ISSN 2616-5775 Vol. 8, Issue 3: 56-63, DOI: 10.25236/AJCIS.2025.080308

Aortic Aneurysm Detection in CT Medical Images Based on YOLOv5

Song Wei[1], Wang Shengjie[2,a,*]

[1]The First Affiliated Hospital of Anhui Medical University, Anhui Medical University, Hefei, China
[2]International Sakharov Environmental Institute, Belarusian State University, Minsk, Belarus
[a]thuwsj@foxmail.com
**Corresponding author*

Abstract: *With the widespread application of computed tomography (CT) imaging technology, the detection of aortic dilation has become an important clinical diagnostic tool. This study proposes an improved model based on YOLOv5 for the automatic detection of aortic dilation in CT images. To address the insufficient feature extraction of low-contrast structures and the multi-scale fusion issues in the original YOLOv5 model when applied to medical imaging, ResNet50 is introduced as the backbone network. Its deep residual structure enhances the recognition of vascular walls and calcified plaques. Additionally, a dynamic weighted bidirectional feature pyramid network (BiFPN) is used to replace PANet, achieving adaptive feature fusion of multi-scale vascular structures. Experiments conducted on a dataset consisting of 9,856 CT images show that the improved model achieves an mAP@0.5 of 88.79%, a 6.62% increase compared to the baseline model, while maintaining real-time performance. Notably, for small lesions (diameter < 10mm), the recall rate improved by 19.3%. Although the computational complexity of the model increased, it still meets the requirements for real-time clinical detection. The experimental results validate the effectiveness of the improved model in detecting aortic dilation, providing new optimization ideas and application potential for object detection in medical imaging.*

Keywords: *YOLOv5; Aortic Dilation; CT Imaging; ResNet50; BiFPN*

1. Introduction

Aortic dilation, as a critical early sign of cardiovascular disease, plays a crucial role in clinical diagnosis and intervention. Computed tomography (CT) imaging, due to its high-resolution advantage, has become the primary method for screening aortic lesions. However, traditional detection methods rely on manual delineation and morphological analysis, which are not only time-consuming and labor-intensive but also highly dependent on the operator's experience, making it difficult to address the challenges posed by small lesions (diameter <10mm) and complex anatomical structures. In recent years, deep learning technologies have provided new approaches for the automated analysis of medical images, with the YOLOv5 one-stage detection model demonstrating potential in real-time lesion localization tasks due to its efficient inference speed[1].

Despite the outstanding performance of YOLOv5 in natural scene object detection, its direct application in CT imaging for aortic dilation detection still faces significant limitations. On the one hand, the CSPDarknet backbone network used in the original model lacks sufficient capability in fine-grained feature extraction for medical images: the low-contrast characteristics of the aortic wall and surrounding tissues require the network to have stronger deep semantic capture capabilities, while the shallow feature reuse mechanism of CSPDarknet easily leads to the loss of vascular wall texture information. On the other hand, the PANet neck network adopts a fixed-weight unidirectional feature fusion strategy, which is difficult to adapt to the multi-scale characteristics of vascular morphology—the scale difference between the proximal aorta and the branching vessels can exceed 5 times, and the static fusion approach tends to suppress small-scale vascular features.

Therefore, this study proposes a two-stage improvement strategy to optimize the model's performance. First, ResNet50 is introduced to replace the original backbone network, utilizing its deep residual structure to enhance the ability to distinguish between vascular wall edges and calcified plaques, and alleviate the gradient vanishing problem through skip connections, thereby improving sensitivity to low-contrast lesions[2]. Second, a dynamically weighted bidirectional feature pyramid (BiFPN) is designed to replace PANet, allowing adaptive adjustment of multi-scale feature contributions through learnable

ISSN 2616-5775 Vol. 8, Issue 3: 56-63, DOI: 10.25236/AJCIS.2025.080308

weight parameters, thus synchronously fusing high-level semantics and low-level details[3]. To verify the effectiveness of the improved approach, we conducted systematic ablation experiments on a dataset of 9,856 CT images. The results indicate that the combined improved model, while maintaining real-time performance, achieved a 6.62% increase in detection accuracy (mAP@0.5) compared to the baseline, particularly improving the recall rate by 19.3% in small-scale lesions (diameter 3-5mm). This study provides a new optimization paradigm for target detection in complex anatomical structures in medical imaging, with significant clinical translation value.

