Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.34002

Comprehensive Evaluation Method for Medical Image Object Detection Models Based on Risk-Constrained Weight Optimization

Feifan Li

Faculty of Medicine, Dalian University of Technology, Dalian, China
3227781293@qq.com

Abstract. Comprehensive evaluation of medical image object detection models is of great significance for clinical application selection. Existing evaluation methods mainly focus on a single accuracy metric, making it difficult to comprehensively reflect the overall performance of models in practical deployment environments. This paper constructs an evaluation system covering three dimensions, namely usability, efficiency, and robustness, with a total of 12 indicators, and proposes a risk-constrained subjective-objective weight optimization fusion method. The proposed method integrates the subjective weights obtained by the analytic hierarchy process, the objective weights obtained by the entropy weight method, and the risk prior weights for medical scenarios. It adaptively determines the reliability coefficients of the three types of weights, thereby obtaining comprehensive evaluation weights that consider expert experience, data variation, and clinical risk. Experimental results on five models, including YOLOv8n, Faster R-CNN, CenterNet, YOLOv6s, and Deformable DETR, show that the ranking results of the proposed method are highly consistent with the actual performance of the models. The method demonstrates good rationality and interpretability, and can provide a quantitative decision-making basis for the selection and optimization of medical image detection models.

Keywords: medical image, object detection, comprehensive performance evaluation, analytic hierarchy process and entropy weight method, risk-constrained optimization

1. Introduction

In recent years, deep learning-driven object detection techniques have made considerable progress in the field of medical image analysis [1, 2]. Various high-performance detection models have emerged, represented by the YOLO series [3-5] and real-time Transformer-based detectors [6, 7]. These models have continuously improved accuracy and speed records on general object detection benchmarks. However, in the high-risk scenario of medical imaging, model deployment still faces complex challenges such as differences in imaging quality, device heterogeneity, and clinical deployment constraints [8, 9]. Different models have distinct advantages in detection accuracy, inference speed, and stability in complex scenarios. Therefore, how to conduct an objective and comprehensive evaluation of these models is a key issue in promoting their practical application.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.34002

Existing studies mostly use a single metric such as mAP to measure model performance, while ignoring the dimensions of efficiency and robustness. In addition, they do not consider the differentiated risk requirements of medical scenarios for different performance indicators [8, 10]. Multi-criteria comprehensive evaluation methods, such as TOPSIS [11], the analytic hierarchy process [12], and VIKOR [13], have been widely applied in engineering evaluation. However, when directly applied to medical image model evaluation, systematic modeling of subjective-objective weight fusion and risk constraints remains insufficient [14].

To address this issue, this paper constructs a multi-level evaluation indicator system for medical image object detection and proposes a risk-constrained optimal fusion method for subjective and objective weights. The aim is to provide a reliable quantitative tool for the comprehensive evaluation of medical detection models.

2. Related work

Evaluation of object detection models.

Object detection tasks usually use mAP as the core evaluation metric [3]. Some studies further introduce efficiency indicators such as inference latency and the number of model parameters to construct comprehensive benchmarks [4, 7, 15]. With the development of lightweight end-to-end detection frameworks [5], the trade-off between efficiency and accuracy for edge deployment has received increasing attention [16]. In terms of medical image evaluation, Maier-Hein et al. [9] systematically reviewed the criteria for metric selection in image analysis validation and pointed out that a single metric is insufficient to comprehensively reflect the clinical applicability of a model. However, existing benchmarks rarely design robustness evaluation dimensions for the special imaging conditions of medical images, such as overexposure, blur, and occlusion [10, 17]. Therefore, the clinical applicability of evaluation conclusions still needs to be improved.

Multi-criteria comprehensive evaluation methods.

The analytic hierarchy process (AHP) quantifies experts' subjective preferences by constructing judgment matrices and has been widely used in multi-criteria decision-making [12]. The entropy weight method determines objective weights according to the amount of information contained in indicator data [18]. The VIKOR method introduces group utility and individual regret indicators to achieve compromise ranking under conflicting criteria, and has been widely studied in engineering and medical decision-making [13]. In recent years, several studies have attempted to integrate subjective and objective weights to improve evaluation robustness [11, 14]. However, explicit modeling of medical risk priors and efficiency constraints for clinical deployment remain relatively limited [16]. Based on the above studies, this paper introduces risk prior constraints and adaptively determines fusion coefficients to meet the specific requirements of medical image evaluation.

