International Journal of Clinical and Experimental Medicine Research, 2025, 9(4), 460-473
https: //www.hillpublisher.com/journals/ijcemr/
ISSN Online: 2575-7970
ISSN Print: 2575-7989

DFEY-Net: Deep Feature-Enhanced YOLOv8 for Pulmonary Nodule Detection

Ji Tian, Ping Yang*

Smart City College, Beijing Union University, Beijing 100101, China.

How to cite this paper: Ji Tian, Ping Yang. (2025). DFEY-Net: Deep Feature-Enhanced YOLOv8 for Pulmonary Nodule Detection. *International Journal of Clinical and Experimental Medicine Research*, 9(4), 460-473.
DOI: 10.26855/ijcemr.2025.07.014

Received: June 12, 2025
Accepted: July 7, 2025
Published: August 4, 2025

***Corresponding author:** Ping Yang, Smart City College, Beijing Union University, Beijing 100101, China.

Abstract

Objective: To address the issue of insufficient integration and expression of multi-level features in pulmonary nodules by existing models, which affects the detection performance of small nodules, an improved pulmonary nodule detection algorithm based on YOLOv8 is proposed. **Methods:** In the backbone network of YOLOv8, the convolution downsampling operation with a stride of 2 is replaced with a Space-to-Depth downsampling operation, enabling more comprehensive extraction of subtle features to improve the detection of tiny nodules. At the end of the backbone network, a squeeze-and-excitation attention module and a global response normalization attention module are incorporated into the Spatial Pyramid Pooling Fast layer, with a fully connected layer established between them to enhance information interaction, thereby improving the model's ability to distinguish key features of pulmonary nodules. In the neck network, a Cross Stage Partial module is employed to integrate multi-level features extracted by the backbone network, achieving richer gradient combinations and reducing interference from redundant information. **Results:** The improved algorithm achieves precision, recall, and mean average precision of 96.1%, 96.3%, and 97.8%, respectively, on the LUNA16 dataset. Compared to the original YOLOv8 model, these metrics are improved by 3.4%, 2.0%, and 1.5%, respectively. **Conclusion:** The improved algorithm demonstrates enhancements across all evaluation metrics, effectively detecting small nodules while maintaining robust performance for larger pulmonary nodules.

Keywords

Pulmonary nodule detection; Convolutional neural network; Small nodule detection; Attention; Feature fusion

1. Introduction

According to the latest 2022 global cancer statistics data released in 2024 by the International Agency for Research on Cancer of the World Health Organization [1]. In 2022, lung cancer caused 1.81 million deaths, accounting for 18.7% of all cancer-related deaths, making it the leading cause of cancer mortality worldwide. Clinical research shows that early-stage lung cancer patients who receive treatment have a 5-year survival rate of approximately 70% to 90%, while for advanced-stage patients, this rate drops below 5%. Since early lung cancer often manifests as pulmonary nodules, early screening and diagnosis of these nodules are crucial for effective treatment. Chest CT scans provide high-resolution images of the lungs. A typical chest CT scan for a single patient consists of around 300 images, requiring healthcare professionals to carefully examine each one to identify nodules and assess their characteristics in order to determine the most appropriate treatment plan [2]. For healthcare professionals, diagnosing lung nodules requires strong expertise and involves high-intensity image interpretation with heavy workloads, making misdiagnosis a significant risk. Therefore, leveraging artificial intelligence to assist healthcare professionals in clinical diagnosis is of great importance.

Before deep learning was fully developed, researchers manually extracted nodule features before making classification predictions. This traditional lung nodule detection method relied heavily on researchers' subjective judgment, required complex parameter tuning, demonstrated poor adaptability across different scenarios, and had low automation levels, making it difficult to achieve satisfactory detection results. Deep learning methods, through training on large datasets, automatically learn features without manual feature engineering, thereby reducing human bias and demonstrating superior adaptability to varying data distributions and scenario changes. Consequently, deep learning has replaced traditional lung nodule detection approaches. Current deep learning-based lung nodule detection algorithms are primarily categorized into two-stage object detection algorithms and single-stage object detection algorithms.

For two-stage object detection algorithms, GUO N [3] et al., proposed a model combining Cascade R-CNN and Feature Pyramid Network (FPN) for multi-scale pulmonary nodule detection. In this model, Residual Network (ResNet) serves as the backbone for feature extraction. To enhance semantic features across layers, upper-layer features are fused with lower-layer features to construct the FPN. Additionally, Cascade R-CNN is incorporated, using different IoU thresholds for detection, achieving an Average Precision (AP) of 0.879. Harsono [4] et al. built upon Faster R-CNN by integrating pre-trained natural image weights with FPN, applying it to multi-scale 3D chest CT datasets. Cai L et al. [5] introduced a 3D visualization diagnostic detection and segmentation method for pulmonary nodules based on Mask R-CNN and a ray casting algorithm. Using ResNet50 as the backbone and FPN to fully integrate multi-scale feature maps, the Region Proposal Network generates candidate regions. The mask matrix is then multiplied with the original medical image sequence to obtain predicted nodule sequences, followed by ray casting to generate 3D nodule models. Two-stage algorithms first identify potential target regions through a Region Proposal Network, then perform classification and bounding box regression. While accurate, their complex architecture and large parameter count hinder practical deployment. In contrast, single-stage algorithms omit the region proposal step, offering faster speeds and lighter weight. The YOLO (You Only Look Once) series represents real-time single-stage detection, with extensive research and improvements by scholars. The YOLO (You Only Look Once) series algorithms represent single-stage real-time object detection methods, and numerous researchers have conducted extensive studies and improvements based on the YOLO framework. Li Xinzheng [6] et al. enhanced the YOLOv2 architecture by incorporating Inception V3 modules, thereby expanding both the width and depth of the network. By leveraging YOLOv2's inherent advantages in object detection, the modified network demonstrated enhanced feature extraction capabilities. Experimental results confirmed that this improved architecture not only streamlined the processing of pulmonary CT images but also increased both nodule detection rates and localization accuracy. Xu K [7] augmented YOLOv3's feature network with Inception ResBlocks to enable more comprehensive feature extraction. The implementation of the GDIoU loss function further improved bounding box regression accuracy, consequently enhancing overall pulmonary nodule detection performance. Building upon YOLOv6, Manokaran et al. [8] developed a pulmonary nodule detection algorithm specifically optimized for low-dose CT scans. Their approach involved initial training on high-dose CT data followed by transfer learning adaptation to low-dose CT images, ultimately achieving superior performance on low-dose CT datasets. Mammeri et al. [9] employed YOLOv7 for pulmonary nodule detection while systematically evaluating the impact of different input image configurations. Their findings indicated optimal performance when utilizing complete, uncropped images. Xiong et al. [10] introduced a novel detection network termed SAB-YOLO, implemented a weighted box fusion algorithm to integrate detection results from both YOLOv7 and SAB-YOLO, and ultimately designed a specialized classifier to effectively reduce false positive rates.