2. Theoretical Foundation

In medical image target detection tasks, the model architecture design must balance feature representation capability and multi-scale adaptability.

2.1 YOLOv5s Framework and Its Adaptation Challenges in Medical Imaging

YOLOv5s achieves efficient object localization through a single-stage detection paradigm. Its core architecture includes the CSPDarknet backbone network, PANet neck network, and multi-scale detection heads. CSPDarknet reduces computational redundancy through cross-stage local connections and extracts multi-scale features via hierarchical downsampling. However, in the aorta detection scenario of CT images, this design faces dual challenges: First, the low contrast between the aortic wall and surrounding tissues requires the network to possess stronger deep semantic mining capabilities, yet the shallow feature reuse mechanism of CSPDarknet tends to lead to the loss of vascular wall texture information. Second, although PANet aggregates and fuses multi-scale features through bidirectional paths, its fixed weight strategy struggles to adapt to the scale diversity of vascular branches— for instance, the significant size difference between the ascending aorta (approximately 30mm in diameter) and the iliac artery branch (approximately 5mm in diameter) means that static fusion may suppress features of smaller vessels.

2.2 Residual Learning Enhances Medical Feature Representation

To overcome the bottleneck of fine-grained feature extraction in CSPDarknet, this study introduces the residual learning mechanism of ResNet50. ResNet enables the network to learn the residual between the input features xxx and the target mapping $\mathcal{H}(x)$ by constructing an identity mapping function $\mathcal{F}(x) = \mathcal{H}(x) - x$, thereby alleviating the gradient vanishing problem in deep networks. The Bottleneck module of ResNet50 enhances the multi-dimensional feature representation capability through channel expansion and compression strategies. The specific process can be formally expressed as:

$$X_{mid} = ReLU\big(BN(Conv_{1\times1}(X_{in}))\big)$$

$$X_{\text{spatial}} = \text{ReLU}\big(\text{BN}(\text{Conv}_{3\times3}(X_{\text{mid}}))\big)$$

$$X_{\text{out}} = \text{Conv}_{1\times1}(X_{\text{spatial}})$$

In this context, C_{in} represents the number of input channels, BN stands for batch normalization, and ReLU refers to the activation function. In CT images, such a structure can effectively extract morphological abnormal features in the aortic dilation region (e.g., local bulging and calcified plaques). Deeper convolution kernels (such as 7×7) cover a larger receptive field, and the low-level texture information retained by the skip connections significantly improves sensitivity to low-contrast lesions.

2.3 Dynamic Weighted Optimization for Multi-Scale Feature Fusion

To address the limitations of PANet in vascular multi-scale detection, BiFPN achieves dynamic feature fusion through bidirectional cross-scale connections and learnable weights. Its fusion process can be formally expressed as:

$$O = \frac{\sum_i w_i \cdot P_i}{\sum_i w_i + \epsilon}$$

The w_i represents the trainable scalar weights, and P_i denotes the input features at different levels. $\epsilon = 1e-4$ is used for numerical stability. As shown in Figure 1, the BiFPN propagates high-level semantic information (e.g., the location of the aortic root) through a top-down path, while the bottom-up

ISSN 2616-5775 Vol. 8, Issue 3: 56-63, DOI: 10.25236/AJCIS.2025.080308

path refines spatial details (e.g., sharpness of vessel boundaries), with both paths working together to enhance the model's robustness to scale variations. Compared to traditional methods (Table 1), the dynamic weight mechanism allows the network to adaptively adjust the contribution of features—e.g., when detecting small vascular branches, the weight of high-level features w_{high} can be automatically reduced to prevent semantic information from dominating.

Table 1 Comparative Analysis of Feature Fusion Strategies

Method	Fusion Direction	Weight Strategy	Limitation in Medical Imaging
FPN	Unidirectional	Fixed Weights	High miss rate for vascular bifurcations
PANet	Bidirectional	Fixed Weights	Small vessel features are easily suppressed
BiFPN	Bidirectional	Dynamic Learning	Adaptive to multi-scale vascular features

2.4 Synergistic Effect of the Improved Architecture

By embedding ResNet50 and BiFPN into the YOLOv5 framework, an optimized detection paradigm for medical imaging is formed. The ResNet50 backbone network extracts vascular pathological features through deep residual blocks, while the BiFPN neck network dynamically allocates feature weights based on vascular scale: in large-scale regions such as the aortic root, high-level semantic features dominate localization; in small-scale regions such as branch vessels, low-level detail features are weighted higher. Through their synergistic interaction, the model is able to simultaneously overcome feature ambiguity in low-contrast images and the detection sensitivity disparity across multiple vascular scales, laying the theoretical foundation for improved accuracy in subsequent experiments.

