasized. In some cases, CAM heatmaps appeared to reveal details not present in the ground truth, suggesting potential annotation limitations.

VII. Conclusions and Future Work

We presented a systematic framework for evaluating explainability methods in coronary artery segmentation using the ARCADE dataset and a U-Net backbone. Our findings show that CAM-based methods, particularly Grad-CAM and Grad-CAM++, provide the most reliable explanations, while perturbation and gradient-integration approaches, such as Score-CAM, IG, and LIME were less effective for thin vascular structures.

Future work will expand this evaluation state-of-the-art architectures such as TransUNet or nnU-Net, larger and more diverse datasets, incorporate additional XAI methods (e.g., SHAP, concept-based explanations), and include expert evaluation by cardiologists to assess clinical usability while they also assess whether CAM-based explanations identify genuine anatomical structures missed during manual annotation. Integrating explainability into real-time clinical workflows and extending the framework to multi-class vessel segmentation are also promising directions.

References

[1] G. A. Roth *et al.*, "Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study," *J. Am. Coll. Cardiol.*, vol. 76, no. 25, pp. 2982–3021, 2020.

[2] M. Kolossváry, B. Szilveszter, P. Kolossváry, I. Karády, and P. Maurovich-Horvat, "Radiomic features are superior to conventional quantitative computed tomographic metrics to identify coronary plaques with napkin-ring sign," *Circ. Cardiovasc. Imaging*, vol. 10, no. 12, pp. 1–10, 2019.

[3] B. K. Nallamothu, M. H. Spertus, D. A. Lansky, J. D. Hofer, and E. R. Bates, "Comparison of clinical interpretation with visual assessment and quantitative coronary angiography in patients undergoing percutaneous coronary intervention in contemporary practice," *Circulation*, vol. 127, no. 17, pp. 1793–1800, 2013.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.

[5] T. T. D. Mahendiran, P. Rajan, S. P. Singh, and A. Verma, "AngioPy segmentation: Deep learning tool for coronary segmentation," *Int. J. Cardiol., in press, 2025.*

[6] M. A. Popov, V. A. Soldatov, and I. A. Solovyev, "Dataset for automatic region-based CAD diagnostics using X-ray angiography images," *Sci. Data*, vol. 20, 2024.

[7] C. Zhao, C. Yang, J. Gao, and S. Xia, "AGMN: Graph matching network for coronary artery semantic labeling," *Pattern Recognit.*, vol. 143, 2023.

[8] C. Rudin, "Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[9] D. Bhati, F. Neha, and M. Amiruzzaman, "A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging," *J. Imaging*, vol. 10, 2024.

[10] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, and D. Parikh, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[11] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[12] H. Wang, Z. Wang, H. Du, Y. Shen, and Y. Pu, “Score-CAM: Score-weighted visual explanations for convolutional neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 111–119, doi: 10.1109/CVPRW50498.2020.00020.

[13] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’ Explaining the predictions of any classifier,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.

[15] J. Zhang and K. A. Chan, “Top-down neural attention by excitation backprop,” *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, 2018.

[16] J. Davis and M. Goadrich, “The relationship between precision–recall and ROC curves,” in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233–240.

[17] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, “What is a good evaluation measure for semantic segmentation?” in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2013, pp. 32.1–32.11.

[18] V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized input sampling for explanation of black-box models,” in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018.

[19] T. Do, P. Huynh, M. Nguyen, and V. Nguyen, “An XAI-based deep learning framework for coronary artery disease diagnosis using SPECT MPI polar map images,” in *Proc. Seventh Int. Conf. Res. Intell. Comput. Eng.*, 2022, pp. 235–241, doi: 10.15439/2022R06.

[20] N. I. Papandrianos, I. A. Panagiotopoulos, G. T. Papageorgiou, and K. N. Sidiropoulos, “An explainable classification method of SPECT myocardial perfusion images in nuclear cardiology using deep learning and Grad-CAM,” *Appl. Sci.*, vol. 12, p. 7592, 2022.

[21] M. Goettling, J. M. Schwenk, F. Kragness, J. Tomaszewski, and R. F. Speier, “xECGArch: A trustworthy deep learning architecture for interpretable ECG analysis considering short-term and long-term features,” *Sci. Rep.*, vol. 14, no. 13122, 2024.

[22] M. Bhandari, R. Singh, S. Gupta, A. Kumar, and R. Sharma, “Explanatory classification of CXR images into COVID-19, pneumonia, and tuberculosis using deep learning and XAI,” *Comput. Biol. Med.*, vol. 150, p. 106156, 2022.

[23] H. S. Anand, R. K. Sharma, and P. Gupta, “Coronary vessel segmentation in X-ray using U-Net,” in *Lecture Notes in Networks and Systems*, vol. 969, Springer, 2024, pp. 57–66.

[24] Z. Gao, Z. Li, H. Ma, and L. Zhang, “Vessel segmentation for X-ray coronary angiography using ensemble methods with deep learning and filter-based features,” *BMC Med. Imaging*, vol. 22, p. 10, 2022.

[25] J. Jiménez-Carretero, R. Ruiz-Sarmiento, J. M. Górriz, and J. Ramírez, “CADICA: A dataset for coronary artery disease analysis in invasive coronary angiography,” arXiv preprint arXiv:2402.00570, 2024.

[26] Z. Wang, Y. Guo, X. Yang, and J. Liu, “XCAD: A dataset for coronary artery segmentation in X-ray angiography,”
IEEE Access, vol. 8, pp. 189030–189041, 2020.

[27] C. Acebes, C. Tejos, and P. Irarrazaval, “The centerline–cross entropy loss for vessel-like structure segmentation,” in *Proc. MICCAI*, 2024, doi: 10.1007/978-3-031-72117-5_30.