Fundamentally, two-stage region proposal algorithms demonstrate superior performance in small target detection. However, they incur significant computational overhead, requiring additional memory to store intermediate parameters, which results in slower processing speeds and higher energy consumption. In contrast, single-stage algorithms address these storage limitations through shared overlapping computations, achieving substantially faster detection speeds. With recent advancements in single-stage network architectures, these models now match the detection accuracy of region proposal methods for small targets while maintaining their inherent efficiency advantages. Notably, single-stage networks achieve real-time detection capabilities—a critical requirement for clinical applications. Consequently, when considering overall performance metrics, single-stage detection algorithms demonstrate clear superiority over their two-stage counterparts in pulmonary nodule detection tasks.

2. Methods

2.1 YOLOv8 algorithm

YOLOv8 is a representative algorithm in the YOLO series, addressing limitations of previous versions by combining rapid processing with high accuracy, making it particularly suitable for medical imaging applications. The algorithm primarily consists of four key components: input processing, backbone network, neck network, and head network. The first step involves image input utilizing Mosaic data augmentation for enhanced data preprocessing. The second step focuses on feature extraction through an advanced backbone network that progressively transforms input images into multi-level feature maps ranging from low-level features containing detailed spatial information to high-level features rich in semantic content. The third step implements feature fusion through a neck network employing a "Feature Pyramid Network + Path Aggregation Network " approach to effectively integrate the hierarchical features extracted by the backbone network. This optimized architecture enables YOLOv8 to maintain real-time detection speeds while achieving the precision required for medical diagnostics, especially valuable for clinical applications like pulmonary nodule analysis that demand both timeliness and accuracy [11]. In the fourth step, the fused multi-level features from the neck network are fed into the head network for detection to obtain the target size and location. Unlike previous versions that relied on anchor-based detection methods, YOLOv8 directly predicts object locations and categories without requiring predefined anchors. Additionally, YOLOv8 decouples the position prediction and category prediction during detection, making network training and inference more efficient. However, for pulmonary nodule detection tasks, YOLOv8 faces challenges including loss of subtle information during asymmetric downsampling operations and insufficient feature fusion in the neck network, leading to missed detection of small nodules. To address these issues, improvements have been made to the downsampling operations in YOLOv8's backbone network by incorporating attention modules, while more advanced feature fusion modules have been adopted in the neck network to enhance YOLOv8's detection capability for small pulmonary nodules.

2.2 The overall network architecture

YOLOv8 is renowned for its high precision and has found extensive applications across various domains, including medical imaging. Its architecture primarily consists of a backbone feature extraction network, a neck feature fusion network, and a detection head. Building upon YOLOv8, this chapter proposes three key improvements to address its limitations in pulmonary nodule detection tasks: First, in the backbone network, we replace stride-2 downsampling operations with Space-to-Depth [12] (SPD) downsampling to mitigate the loss of subtle features caused by asymmetric sampling (where different points are sampled unevenly) during feature extraction. Second, at the final SPPF layer of the backbone, we introduce a Spatial Pyramid Pooling Feature Enhance (SPPFE) module that combines Squeeze-and-Excitation (SE) attention and Global Response Normalization (GRN) attention. This enhancement strengthens the model's evaluation of channel-wise importance while establishing connections between the two attention mechanisms through fully connected layers [13], thereby improving information flow. The modification enhances the model's ability to aggregate critical nodule features from both spatial and channel dimensions. Finally, in the neck network, we substitute the original C2f module with a Cross Stage Partial (CSP) [14] module to fuse multi-level features extracted by the backbone. This change enables richer gradient combinations while reducing interference from redundant information.

The improved architecture is illustrated in Figure 1, where UP modules represent upsampling operations that reshape high-level features for fusion with other feature levels. The principles and structures of the SPD, SPPFE, and CSP modules are detailed in Sections 2.3, 2.4, and 2.5, respectively.

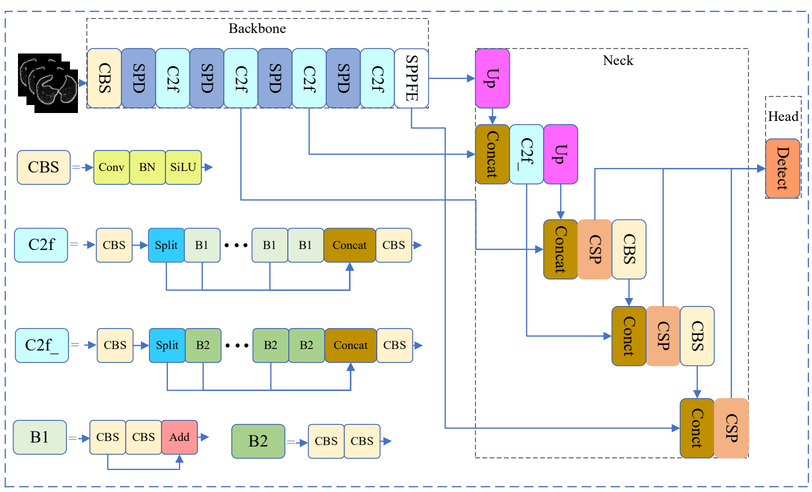


Figure 1. Overall structure of the DFEY-Net network.

2.3 Space-to-Depth downsampling operation

In the original YOLOv8, downsampling is performed using 3×3 convolutional kernels with a stride of 2. This operation constitutes asymmetric sampling, meaning different pixel points are sampled varying numbers of times. As shown in Figure 2, the circles represent pixels in the feature map to be downsampled, while the black squares indicate the coverage area of the 3×3 convolutional kernel during each stride-2 operation. The illustration reveals that green, yellow, and purple pixels are sampled 1, 2, and 4 times, respectively, during convolution, consequently influencing 1, 2, and 4 corresponding points in the subsequent feature map.