3. Evaluation method

3.1. Evaluation indicator system

For the comprehensive evaluation of medical image object detection models, this paper constructs an evaluation system consisting of three first-level indicators: usability, efficiency, and robustness. Usability mainly reflects the detection accuracy of the model, efficiency mainly reflects the inference speed and computational resource consumption of the model, and robustness mainly reflects the stability of the model under complex imaging conditions. The specific indicator system is shown in Table 1.

DOI: 10.54254/2755-2721/2026.34002

Table 1. Evaluation indicator system

Category	Metric	Definition	Dir.
Usability	mAP	Mean average precision over all categories	↑
	AP@0.5	Average precision at IoU threshold 0.5	↑
	AP@0.75	Average precision at IoU threshold 0.75	↑
Efficiency	FPS	Number of processed frames per second	↑
	FLOPs	Floating-point operations for inference	↓
	GPU Memory	GPU memory usage during inference	↓
Robustness	Overexposure mAP	mAP under overexposure conditions	↑
	Underexposure mAP	mAP under underexposure conditions	↑
	Occlusion mAP	mAP under target occlusion conditions	↑
	Gaussian Noise mAP	mAP after adding Gaussian noise	↑
	Gaussian Blur mAP	mAP after adding Gaussian blur	↑
	Contrast mAP	mAP under contrast variation conditions	↑

3.2. Indicator quantification and score calculation

Since different indicators have different dimensions and value ranges, this paper first transforms the original indicators into unified quantitative scores. For benefit-type indicators, a larger indicator value indicates better model performance. For cost-type indicators, a smaller indicator value indicates better model performance. Let x_{ij} denote the quantitative score of the i-th model on the j-th indicator. The comprehensive evaluation score of the model can be expressed as

$$S_i = \sum_{j=1}^{n} w_j x_{ij} \tag{1}$$

where w_j denotes the fused weight of the j-th indicator and satisfies

$$\sum_{j=1}^{n} w_j = 1, \quad w_j \geq 0 \tag{2}$$

3.3. Subjective weight calculation

This paper adopts the analytic hierarchy process to calculate subjective weights. A judgment matrix A is constructed according to the relative importance among indicators, and the subjective weight vector w_s is obtained using the eigenvalue method. The consistency of the judgment matrix is measured by the consistency ratio CR:

DOI: 10.54254/2755-2721/2026.34002

$$CR = \frac{CI}{RI} \tag{3}$$

where

$$CI = \frac{\lambda_{\max} - n}{n - 1} \tag{4}$$

A smaller CR indicates better consistency in expert judgments and higher credibility of the subjective weights.

3.4. Objective weight calculation

To reflect the variability of the indicator data itself, this paper adopts the entropy weight method to calculate objective weights. Let p_{ij} denote the normalized proportion of the i-th model on the j-th indicator. The information entropy of the j-th indicator is given by

$$e_j = -k \sum_{i=1}^{m} p_{ij} \ln p_{ij} \tag{5}$$

where k = 1/ln m, and m is the number of models involved in the evaluation. The objective weight is further obtained as

$$w_{o,j} = \frac{1 - e_j}{\sum_{j=1}^{n} (1 - e_j)} \tag{6}$$

Thus, the objective weight vector w_o is obtained.

3.5. Modeling of risk prior weights

The evaluation of medical image models should consider not only statistical performance but also the impact of different indicators on practical application risks. For example, detection accuracy and robustness in complex scenarios directly affect diagnostic reliability, while FPS, FLOPs, and GPU memory usage mainly reflect deployment efficiency. Therefore, this paper introduces the risk prior weight w_r to describe the risk sensitivity of different indicators:

$$w_r = \text{Normalize}\,(r) \tag{7}$$

where r denotes the indicator risk score vector.