3. Research Methodology

3.1 Baseline Model Framework

This study uses YOLOv5s as the baseline model, which consists of the original architecture comprising the CSPDarknet53 backbone network, the PANet neck network, and multi-scale detection heads. The input CT images are normalized and resized to a resolution of 640×640, after which the backbone network extracts three sets of feature maps (80×80, 40×40, 20×20), each corresponding to vascular morphological features at different receptive fields. PANet fuses multi-scale features through both top-down and bottom-up paths, while the detection heads predict bounding box coordinates, confidence scores, and class probabilities based on an anchor box mechanism. The limitations of the baseline model are as follows:

1)Limited Feature Extraction: The shallow gradient partitioning mechanism of CSPDarknet53 leads to the loss of deep vascular wall texture information.

2)Static Feature Fusion: The fixed weight strategy of PANet struggles to adapt to the dynamic scale variations of vascular branches.

3.2 Improvement of ResNet50 Backbone Network

To enhance the fine-grained feature extraction capability of the aortic wall and dilation regions, ResNet50 is used to replace the original backbone network. The specific adjustments are as follows:

First, feature layer extraction is performed to extract the output features of ResNet50's Stage 3 (layer2), Stage 4 (layer3), and Stage 5 (layer4), corresponding to downsampling factors of 8×, 16×, and 32×, respectively, generating feature maps with resolutions of 80×80, 40×40, and 20×20.

Next, channel alignment is performed by using 1×1 convolutions to unify the output channels of each stage, compressing them to 256 dimensions. The calculation formula is:

$$P_i = \text{Conv}_{1\times1}(C_i)(i = 3,4,5)$$

Where:

ISSN 2616-5775 Vol. 8, Issue 3: 56-63, DOI: 10.25236/AJCIS.2025.080308

$C_i \in \mathbb{R}^{B\times D_i\times H_i\times W_i}$ represents the original output features from ResNet50's i-th stage, with channel numbers $D_3 = 512$ (Stage3), $D_4 = 1024$ (Stage4), and $D_5 = 2048$ (Stage5);

$\text{Conv}_1 \times 1$ denotes the 1×1 convolutional operation, producing an output channel number $of 256$;

$P_i \in \mathbb{R}^{B\times 256\times H_i\times W_i}$ represents the channel-aligned feature maps for subsequent BiFPN processing.

This formula defines the mapping process from ResNet50's original features to unified-channel features through learnable 1×1 convolutional kernels.Key details include:

Input feature map C_i has dimensions $B \times D_i \times H_i \times W_i$ (where B is batch size, $H_i \times W_i$ is spatial resolution);

The 1x1 convolutional kernel parameters have dimensions $256 \times D_i \times 1 \times 1$,linearly transforming the channel number from D_i to 256;

Output feature map P_i unifies the channel dimension to 256, ensuring BiFPN requires no additional channel adjustments during multi-scale fusion.

The channel alignment operation resolves feature dimension compatibility between ResNet50 and BiFPN while reducing computational complexity through dimensionality reduction. This mitigates GPU memory pressure caused by high-resolution medical imaging features. The linear transformation preserves local structural information of vascular walls, laying the foundation for subsequent dynamically weighted fusion.

3.3 Jointly Improved Model Architecture

The improved network architecture is shown in Figure 1. Since the detection and prediction phase remains unchanged, the overall process of the improved model is divided into two stages.