Considering small pulmonary nodules in detection tasks - typically quasi-spherical structures approximately 3mm in diameter (fewer than 10 pixels)—these may ultimately be represented by just a single pixel after multiple downsampling operations. If purple pixels in the feature map correspond to background information while green pixels contain nodule features, this sampling strategy excessively samples irrelevant background areas while neglecting nodule information, leading to missed detection of small nodules. Furthermore, when max pooling is employed for downsampling, it retains only the maximum pixel value within each kernel's coverage area while discarding other pixel information, resulting in data loss that similarly compromises model judgment.

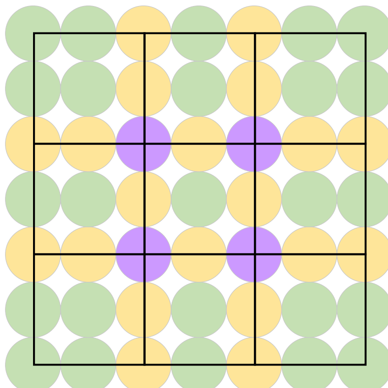


Figure 2. Convolution downsampling operation with a step size of 2.

This paper employs the SPD operation to reshape pixels from spatial to depth dimensions. The input tensor X is partitioned into four segments along the spatial dimensions (width and height), which are then concatenated along the channel dimension. This operation reduces the tensor's width and height by half while quadrupling the channel count. The processed tensor is subsequently fed into the C2f module shown in Figure 5, where the first convolutional layer restores the original channel dimensions. Essentially, this method calculates a weighted average of four pixel values from the original feature map to generate each new pixel, thereby avoiding asymmetric sampling while resolving the max pooling operation's limitation of disregarding non-maximum pixel values. The schematic diagram of the SPD operation is illustrated in Figure 3.

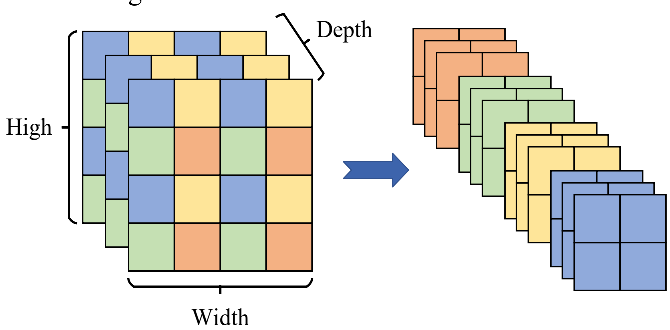


Figure 3. SPD operation.

2.4 Spatial Pyramid Pooling Feature Enhancement

In the original YOLOv8 architecture, the SPPF module connects the backbone network and the neck feature fusion network. The feature maps input to SPPF first pass through a CBS module, which reduces channel dimensions to decrease computational load while extracting preliminary features. Subsequently, the output from the CBS operation is concatenated with outputs that have undergone both CBS processing and varying numbers of pooling operations. Finally, the concatenated features are fused through another CBS module to produce the final output. The SPPF structure is illustrated in Figure 4.

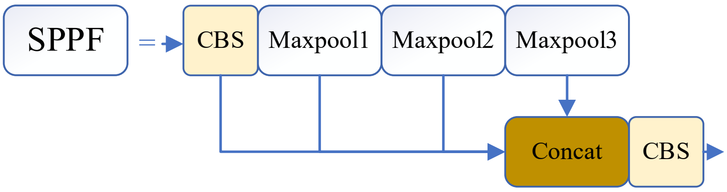


Figure 4. SPPF module.

We note that in the SPPF module, two max pooling layers with 5×5 kernels, stride 1, and padding 2 connected in series are equivalent to a single max pooling layer with a 9×9 kernel. The schematic diagram of two max pooling operations in series is shown in Figure 5, where green dots represent pixels in the feature map. The red box in the figure indicates the coverage area of one max pooling operation in Maxpool1 in Figure 4, with the pixel value at point 1 being the maximum value within the red box area. The blue box in Figure 5 represents the coverage area of one pooling operation in Maxpool2. When the red box slides to where its center point lies within the blue box area, the coverage area of the red box becomes identical to the black box. Since each point in the blue box represents the maximum pixel value within a 5×5 box centered at that point, point 2 is simultaneously the maximum pixel value within both the blue box area and the black box area. The black box represents the coverage area of a max pooling operation with a 9×9 kernel, meaning that two pooling operations with 5×5 kernels accomplish the equivalent of one max pooling operation with a 9×9 kernel. Similarly, three max pooling layers with 5×5 kernels, stride 1, and padding 2 connected in series are equivalent to a max pooling layer with a 13×13 kernel.

As the final layer of the feature extraction network, the SPPF layer enhances important features at different scales at the spatial level, effectively improving the network's learning capability. Building on this, we leverage the advantages of channel attention to further summarize features along the channel dimension, increasing the network's attention to important information across different channels, thereby further enhancing the network's feature extraction capability. Therefore, this paper integrates the CA (Channel Attention) module into the SPPF module to obtain the SPPFE module, aiming to improve the network's detection performance. The structure of SPPFE is shown in Figure 6, and the operation performed by GRN in Figure 6 is shown in Equation (1).

$$X_i = \gamma X_i N(G(X)_i) + \beta + X_i \tag{1}$$

In the equation, Xi represents each feature value in the feature map, $G(\cdot)$ denotes a global feature aggregation function using the L2 norm, $N(\cdot)$ represents a feature normalization function, γ is a learnable weight parameter, and β is a learnable bias parameter.

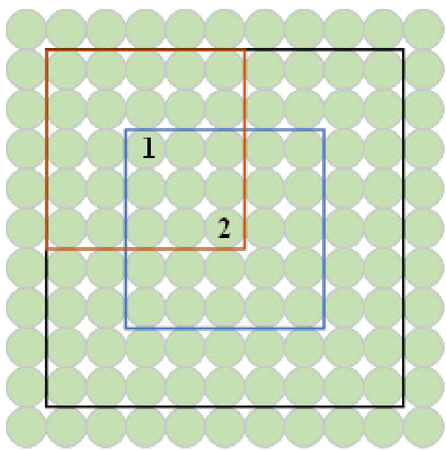


Figure 5. Two Max pooling series operations.

The SPPFE module is obtained by adding a CA module before the second CBS module in the SPPF module. The CA module uses a fully connected layer to expand the number of channels in the feature map processed by SE attention to four times the original and connects all neurons with global response normalization, enabling sufficient weight adjustment and output normalization. Subsequently, a fully connected layer is used to restore the channel count and assign a weight before performing a residual connection with the input to obtain the final output.

The key of the CA module lies in connecting all neurons from SE attention and global response normalization to combine the advantages of both, allowing the model to accurately focus on critical information and exclude interference from irrelevant information.