DOI: 10.54254/2755-2721/2026.34002

3.6. Risk-constrained optimal fusion of subjective and objective weights

To integrate subjective experience, objective data, and risk information, this paper constructs a risk-constrained weight optimization model:

$$\min_{w} a \parallel w - w_s \parallel_2^2 + b \parallel w - w_o \parallel_2^2 + c \parallel w - w_r \parallel_2^2 \tag{8}$$

subject to

$$\sum_{j=1}^{n} w_j = 1, \quad w_j \geq 0 \tag{9}$$

Here, a, b, and c denote the reliability coefficients of the subjective weights, objective weights, and risk weights, respectively. Since w_s, w_o, and w_r are all normalized weight vectors, the above optimization problem has the following closed-form solution:

$$w^* = \frac{a w_s + b w_o + c w_r}{a + b + c} \tag{10}$$

This paper further adaptively determines the three coefficients according to consistency, variability, and risk information. The subjective reliability is defined as

$$a = \exp\left[-\gamma \max\left(0, CR - \tau\right)\right] \tag{11}$$

where τ is the consistency threshold and γ is the decay coefficient. The objective reliability is determined by the average degree of variability:

$$b = \frac{1}{n} \sum_{j=1}^{n} (1 - e_j) \tag{12}$$

The risk constraint strength is obtained by coupling the risk prior weights with the objective weights:

$$c = \sqrt{\sum_{j=1}^{n} w_{r,j} w_{o,j}} \tag{13}$$

DOI: 10.54254/2755-2721/2026.34002

The final w* is used as the comprehensive evaluation weight.

4. Experimental results and analysis

4.1. Experimental setup

This paper selects five object detection models, namely YOLOv8n, Faster R-CNN, CenterNet, YOLOv6s, and Deformable DETR, for comprehensive evaluation. The experimental indicators include detection accuracy, inference efficiency, computational complexity, and robustness under complex imaging conditions. All indicators are mapped to the range of [0,100] according to predefined scoring rules, and the final model scores are calculated based on the risk-constrained optimized fusion method proposed in this paper.

4.2. Comprehensive evaluation results

The comprehensive scores of the models obtained using the proposed method are shown in Figure 1.

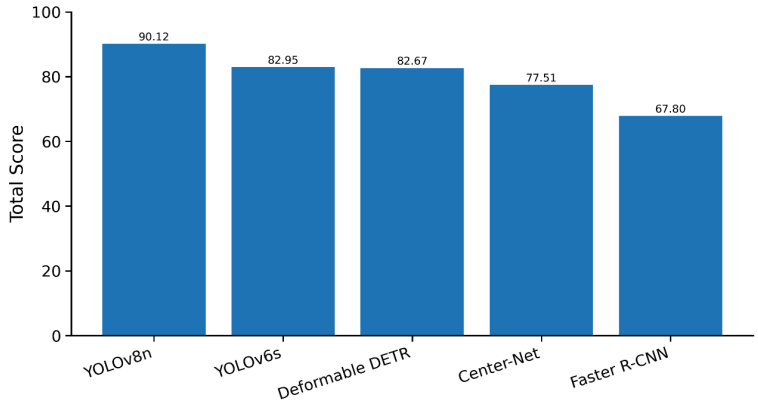

Figure 1. Overall evaluation scores by risk-constrained fusion

As shown in Figure 1, YOLOv8n achieves the highest comprehensive score, with a score of 90.12, indicating that it has the best overall performance. YOLOv6s and Deformable DETR obtain relatively close scores, ranking second and third, respectively. CenterNet performs at a moderate level, while Faster R-CNN obtains the lowest score. These results indicate that the proposed method can effectively distinguish the comprehensive performance differences among different models.

4.3. Analysis of first-level indicator results

The scores of each model on the first-level indicators are shown in Figure 2.

DOI: 10.54254/2755-2721/2026.34002

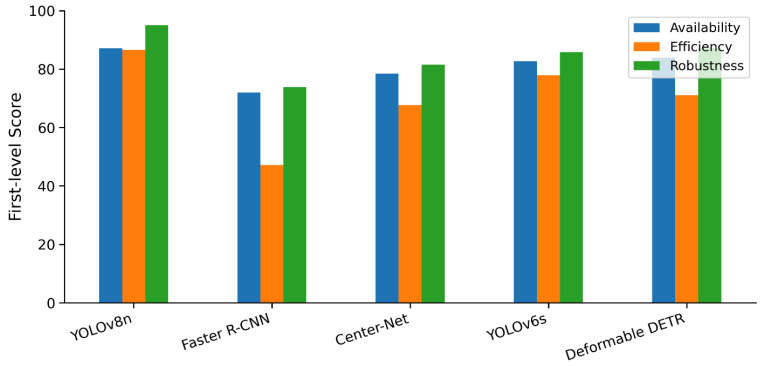

Figure 2. First-level indicator scores by the optimization-risk method

It can be seen from Figure 2 that YOLOv8n performs well in usability, efficiency, and robustness, and therefore obtains the highest comprehensive score. Faster R-CNN performs acceptably in terms of usability, but its efficiency score is lower, resulting in a lower overall ranking. Deformable DETR shows strong robustness, but its efficiency indicators are lower than those of the YOLO-series models. YOLOv6s has good advantages in efficiency and demonstrates relatively balanced overall performance.