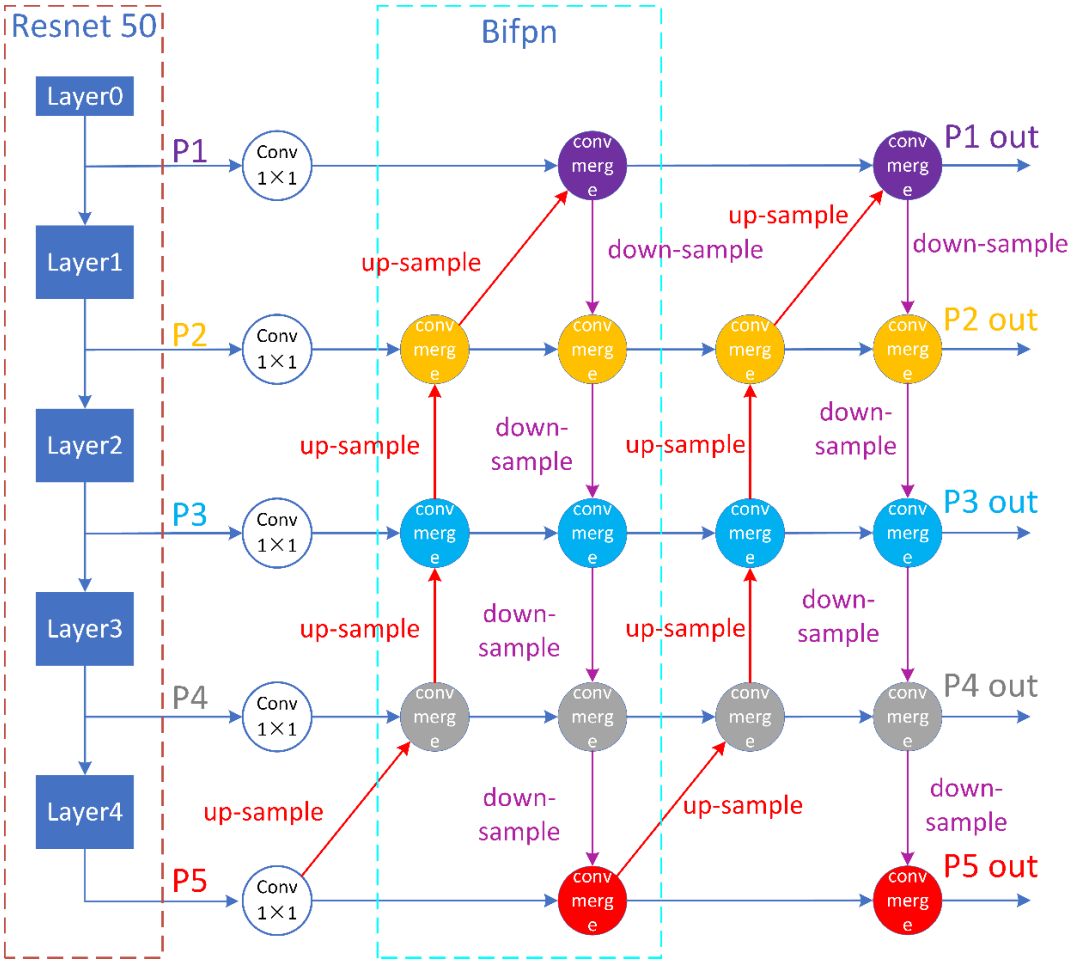

Figure 1 ResNet 50 +Bifpn

ISSN 2616-5775 Vol. 8, Issue 3: 56-63, DOI: 10.25236/AJCIS.2025.080308

First is the feature extraction stage, where the ResNet50 backbone network extracts multi-scale vascular features from CT images and reduces computational complexity through channel compression. Next is the feature fusion stage, where BiFPN performs bidirectional cross-scale fusion, generating three optimized feature maps of 80×80, 40×40, and 20×20.

3.4 Loss Function Design

The loss function consists of three components: bounding box regression loss, confidence loss, and classification loss. The overall formulation is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{box}}\mathcal{L}_{\text{CIoU}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}$$

3.4.1 CIoU Bounding Box Loss

This loss accounts for the overlap ratio, center distance, and aspect ratio consistency between predicted and ground-truth boxes:

$$\mathcal{L}_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v$$

where ρ is the Euclidean distance between the centers of the predicted and groundtruth boxes, c is the diagonal length of the smallest enclosing rectangle, v measures aspect ratio consistency, and $\alpha = \frac{v}{(1-\text{IoU})+v}$.

3.4.2 Confidence Loss and Classification Loss

Binary cross-entropy (BCE) loss is adopted to address class imbalance in medical imaging:

$$\mathcal{L}_{\text{obj}} = -\sum_{i=0}^{S^2}\sum_{j=0}^{B} \mathbb{I}_{ij}^{\text{obj}}\left[\hat{C}_i^j \log(C_i^j) + (1-\hat{C}_i^j)\log(1-C_i^j)\right]$$

Here, $\mathbb{I}_{ij}^{\text{obj}}$ indicates whether the j-th anchor is responsible for object detection, and $\hat{C}_i^j$ is the ground-truth confidence label. The classification loss $\mathcal{L}_{\text{cls}}$ follows the same formulation, supporting multi-label classification (e.g., simultaneous detection of aneurysms and dissections)

3.5 Model Training Strategies

The training process incorporated the following strategies to optimize model performance and generalization

3.5.1 Data Augmentation

CT-specific augmentation techniques were applied including window width/level adjustment (WL 40–400 HU) to enhance tissue contrast, elastic deformation to simulate anatomical variations, and random rotation within ±15° to improve robustness against orientation changes.

