# One class classification-based quality assurance of organs-at-risk delineation in radiotherapy

Yihao Zhao[1], Cuiyun Yuan[2], Ying Liang[2], Yang Li[2], Chunxia Li[2], Man Zhao[2], Jun Hu[1], Ningze Zhong[1], Chenbin Liu[2]*

[1]School of Electronic and Communication Engineering, Sun Yat-sen University, China
Zhaoyh69@mail2.sysu.edu.cn
[2]National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
liuchenbin@cicams-sz.org.cn

## Abstract

*The delineation of tumor target and organs-at-risk (OARs) is critical in the radiotherapy treatment planning. It is also tedious, time-consuming and prone to subjective experiences. Automatic segmentation can be used to reduce the physician's workload. However, the quality assurance of the segmentation is an unmet need in clinical practice. In this study, we developed an automatic model that detects the errors of the contouring using one-class classifier. The OARs included left and right lungs, heart, esophagus, and spinal cord. Each data includes the ground truth, which is manually contoured by experienced doctor, and contour generated by a contouring software. We used three metrics to determine whether the contour of an OAR is "high" or "low" quality. A resnet-152 network performed as a feature extractor, and a one class support vector machine determines the quality of the contour. We generated certain contour errors to evaluate the generalizability of this method. Furthermore, to enhance the interpretability of this method, we conducted a set of experiments to assess its detection limit and discussed the correlation between this limit and metrics such as volume, DSC, HD95, and MSD. The proposed method showed significant improvement over binary classifiers in handling various types of errors. The relationship between the detection limit and multiple factors of the OARs indicates that our method is highly interpretable. Moreover, the model's fast execution speed can significantly reduce the burden on physicians.*

## 1. Introduction

According to the extent of user interaction, segmentation techniques could be categorized into manual, automatic, and semi-automatic methods. There are mainly two types of automatic contouring methods, atlas-based and deep learning methods. Atlas-based methods created atlas from previously annotated dataset, deformed the atlas templates to the target images, and generated the target anatomical structures.[1, 2] Deep learning methods applied multiple level of filters and max-pooling processes to extract image features (encoder), and inflated the encoder's output into a segmentation mask using convolution and up-sampling (decoder). Deep learning has demonstrated superior performance in image segmentation, including U-Net [3], 3D U-Net [4], V-Net[5], Seg-Net [6], DeepMedic[7], DeepLab[8], VoxResNet[9] and Mask RCNN [10]. However, deep learning-based systems were considered as black boxes, challenging to interpret, and prone to errors. Therefore, we need to establish a Quality Assurance (QA) system. Several studies suggest issues with the accuracy of automatic delineation models in complex scenarios and poor generalization performance. Additionally, adopting a higher-precision model may result in shorter runtime and the challenge of increased runtime.

To assess the quality of auto-segmentation, conventional practice applied an independent test dataset and diverse metrics such as DSC, $HD_{95}$ and MSD. [11] These metrics assessed the consistence between predicted contours and the reference "ground truth", which was manually delineated by the experienced physicians. DSC reflects the degree of overlap between two delineations, HD95 reflects the maximum error, and MSD reflects the average boundary error. To mitigate the substantial time and resource costs involved in generating a "ground truth" segmentation for every case, there is a need to develop an efficient and labor-effective method for identifying low-quality segmentations

## 2. Related Work

Previous researchers have designed methods for automated quality control.[12-21] These methods included feature-based methods, [14, 16, 22, 23], and deep learning based methods.[13, 15, 19, 20] feature-based methods employed in quality prediction involved the extraction of 2-D features, such as Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and texture features and 3-D features like contour volume, surface area, and orientation for predicting contour quality employed. Feature-based methods required the OAR delineations by physicians before quality can be assessed by comparing them with the automatically generated delineations. Therefore, they either consumed a significant amount of internal medicine doctors' time or consumed a large amount of computational resources. In recent years, the convolutional neural network (CNN) belongs to a category of deep neural networks commonly utilized for the analysis of visual imagery. [24, 25] Rhee et al. [13] proposed a deep-learning based system that first generate a contour and calculate the 11 quantitative metrics of contour and the test contour. However, some datasets in this study originate from the same automated delineation system, which may result in challenges when it comes to recognizing delineations from other systems. Chen et al. [20] proposed a deep learning based method that classifies the contouring into "good", "medium" and "bad" automatically by dice value. Duan et al. [19] introduced a method that is based on k-fold cross-validation dataset, which addressing the issue of a limited dataset. While their methods demonstrate commendable accuracy and efficiency, there are still areas in which we can make improvements.

Nevertheless, they use the whole CT image as the input of the network, but the OARs occupy only a small region of the image. CNN may extract a lot of useless information in the other region. We need to find a way to make the network focus on the regions where the OARs located in. Uijlings et al. propose a method that is generate candidate boxes and then proceed with identification.[26] It is called Region-CNN(R-CNN). With R-CNN, we can pay more attention on the object we want to analysis. He, K et al proposed a method called Mask R-CNN to simultaneously generating a high-quality segmentation mask for each instance, which further improved the performance of R-CNN.[10] In the quality assurance of the OARs segmentation, the mask is given by either the automatic model or by the doctor. We use a method similar to mask R-CNN to improve accuracy. What's more, as we mentioned above, none of the researchers developed a method that can detect types of different errors. Their methods can only adapt to specific contouring networks and lack of universality. One class support vector machine (OCSVM) provides a method to train the network in the absence of counter-examples. [27-30] It is widely used in abnormal detection. We noticed that one-class classifier was not utilized for quality control in organ delineation. By incorporating a one-class classifier, we can address the challenge of identifying various errors and enhance the compatibility of the quality control process. Oza et al. [31] introduced a method to add zero centered gaussian noise to the latent space as the pseudo-negative class and train the network. With the one class convolution neural network, we can train a network to detect different types of errors with high quality contours.