The SE attention module consists of three main steps: squeeze, excitation, and feature calibration. First, global average pooling is used to "squeeze" the pixels of each channel in the input features to obtain the global features of each channel. Then, fully connected layers and activation functions (ReLU, Sigmoid) are used to "excite" the global features, capturing inter-channel dependencies and generating channel weights. Finally, the input features are "calibrated" by element-wise multiplication with the channel weights to produce the final output. This mechanism enables the network to focus more on task-relevant channels while ignoring unimportant information, reducing interference from redundant features and thereby enhancing feature representation.

Global response normalization, as shown in Equation (1), mainly consists of three parts: global feature aggregation, feature normalization, and feature calibration. First, global response normalization uses the L2 norm to "aggregate" all features within each channel, obtaining the weight for each channel. Then, "feature normalization" is applied to the channel weights to constrain their values between 0 and 1. Finally, the weights obtained in the second step are multiplied by the input, scaled by a parameter γ, and added to the input through a residual connection, followed by the addition of a bias term β to complete the "feature calibration" of the input. This method adjusts the feature responses across different channels, helping to improve channel contrast and selectivity.

In summary, the squeeze-and-excitation attention mechanism extracts channel importance through global average pooling, while global response normalization calculates channel importance using the L2 norm. These two approaches extract key channel information from different perspectives and are complementary. By using fully connected layers to link all neurons of these two attention mechanisms in the CA module, the two attention modules can fully interact, combining their respective advantages to enhance the network's ability to distinguish critical information and thereby improve detection performance.

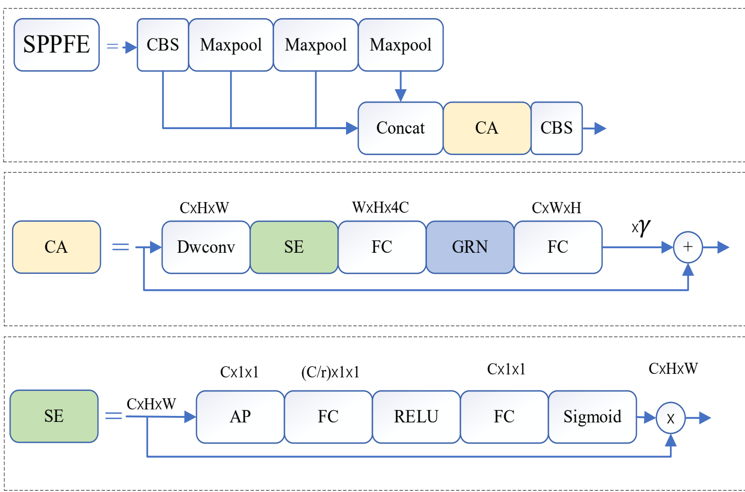


Figure 6. Schematic diagram of the SPPFE module structure.

2.5 Cross-Stage Partial Model

The backbone network of YOLOv8 extracts features from different hierarchical levels and transmits these features to the neck network for multi-scale fusion, thereby enhancing the model's ability to discriminate complex targets. In the original YOLOv8, features from different levels are first concatenated along the channel dimension and then fused through the C2f module. The specific structure of the C2f module is shown in Figure 7: the input features are processed by the CBS module and then evenly split into two parts along the channel dimension. The first part is processed through n serially connected B2 modules, each containing two output paths: the first path transmits features to the next B2 module, while the second path is concatenated with the second paths of other B2 modules and the second part of the input features. Finally, the concatenated features are fed into the CBS module at the end of the C2f module to adjust the channel count and generate the output.

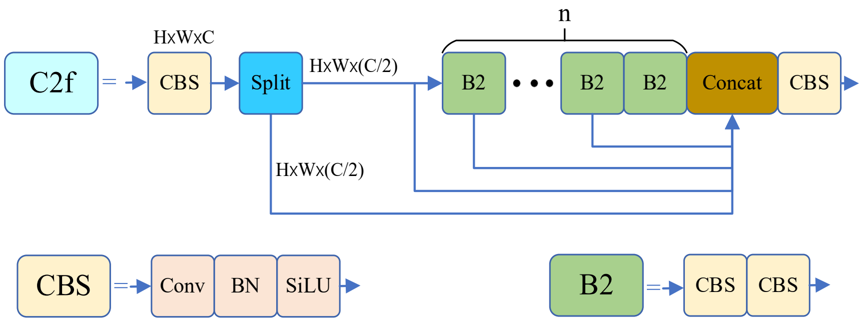


Figure 7. C2f module.

The C2f module achieves feature reuse through parallel connections of all B2 module output paths, but this design also introduces issues of feature redundancy and gradient dispersion [15, 16]. Specifically, in deep networks, since the output of each B2 module is directly concatenated at the module's end, feature information gets repeatedly reused. This not only increases computational complexity but may also introduce redundant information, interfering with the model's extraction of key features. Additionally, during backpropagation, gradients disperse across multiple paths, reducing training efficiency and weakening the model's perception of subtle features. These issues are particularly prominent in small target detection tasks like pulmonary nodule detection, where tiny nodules with indistinct features can easily be overwhelmed by redundant information. Consequently, this design of the C2f module limits the model's ability to capture fine features to some extent, affecting detection performance.

To address this, improvements were made to YOLOv8's neck network by replacing the original C2f module with a CSP module. The structure of the CSP module is shown in Figure 8, designed to optimize the feature fusion process and enhance the model's perception of subtle features. Specifically, input features entering the CSP module are split into two independent branches: the first branch passes through a CBS module and is directly transmitted to the module's end; the second branch connects two residual modules composed of Rep and CBS in series, with an additional direct path to the module's end introduced between the two residual modules. This multi-branch residual connection significantly enhances information interaction and transmission between feature layers, preventing the loss of subtle feature information during fusion, thereby improving the model's detection capability for small pulmonary nodules.

Furthermore, the Rep module in the CSP module adopts a reparameterization design. During training, the Rep module fuses rich feature information through a dual-branch structure; during inference, the dual branches are merged into a single path via structural reparameterization technology, significantly reducing computational load and memory usage, thereby lowering inference time. This design not only ensures effective feature extraction during training but also improves inference efficiency, making the model more practical in real-world applications.