4.4. Radar chart analysis

The radar chart of the first-level indicators is shown in Figure 3.

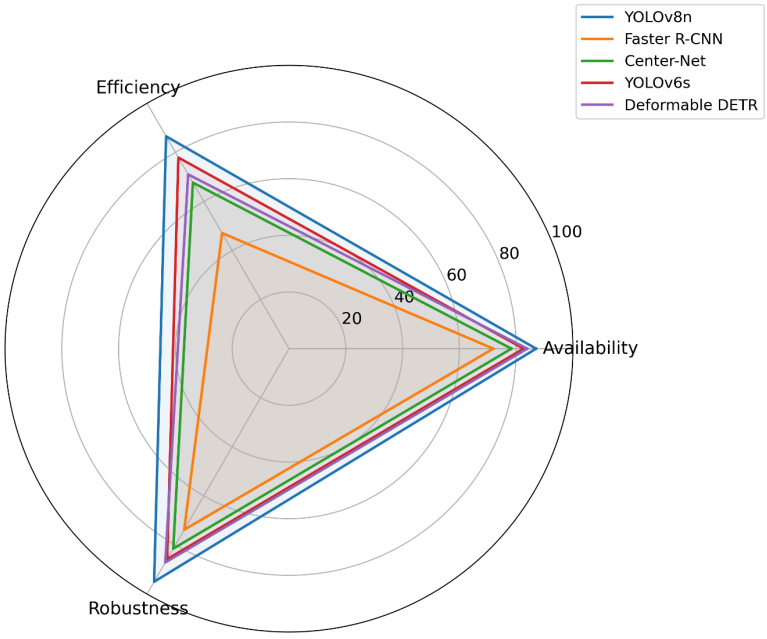

Figure 3. First-level indicator radar chart

As shown in Figure 3, YOLOv8n maintains advantages across all three first-level indicators, showing good overall balance. Faster R-CNN has an obvious shortcoming in the efficiency dimension, indicating that its application in practical deployment scenarios is limited to some extent.

Furthermore, the radar chart of the second-level indicators is shown in Figure 4.

DOI: 10.54254/2755-2721/2026.34002

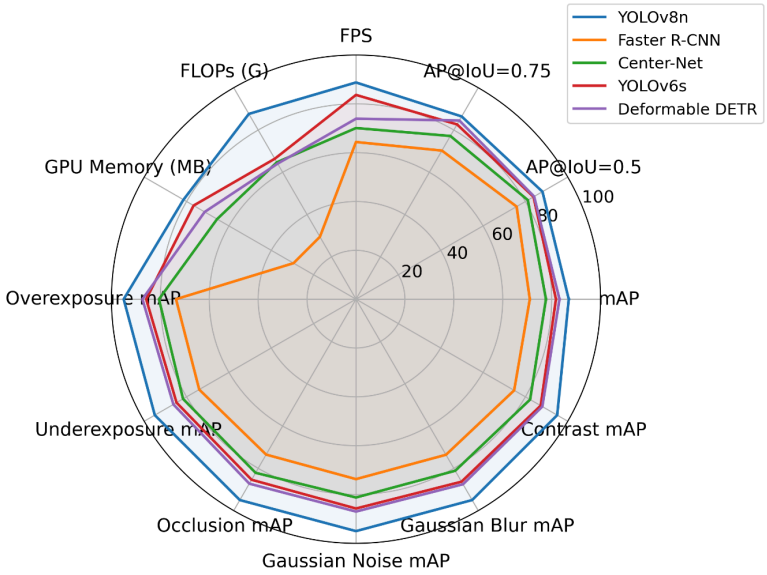

Figure 4. Second-level indicator radar chart

As shown in Figure 4, YOLOv8n achieves high scores on most second-level indicators, especially in terms of efficiency and robustness under complex scenarios. Deformable DETR and YOLOv6s are relatively close in some accuracy and robustness indicators, but Deformable DETR has relatively weaker efficiency indicators. Faster R-CNN shows obvious disadvantages in efficiency-related indicators such as FPS, FLOPs, and GPU memory usage.