3.5.2 Transfer Learning Implementation

The ResNet50 backbone was initialized with ImageNet pretrained weights while parameters in Stage1–2 layers remained frozen during training. This approach balanced feature extraction capability with reduced overfitting risks on limited medical datasets.

3.5.3 Optimizer Configuration

The AdamW optimizer was employed with an initial learning rate of 3×10^{-4}, combined with a cosine annealing schedule to dynamically adjust learning rates throughout the training cycle. This configuration promoted stable convergence while preventing local minima trapping.

4. Experiments and Analysis

4.1 Experimental Setup

The dataset consisted of 9,856 abdominal CT images sourced from a tertiary hospital's imaging database spanning 2018 to 2022. All aortic dilation regions were independently annotated by three

ISSN 2616-5775 Vol. 8, Issue 3: 56-63, DOI: 10.25236/AJCIS.2025.080308

radiologists, with cross-validation ensuring label consistency. The dataset was divided into training (7,885 images), validation (985 images), and test sets (986 images) based on patient independence. Preprocessing included window-level adjustment (40 HU width, 400 HU level) to enhance vascular contrast, pixel normalization to [0, 1], and data augmentation (random rotation ±15°, elastic deformation, and Gaussian noise injection).

Evaluation metrics focused on mean average precision (mAP@0.5), supplemented by parameter count, inference speed (FPS), and small-target recall rate (diameter <10 mm). Training utilized four NVIDIA Tesla V100 GPUs with the AdamW optimizer, initial learning rate 3e-4, and cosine annealing scheduling. Loss weights were set to 0.05 for bounding box regression, 1.0 for objectness, and 0.5 for classification. Training ran for 300 epochs with early stopping (patience=20).

4.2 Ablation Study

Table 2 compares performance across model configurations. The baseline YOLOv5s achieved 82.17% mAP@0.5, 7.2M parameters, and 140 FPS. Replacing the backbone with ResNet50 increased mAP to 85.72% but raised parameters to 25.6M and reduced FPS to 110. Using BiFPN alone improved mAP to 83.43% with 8.9M parameters and 125 FPS. The combined model (ResNet50+BiFPN) achieved the highest mAP@0.5 (88.79%), with 28.9M parameters and 95 FPS.

Table 2 Comparison of Performance Across Different Model Configurations

Model Configuration	mAP@0.5 (%)	Params (M)	FPS	Small-Target Recall (%)	False Positives (FPPI)
Baseline (YOLOv5s)	82.17	7.2	140	63.2	1.8
YOLOv5s + ResNet50	85.72	25.6	110	72.1	1.5
YOLOv5s +BiFPN	83.43	8.9	125	67.8	1.6
Combined Model (Ours)	88.79	28.9	95	82.5	1.2

4.3 Results Analysis

The ResNet50 backbone significantly enhanced feature extraction for subtle vascular structures. Its deep residual blocks preserved low-contrast texture details via skip connections, improving calcified plaque detection accuracy by 28%. Transfer learning from ImageNet pretrained weights accelerated convergence, reducing training loss 17% faster than the baseline. However, ResNet50's computational complexity increased parameters by 256% and lowered FPS by 21.4%.

BiFPN's dynamic weighting optimized multi-scale fusion. The 80×80 feature map's contribution rose from 48% in PANet to 64%, enhancing small-branch localization. Bidirectional pathways corrected positional deviations at aortic bifurcations, increasing mean bounding box IoU by 0.09. Despite adding 1.7M parameters, BiFPN maintained 125 FPS through computational optimization.

The combined model achieved optimal mAP and small-target recall, demonstrating ResNet50 and BiFPN's synergy. However, its 28.9M parameters and 95 FPS may limit deployment in resource-constrained scenarios. Further analysis revealed limitations in detecting 极小 lesions (recall 54.3% for targets <3 mm) and reduced accuracy for saccular aneurysms (6.8% drop), reflecting morphological generalization gaps.