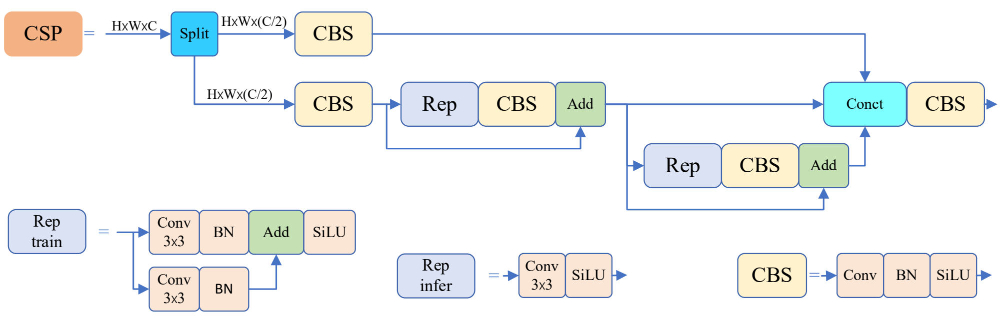


Figure 8. CSP module.

3. Results

3.1 Data, Metrics, and Parameters

3.1.1 Lung CT image dataset

The experiment utilized the LUNA16 dataset [17], which was established by the National Institutes of Health for evaluating the performance of automated pulmonary nodule detection algorithms in a challenge competition. The organizers initiated this challenge with the aim of promoting the development of computer-aided diagnosis systems for pulmonary nodules, thereby assisting physicians in more accurate nodule diagnosis and reducing their diagnostic workload.

The LUNA16 dataset contains lung CT scan data from 888 patients, involving 1,186 pulmonary nodules, totaling 227,225 CT images along with corresponding nodule annotation information. The original dataset consists of files in MHD and raw formats. The MHD files store CT image dimensions, CT value ranges, and nodule annotation data, while the raw files contain the actual CT images. Prior to training, the CT images were converted into standard PNG-format images, and text-format label files were generated based on the annotation information.

3.1.2 Data Preprocessing

This paper uses traditional methods to perform lung parenchyma segmentation on the sliced LUNA16 dataset, which can eliminate the influence of background elements in CT images, such as fat, air, and non-lung regions. The main steps are as follows:

(1) Normalize the image by standardizing pixel values to have a mean of 0 and standard deviation of 1, facilitating subsequent processing;

(2) Extract the central region of the image, as lungs are typically located in the central area, which can reduce interference from non-tissue components (such as background and scanning table);

(3) Apply K-means clustering to the central region to obtain foreground and background thresholds, dividing the image into foreground (radiolucent tissues like lungs) and background (radiopaque tissues like bones and soft tissues), with foreground pixels assigned a value of 1 and background pixels assigned 0;

(4) Perform morphological operations (erosion and dilation) to remove noise and fill cavities, making the lung region more complete;

(5) Generate the final lung mask;

(6) Complete lung parenchyma segmentation using the mask.

Figure 9 shows the processing results at each stage of lung parenchyma segmentation.

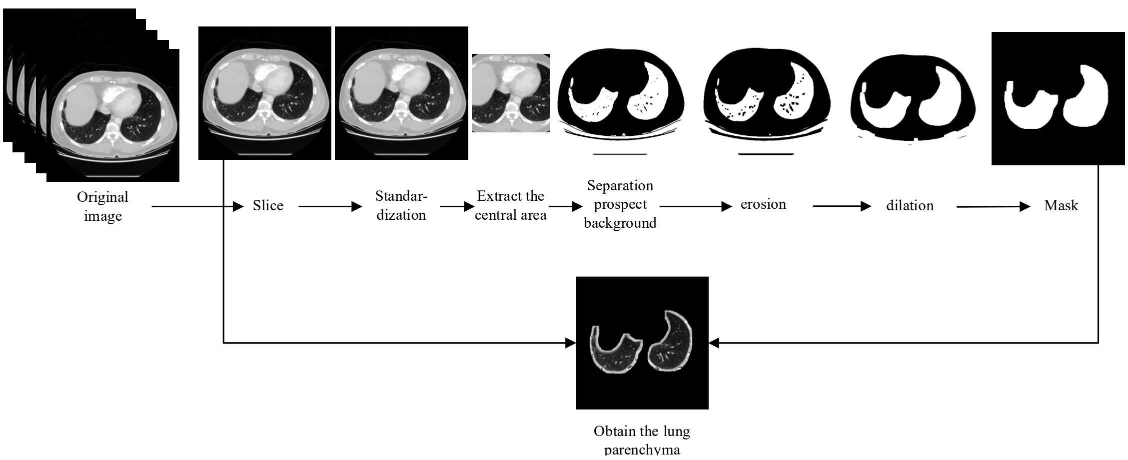


Figure 9. Treatment effect of each stage of lung parenchymal segmentation.

To enhance the model's generalization capability, the MedAugment [18] data augmentation method was employed to expand the lung parenchyma-segmented images prior to training, followed by dataset partitioning and training-validation using ten-fold cross-validation. MedAugment systematically filters out augmentation operations from natural image processing that are unsuitable for medical image analysis (e.g., inversion, equalization), then categorizes the augmentation operations into pixel-level and spatial-level transformations. Two augmentation spaces, denoted as Ap and As, were constructed, with their specific operations detailed in Table 1.

Table 1. Specific Operations of Ap and As

Ap	Brightness	Contrast	Posterization	Sharpness	Gaussian Blur	Gaussian Noise
As	Rotation	Horizontal Flip	Vertical Flip	Scaling	Horizontal Translation	Vertical Translation

Due to the high sensitivity of medical images to attributes such as brightness, and considering that consecutive operations in Ap may damage medical image information, a novel sampling strategy was adopted to select operations from Ap and As. Taking single-image augmentation as an example, the specific procedure is as follows:

(1) Randomly select 1 or 0 method from Ap;

(2) Randomly select 2 or 3 methods from As;

(3) Combine the methods randomly selected in steps (1) and (2) to process the image, completing one augmentation cycle.