5. Conclusion

This paper proposes a comprehensive evaluation method for medical image object detection models by constructing an indicator system covering usability, efficiency, and robustness. The method integrates AHP-based subjective weights, entropy-based objective weights, and medical risk prior weights through a risk-constrained optimization framework, thereby obtaining comprehensive weights that consider expert knowledge, data variability, and clinical risk. Experiments on five detection models show that the evaluation results are consistent with the actual performance characteristics of the models, with YOLOv8n achieving the best overall performance and Faster R-CNN ranking lower due to its efficiency limitations. The proposed method provides an interpretable and quantitative decision-making tool for model selection and optimization in medical image detection. Future work will include more models, datasets, and dynamic risk-aware weight adjustment mechanisms.

References

[1] T. Ye, W. Yang, X. Chen, et al., "A survey of deep learning-based object detection methods in medical image analysis, " Computers in Biology and Medicine, vol. 178, p. 108773, 2024.
[2] X. Zhao, F. Liu, S. Li, and A. Ding, "Medical image object detection with transformer: A review, " IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 8, pp. 3801-3813, 2023.
[3] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8, " GitHub, 2023. [Online]. Available: https: //github.com/ultralytics/ultralytics
[4] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information, " in Proc. Eur. Conf. Comput. Vis. (ECCV), 2024.

DOI: 10.54254/2755-2721/2026.34002

[5] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection, " in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2024.
[6] Z. Liu, H. Lin, Y. Cao, et al., "Swin Transformer V2: Scaling up capacity and resolution, " in Proc. IEEE/CVF CVPR, 2022, pp. 12009-12019.
[7] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection, " in Proc. IEEE/CVF CVPR, 2024, pp. 16965-16974.
[8] S. Huang, J. Yang, S. Fong, and Q. Zhao, "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges, " Cancer Letters, vol. 471, pp. 61-71, 2020.
[9] L. Maier-Hein, A. Reinke, P. Godau, et al., "Metrics reloaded: Recommendations for image analysis validation, " Nature Methods, vol. 21, no. 2, pp. 195-212, 2024.
[10] F. Di Salvo, S. Doerrich, and C. Ledig, "MedMNIST-C: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions, " in Proc. MICCAI Workshop Advancing Data Solutions Med. Imaging AI (ADSMI), 2024.
[11] L. Wang, X. Zhang, and H. Li, "An integrated TOPSIS-entropy weight method for multi-criteria evaluation of deep learning models in medical applications, " Expert Systems with Applications, vol. 238, p. 121823, 2024.
[12] Y. Zhang, Y. Xu, and M. Zhou, "A hybrid AHP-entropy weight approach for evaluating the performance of intelligent medical imaging systems, " Biomedical Signal Processing and Control, vol. 89, p. 105741, 2024.
[13] J. Wieckowski, W. Salabun, B. Kizielewicz, et al., "Recent advances in multi-criteria decision analysis: A comprehensive review of applications and trends, " Int. J. Knowl.-Based Intell. Eng. Syst., vol. 27, pp. 367-393, 2023.
[14] S. Chakraborty, R. D. Raut, T. M. Rofin, and S. Chakraborty, "A comprehensive and systematic review of multi-criteria decision-making methods and applications in healthcare, " Healthcare Analytics, vol. 4, p. 100232, 2023.
[15] C. Li, L. Li, H. Jiang, et al., "YOLOv6: A single-stage object detection framework for industrial applications, " arXiv preprint arXiv: 2209.02976, 2022.
[16] P. Mittal, "A comprehensive survey of deep learning-based lightweight object detection models for edge devices, " Artificial Intelligence Review, vol. 57, p. 242, 2024.
[17] R. Geirhos, C. R. M. Temme, J. Rauber, et al., "Generalisation in humans and deep neural networks, " in Proc. NeurIPS, 2022, pp. 7538-7550.
[18] J. Li, G. Li, and B. Sun, "Entropy weight method for objective evaluation of medical AI model performance under uncertainty, " Journal of Healthcare Engineering, vol. 2023, p. 6685210, 2023.