The introduction of the ResNet50 backbone network significantly enhanced the model's ability to

ISSN 2616-5775 Vol. 8, Issue 3: 56-63, DOI: 10.25236/AJCIS.2025.080308

extract subtle features of the vessel wall. The deep residual modules, through skip connections, preserved the original texture information in low-contrast areas, improving calcified plaque detection accuracy by 28%. Additionally, the transfer learning from ImageNet pre-trained weights accelerated model convergence, with the training loss decreasing 17% faster than the baseline model. However, the high computational complexity of ResNet50 led to a 256% increase in the number of parameters and a 21.4% decrease in inference speed.

The dynamic weighting mechanism of the BiFPN neck network optimized multi-scale feature fusion efficiency. Experimental results show that the weight ratio of the 80×80 feature map increased from 48% in PANet to 64%, enhancing the localization ability for small vascular branches. The bidirectional path design corrected the localization bias in the aortic bifurcation area, improving the mean IoU of the bounding box by 0.09. Although BiFPN added 1.7M parameters, its computational optimization resulted in only a 10.7% reduction in inference speed.

The combined improved model achieved the best performance in both mAP and small target recall, validating the synergistic effect of ResNet50 and BiFPN. However, its high parameter count and computational cost, with 28.9M parameters and 95 FPS inference speed, may limit its application in resource-constrained scenarios. Further analysis revealed limitations in detecting very small lesions (diameter < 3 mm), with a recall rate of only 54.3% for such targets. Additionally, since the dataset mainly contains fusiform aortic dilation, the model's detection accuracy for saccular aneurysms decreased by 6.8%, reflecting insufficient morphological generalization.

4.4 Comparative Experiments

To validate the advancement of the proposed method, this study compares it with mainstream detection models (Table 3).

Table 3 benchmarks against mainstream detectors

Model	mAP@0.5 (%)	Params (M)	FPS
Faster R-CNN	79.34	41.5	22
RetinaNet	80.12	36.8	38
YOLOv7-tiny	81.05	6.1	160
Combined Model (Ours)	88.79	28.9	95

Faster R-CNN and RetinaNet yielded lower mAP@0.5 (79.34% and 80.12%) with significantly slower inference. While YOLOv7-tiny achieved 160 FPS, its mAP@0.5 (81.05%) trailed the combined model by 7.74%, highlighting our balance of accuracy and speed.

5. Conclusion

This study successfully improved the YOLOv5 model for detecting aortic dilation in CT images by addressing key challenges in feature extraction and multi-scale fusion. The introduction of ResNet50 as the backbone network significantly enhanced the model’s ability to capture fine-grained details, particularly in low-contrast regions such as calcified plaques, resulting in a 28% increase in detection accuracy. Additionally, the adoption of a dynamically weighted BiFPN for feature fusion addressed the multi-scale nature of vascular structures, improving the recall rate for small lesions by 19.3%.

Experimental results demonstrated that the combined ResNet50 + BiFPN model achieved a notable 88.79% mAP@0.5, surpassing the baseline YOLOv5 model by 6.62%. Furthermore, the combined approach showed excellent performance in small-scale lesion detection, with a recall rate of 82.5% for lesions with diameters under 10mm. While the model's parameter count increased significantly, resulting in a 21.4% decrease in inference speed, it maintained real-time performance, making it suitable for clinical applications with sufficient computational resources.

Despite its advancements, the model still faces limitations in detecting very small lesions (<3 mm) and saccular aneurysms, which highlights the need for further refinement in handling diverse morphological variations. Future work could explore additional strategies for optimizing the model's

ISSN 2616-5775 Vol. 8, Issue 3: 56-63, DOI: 10.25236/AJCIS.2025.080308

efficiency and expanding its generalization capability to handle a broader range of vascular anomalies.

Overall, the proposed model offers a promising solution for the automated detection of aortic dilation in CT images, providing a foundation for further improvements in medical imaging applications.

References

[1] Jocher G, Stoken A, Borovec J, et al. ultralytics/yolov5: v3. 0[J]. Zenodo, 2020.
[2] Elpeltagy M, Sallam H. Automatic prediction of COVID− 19 from chest images using modified ResNet50[J]. Multimedia tools and applications, 2021, 80(17): 26451-26463.
[3] Chen J, Mai H S, Luo L, et al. Effective feature fusion network in BIFPN for small object detection[C]//2021 IEEE international conference on image processing (ICIP). IEEE, 2021: 699-703.