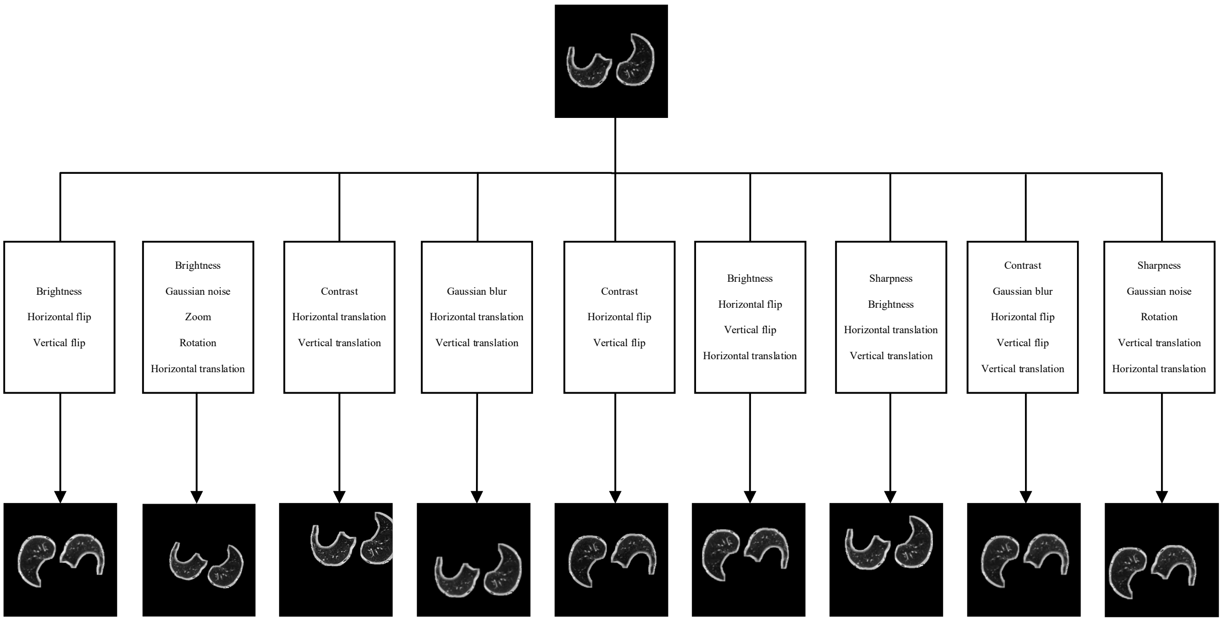


Figure 10. Demonstration of the data expansion effect of Medaugment.

3.1.3 Evaluation Metric

This paper adopts common object detection evaluation metrics: Precision, Recall, and Average Precision (AP). Precision represents the percentage of correctly detected positive samples among all detected positive samples. Recall refers to the percentage of correctly detected positive samples among all actual positive samples. AP is the area under the Precision-Recall curve plotted with precision as the vertical axis and recall as the horizontal axis at different confidence thresholds. The mathematical expressions for Precision, Recall, and AP are shown in Equations (2), (3), and (4), respectively:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{2}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{3}$$

$$\mathrm{AP} = \int_0^1 \mathrm{P(R)dR} \tag{4}$$

Among them, TP is the number of actual positive samples that are correctly identified as positive, FP is the number of actual negative samples that are incorrectly identified as positive, FN is the number of actual positive samples that are incorrectly identified as negative. In this paper, pulmonary nodules are considered positive samples, and the background is considered negative samples. R represents the recall rate as the independent variable, and P(R) represents the precision corresponding to the recall rate at the current confidence level.

3.1.4 Experimental configuration and initial parameters

Table 2. Training Parameter Settings

Training parameter name	Parameter value
Initial learning rate	0.01
Cosine Annealing Hyperparameters	0.01
weight decay coefficient	0.0005

Table 2 Continued

Training parameter name	Parameter value
Batch size	16
Training batch	300
Image input size	640×640

The implementation was developed using Python 3.9 and PyTorch 1.12 framework, with network training and testing conducted on a system equipped with 32 GB RAM, NVIDIA GeForce RTX 3090Ti GPU, Intel Core i6-12600KF 3.70 GHz CPU processor, and Windows 10 operating system. Other detailed parameter configurations are specified in Table 2.

3.2 Experimental Results and Analysis

3.2.1 Experiments on the effectiveness of various improvement strategies

To investigate the impact of each improvement on algorithm performance, this paper conducted ablation experiments for various modifications, with the experimental results presented in Table 3. Table 3 compares the effects of using SPD operations for downsampling in the YOLOv8 backbone network, replacing the SPPF module at the end of the backbone network with the SPPFE module, and employing the CSP module for feature fusion in the neck network.

Table 3. Impact of SPD, AFPN, and GCAA on Model Performance (%)

Network	Precision	Recall	AP
YOLOv8	92.7	94.3	96.3
YOLOv8+SPD	93.7	95.0	97.1
YOLOv8+SPD+CSP	95.0	95.4	97.4
YOLOv8+SPD+SPPFE+CSP	96.1	96.3	97.8

As shown in Table 3, when replacing the stride-2 convolutional downsampling operation in YOLOv8 with the SPD operation, all metrics improved because the partial information lost in the stride-2 convolution was preserved. Using the CSP module to replace the C2f module in the YOLOv8 neck network enabled more thorough fusion of features across different levels, leading to increases in precision, recall, and AP. Finally, after integrating both squeeze-and-excitation attention and global feature response attention into the SPPF module to form the SPPFE module, precision, recall, and AP increased by 1.1%, 0.9%, and 0.4%, respectively. This demonstrates that connecting squeeze-and-excitation attention with global response normalization via fully connected layers effectively combines their advantages, successfully integrates critical information, and positively contributes to the algorithm's improvement.

3.2.2 Comparative experiments of different detection methods

Table 4. Performance Comparison Between DFEY-Net and Other Detection Models

Network	Precision	Recall	AP
FasterRCNN [19]	73.2	72.1	71.0
SSD [20]	67.3	64.8	65.2
YOLOv3	93.6	93.0	94.9
YOLOv4 [21]	68.5	54.9	57.3
YOLOv5m	82	84.8	85.2
YOLOv7 [22]	71.2	69.4	68.2
YOLOv8m	92.7	93.8	95.5

Table 4 Continued

Network	Precision	Recall	AP
YOLOv9m [23]	89.7	91.0	94.5
YOLOv10m [24]	94.8	93.2	97.3
YOLO11m	93.3	93.9	96.5
Ref. [25]	93.3	94.1	95.9
Ref. [26]	-	96.7	95.9
Ref. [27]	-	-	73.0
Ref. [28]	-	94.4	95.39
Ref. [29]	95.41	94.02	95.26
Ref. [30]	92.3	88.5	-
Ref. [31]	90.8	89.8	92.7
DFEY-Net	96.1	96.3	97.8

Table 4 presents the performance comparison between the YOLOv8-based pulmonary nodule algorithm and the aforementioned models. The results indicate that YOLOv3, YOLOv8m, YOLOv9m, YOLOv10, references [25-28], and DFEY-Net demonstrate superior performance in pulmonary nodule detection tasks, with precision, recall, and AP metrics consistently exceeding 90%. In contrast, YOLOv4, YOLOv7, Faster RCNN, SSD, and reference [27] show relatively mediocre performance for pulmonary nodule detection, with all metrics around 70%, making them inadequate for this diagnostic task. Compared with other detection models, DFEY-Net achieves the highest precision and AP values. This superior performance primarily benefits from two factors: firstly, DFEY-Net builds upon the already robust YOLOv8 framework as its foundation; secondly, each implemented improvement effectively addresses specific limitations of YOLOv8, collectively contributing to the enhanced performance.

