Supplementary

Table S1 End-to-end deep learning approaches for stenosis detection

Authors	Images	Model	Detected stenosis
Classification			
Ovalle-Magallanes *et al.* (13)	250 XCA images	Squeeze-and-excitation attention, efficient channel attention, convolutional block attention	Stenosis, nonstenosis (accuracy =87.87%, recall=86.10%, F1-score =87.32%)
Ovalle-Magallanes *et al.* (14)	250 XCA images	Hybrid classical-quantum network	Stenosis, nonstenosis (accuracy =91.80%, recall =94.92%, and F1-score =91.80%)
Ovalle-Magallanes *et al.* (17)	250 real XCA images, 10,000 synthetic images	Hierarchical Bézier-based generative model	Stenosis, nonstenosis (accuracy =89.34%, recall =90.31%, sensitivity =87.46%, F1-score =88.80%)
Ovalle-Magallanes *et al.* (16)	Dataset 1: 6769 XCA images Dataset 2: 26,699 XCA images	Squeeze-and-excitation attention mechanism and depthwise separable convolution	Stenosis, nonstenosis (dataset 1: accuracy =95.49%, sensitivity =63.20%, precision =59.91%, F1-score =61.03%; dataset 2: accuracy =95.43%, sensitivity =87.92%, precision =89.44%, F1-score=89.44%)
Ovalle-Magallanes *et al.* (15)	250 real XCA images, 10,000 synthetic images	InceptionV3	Stenosis, nonstenosis (accuracy =95%, precision =93%, F1-score =95%)
Moon *et al.* (18)	452 RCA images	InceptionV3 with attention modules	Stenosis, nonstenosis (AUC =0.971)
Stralen *et al.* (19)	16,980 RCA images	EfficientDet	Stenosis, nonstenosis (precision =67%)
Cong *et al.* (23)	194 XCA videos	LSTM and InceptionV3	<25% stenosis, 25–99% stenosis, total occlusion (AUC 0.91/0.85 for RCA/LCA)
Danilov *et al.* (25)	8325 XCA images	Faster-RCNN ResNet-50 V1	Small, medium, and large stenosis (precision =92%)
Bounding box			
Rodrigues *et al.* (21)	1593 XCA images	RetinaNet	Stenosis, nonstenosis (72%/70% recall for RCA/LCA)
Han *et al.* (22)	233 XCA images	Proposal-shifted spatial-temporal transformer	Stenosis, nonstenosis (F1-score =90.88%)
Pang *et al.* (33)	166 XCA images	ResNet-50 with feature pyramid network	Stenosis, nonstenosis (precision =94.87%, sensitivity =82.22%, F1-score =88.1%)
Cong *et al.* (24)	13,744 XCA images from 230 patients	InceptionV3, class activation map with feature pyramid network	<25%, >25%, and/or total occlusion (accuracy =85%, sensitivity =96%, AUC =0.86)

https://dx.doi.org/10.21037/qims-24-2415

Appendix 1 Adaptation of backbone architectures for semantic segmentation

To ensure a fair and consistent comparison across architectures, we adapted residual neural network (ResNet), MobileNetV2, and Vision Transformer models, originally designed for image classification, to perform dense semantic segmentation. These adaptations involved removing the final classification layers and constructing decoder modules capable of producing full-resolution feature maps. All models were configured to output feature maps with a spatial resolution of 512 × 512 × 32 channels, which were then processed by task-specific output heads.

The ResNet backbone (ResNet18) was used as the encoder. The final fully connected classification layer was removed. The encoded feature maps were passed through a custom decoder consisting of a series of upsampling blocks. Each block included a bilinear upsampling layer followed by two convolutional layers with rectified linear unit (ReLU) activation. The number of channels was progressively halved at each stage to reconstruct spatial detail. The final decoder output was a 512 × 512 × 32 feature map used for downstream segmentation tasks.

The MobileNetV2 encoder was constructed using the MobileNetV2 architecture with the global pooling and classification layers removed. A lightweight decoder comprising transpose convolution layers and convolutional blocks was appended to upsample the feature maps. At each scale, feature maps were processed by residual convolution blocks with batch normalization and ReLU activation. Upsampling was performed progressively (e.g., from 8 × 8 to 16×16, etc.) until reaching 512 × 512. The final decoder output was a 32-channel feature map aligned with the input resolution.

The Vision Transformer model was adapted by removing the classification head and reshaping the transformer's patch-wise output into a spatial feature map. A custom decoder composed of successive ConvTranspose2d layers and convolutional blocks was applied to upsample the transformer outputs from 28 × 28 to the original 512 × 512 resolution. The decoder included convolutional layers with ReLU activation to refine features at each scale. The final output was a 512 × 512 × 32 feature map, consistent with the decoder outputs of other architectures.

All models were adapted to produce dense spatial feature maps that matched the input resolution and feature dimensionality. This standardization allowed for a direct and meaningful comparison between classification-based backbones and fully convolutional segmentation models such as U-Net and U-Net++, which inherently support pixel-wise prediction.

https://dx.doi.org/10.21037/qims-24-2415