Table 5. Size Comparison Between DFEY-Net and Other Detection Models

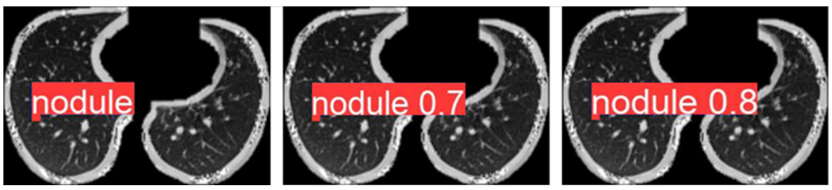
Network	Parameter /M	Gflops	Weight file size /MB
FasterRCNN	78	159	163
SSD	47	100.2	102
YOLOv3	62	155.3	117.8
YOLOv4	52	119.7	137.6
YOLOv7	37	105.1	72.1
YOLOv8m	26	78.9	49.6
YOLOv9m	26	76.3	38.8
YOLOv10m	15	59.1	63.8
YOLO11m	20.1	68.0	38.8
Ref. [25]	64.152	127.5	-
Ref. [26]	43.6	-	-
Ref. [27]	160.5	-	-
DFEY-Net	34	81.7	68.8

Table 5 compares the model complexity of different pulmonary nodule detection algorithms in terms of parameter count, floating-point operations (FLOPs), and weight file size. As shown in Table 5, DFEY-Net exhibits increases in parameter count, FLOPs, and weight file size compared to YOLOv8. This is attributed to the incorporation of both squeeze-and-excitation attention modules and global response normalization modules connected via fully connected layers into the original YOLOv8's SPPF module, which enhances model performance while also increasing complexity. Overall, DFEY-Net's complexity ranks mid-level among the listed models—while more complex than YOLOv9 and YOLOv10, it remains relatively lightweight compared to references [25-27].

When considering both model performance and complexity, DFEY-Net achieves the best performance, with precision and AP values surpassing existing algorithms. However, there remains room to reduce DFEY-Net's model complexity. Future work will focus on developing lung nodule detection algorithms that balance high accuracy and lightweight design.

3.2.3 Detection Performance Comparison

Detection results can more intuitively demonstrate the differences between algorithms. Figures 11 to 17 show the detection performance of YOLOv8 and DFEY-Net for pulmonary nodules of different sizes. Figures 11, 12, 13, and 14 display detection results for small nodules (3-5 mm), while Figures 15 and 16 show medium-sized nodules (5-10 mm), and Figure 17 presents results for large nodules (10-30 mm). Each figure contains three images: the actual nodule, YOLOv8's predicted nodule, and DFEY-Net's predicted nodule. The results reveal that in Figures 11, 12, and 13, YOLOv8 exhibits lower confidence scores in nodule detection compared to DFEY-Net. Figure 14 demonstrates that YOLOv8 missed detecting a 3-4 mm small nodule, while DFEY-Net successfully detected it with higher confidence. Additionally, Figure 15 shows that YOLOv8 incorrectly identified the background as a nodule, whereas DFEY-Net did not make this error. In summary, DFEY-Net demonstrates more accurate detection of pulmonary nodules across all size ranges and achieves higher confidence scores compared to the original YOLOv8, making it more valuable for practical applications.

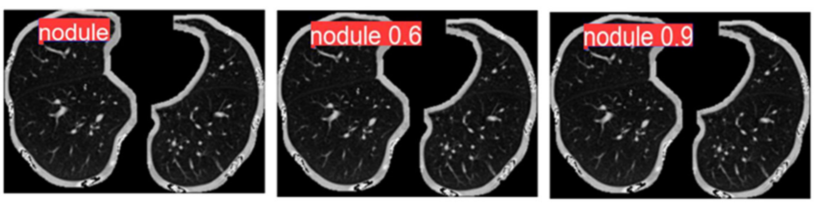


(a) Lable (b) YOLOv8 (c) DFOY-Net

Figure 11. 3-5mm Pulmonary Nodule Case 1

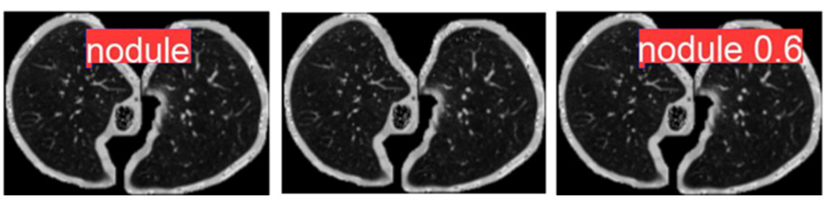
(a) Lable (b) YOLOv8 (c) DFOY-Net

Figure 12. 3-5mm Pulmonary Nodule Case 2



(a) Lable (b) YOLOv8 (c) DFOY-Net

Figure 13. 3-5mm Pulmonary Nodule Case 3.



(a) Lable (b) YOLOv8 (c) DFOY-Net

Figure 14. 3-5mm Pulmonary Nodule Case 4.

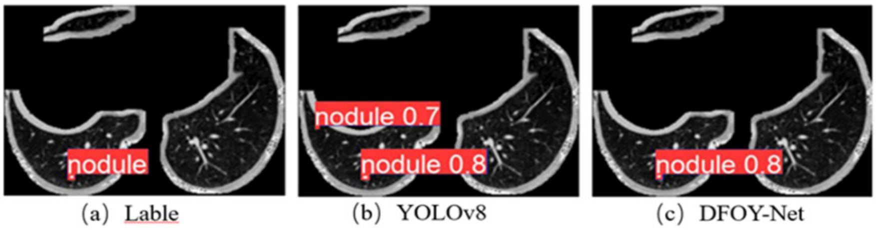


Figure 15. 5-10mm Pulmonary Nodule Case 1.

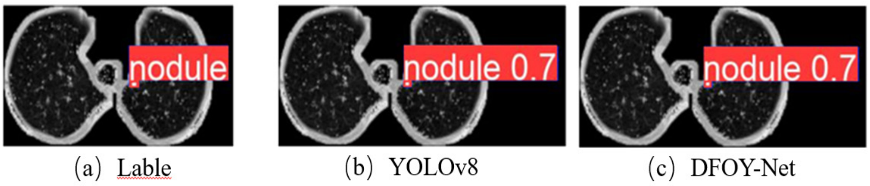


Figure 16. 5-10mm Pulmonary Nodule Case 2.

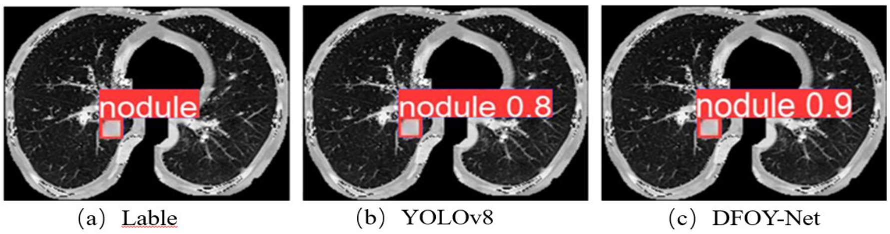


Figure 17. 10-30mm Pulmonary Nodule.

4. Discussion

This study addresses the insufficient multi-level feature integration and representation capability of existing models in pulmonary nodule detection, particularly for small nodules, by enhancing the YOLOv8 model. The improvements include: (1) Replacing stride-2 convolutional downsampling with Space-to-Depth downsampling to better preserve subtle features; (2) Introducing a dual-attention mechanism at the backbone network's terminal with fully connected layers to strengthen feature discrimination; (3) A more advanced feature fusion module is adopted in the neck network for richer gradient combinations. Experimental results demonstrate that the enhanced algorithm maintains detection performance for large nodules while significantly improving small nodule detection accuracy, with all evaluation metrics surpassing the original model, showing strong clinical applicability.

References

[1] Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-63.

[2] Al-Shabi M, Lee HK, Tan M. Gated-dilated networks for lung nodule classification in CT scans. IEEE Access. 2019;7:178827-38.

[3] Guo N, Bai Z. Multi-scale pulmonary nodule detection by fusion of cascade R-CNN and FPN. In: 2021 International Conference on Computer Communication and Artificial Intelligence (CCAI). IEEE; 2021:16-9.

[4] Harsono IW, Liawatimena S, Cenggoro TW. Lung nodule detection and classification from Thorax CT-scan using RetinaNet with transfer learning. J King Saud Univ Comput Inf Sci. 2022;34(3):567-77.

[5] Cai L, Long T, Dai Y, et al. Mask R-CNN-based detection and segmentation for pulmonary nodule 3D visualization diagnosis. IEEE Access. 2020;8:44400-9.

[6] Li X, Jin W, Li G, et al. Asymmetric Convolutional Kernel YOLO V2 Network for Pulmonary Nodule Detection in CT Images. Chin J Biomed Eng. 2019;38(4):401-8.

[7] Xu K, Jiang H, Tang W. A New Object Detection Algorithm Based on YOLOv3 for Lung Nodules. In: Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence. 2020:233-9.

[8] Manokaran J, Mittal R, Ukwatta E. Pulmonary nodule detection in low dose computed tomography using a medical-to-medical transfer learning approach. J Med Imaging. 2024;11(4):044502.

[9] Mammeri S, Amroune M, Haouam MY, et al. Early detection and diagnosis of lung cancer using YOLO v7, and transfer learning. Multimed Tools Appl. 2024;83(10):30966-80.

[10] Xiong Y, Deng L, Wang Y. Pulmonary nodule detection based on model fusion and adaptive false positive reduction. Expert Syst Appl. 2024;238:121890.

[11] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:8759-68.

[12] Sunkara R, Luo T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer; 2022:443-59.

[13] Ding X, Zhang Y, Ge Y, et al. UniRepLKNet: A Universal Perception Large-Kernel ConvNet for Audio Video Point Cloud Time-Series and Image Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024:5513-24.

[14] Zhao G, Ge W, Yu Y. GraphFPN: Graph feature pyramid network for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:2763-72.

[15] Zhang K, Li Z, Zhang F, et al. Pan-sharpening based on transformer with redundancy reduction. IEEE Geosci Remote Sens Lett. 2022;19:1-5.

[16] Hidayatullah P, Syakrani N, Sholahuddin MR, et al. YOLOv8 to YOLO11: A Comprehensive Architecture In-depth Comparative Review. arXiv. 2025;2501.13400.

[17] Setio AAA, Traverso A, De Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med Image Anal. 2017;42:1-13.

[18] Liu Z, Lv Q, Li Y, et al. Medaugment: universal automatic data augmentation plug-in for medical image analysis. arXiv. 2023;2306.17466.

[19] Girshick R. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. 2015:1440-8.

[20] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer; 2016:21-37.

[21] Bochkovskiy A, Wang CY, Liao HYM. Yolov4: Optimal speed and accuracy of object detection. arXiv. 2020;2004.10934.

[22] Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:7464-75.

[23] Wang CY, Yeh IH, Liao HYM. Yolov9: Learning what you want to learn using programmable gradient information. In: European Conference on Computer Vision. Cham: Springer; 2025:1-21.

[24] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection. arXiv. 2024;2405.14458.

[25] Chen M, Li Y, Yun J, et al. Pulmonary nodule detection algorithm based on KCCS-YOLOv4. J Changchun Univ. 2023;44(5):424-33. doi:10.15923/j.cnki.cn22-1382/t.2023.5.07.

[26] Liu K. Stbi-yolo: A real-time object detection method for lung nodule recognition. IEEE Access. 2022;10:75386-94.

[27] Ji Z, Wu Y, Zeng X, et al. Lung nodule detection in medical images based on improved YOLOv5s. IEEE Access. 2023;11:76371-87.

[28] Liu Y, Ao Y. Deformable attention mechanism-based YOLOv7 structure for lung nodule detection. J Supercomput. 2024;80(17):25450-69.

[29] Wu X, Zhang H, Sun J, et al. YOLO-MSRF for lung nodule detection. Biomed Signal Process Control. 2024;94:106318.

[30] Şaman C, Çelikbaş Ş. YOLOv8-based lung nodule detection: a novel hybrid deep learning model proposal. Int Res J Eng Technol. 2023;10(8):230-7.

[31] Wang M, Yang Z, Zhao R, et al. CPLOYO: A Pulmonary Nodule Detection Model with Multi-Scale Feature Fusion and Nonlinear Feature Learning. arXiv. 2025;2503.10045.