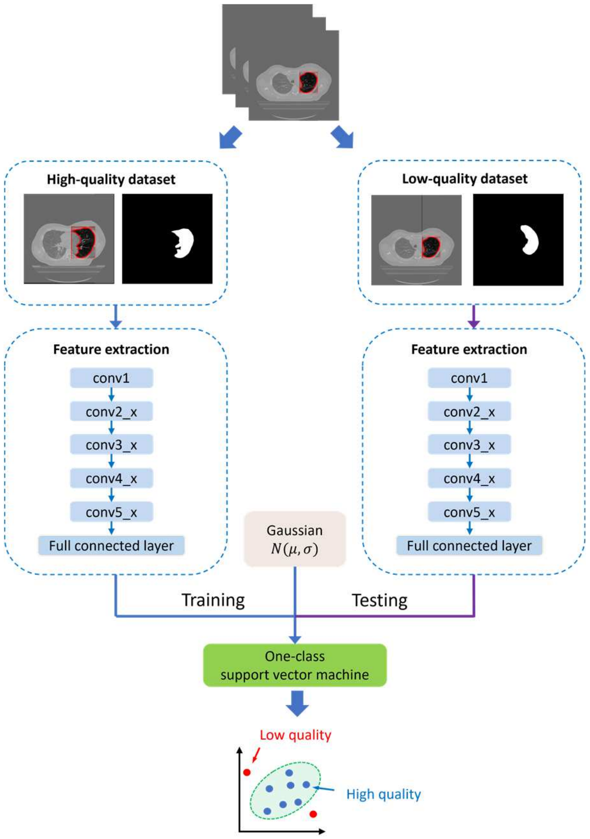

Figure 1. The whole workflow of the proposed quality assurance model.

As shown in Figure 1, this section provides an overview of several key components in our process, including data acquisition, contour evaluation, data preprocessing, feature extraction, machine learning model, and evaluation metrics.

# 3. Materials and Methods

## 3.1. Data Acquisition

The patient data utilized in our study comprised CT images and manually delineated contours from the AAPM Thoracic CT Segmentation Challenge competition, encompassing a total of 60 cases [32]. The CT images were reconstructed to encompass the entire thoracic region, with the number of slices ranging from 103 to 279. These CT scans had a consistent field of view of 50 cm and a reconstruction matrix size of 512 x 512. There was variation in slice spacing across institutions, ranging from 1 mm to 3 mm. The reported pixel size of the images ranged between 0.98 mm and 1.37 mm, with a median value of 0.98 mm.

The gold standard atlas for organ-at-risk (OAR) delineation in this study consisted of manual contours. The OARs included the left and right lungs, heart, esophagus, and spinal cord. These manual contours were created by expert clinicians following the contouring atlas guideline outlined in RTOG 1106 [33]. They served as a reference for accurately identifying and delineating these anatomical structures and were considered as the "ground truth" for comparison. To automatically generate contours, a commercially available contouring software, AccuContour$^{TM}$ (Manteia Medical Technologies Co. Ltd., Xiamen, China), was employed, utilizing deep learning techniques. The auto-generated contours were divided into two datasets based on quality: a high-quality contour dataset and a low-quality contour dataset. The objective of this study was to develop a quality assurance model capable of identifying low-quality contours. A total of 60 cases were included in this study, with 48 cases allocated to the training set and 12 cases to the test set. The selection of cases for both sets was performed randomly.

### 3.2. Contour Evaluation

We used Dice similarity coefficient (DSC) [34], the maximum Hausdorff distance ($HD_{95}$) [35], and mean surface distance (MSD) [36] to measure the contour quality. They were calculated by the follows:

$$DSC(GT\ ,\ AGC) = \frac{2|\mathrm{GT}\cap\mathrm{AGC}\ |}{|\mathrm{GT}|+|\mathrm{AGC}\ |} \quad (1)$$

where GT is the ground truth and AGC is the automatically generated contours.

$$d(X \rightarrow Y) = \max_{x\epsilon X} \min_{y\epsilon Y} \left(d^{X\rightarrow Y}\right) \quad (2)$$

$$\mathrm{HD}_{95}(GT, AGC) = max(\mathrm{d}(\mathrm{GT} \rightarrow AGC),\mathrm{d}(\mathrm{AGC} \rightarrow GT)) \quad (3)$$

where d is the one-sided Euclidean distance from point set X to point set Y. $HD_{95}$ is the longest bidirectional distance between the ground truth and automatically generated contours.

$$\mathrm{MSD}(GT, AGC) = \frac{1}{N_{GT} + N_{AGC}} \left( \sum_{x\in GT} \min_{y\in AGC} \|x - S(AGC)\| + \sum_{y\in AGC} \min_{x\in GT} \|y - S(GT)\| \right) \quad (4)$$

where $N_{GT}$ and $N_{AGC}$ are the number of the pixels in the contour of ground truth and automatically generated contours respectively. ‖·‖ denotes the Euclidean distance. S(·) denote the point set of surface voxels.

Based on the aforementioned measurement, we established the criteria for the assessment of contour quality. The high-quality contour was defined as one that met all the following requirements:

$$DSC > mean_{DSC} - \sigma_{DSC} \quad (5)$$

$$HD_{95} < \mathrm{mean}_{HD95} - \sigma_{HD95} \quad (6)$$

$$MSD < \mathrm{mean}_{MSD} - \sigma_{MSD} \quad (7)$$

In the equations (5)-(7), $\mathrm{mean}_{DSC}$, $\mathrm{mean}_{HD}$, and $\mathrm{mean}_{MSD}$ are the average value of DSC, $HD_{95}$ and MSD calculated in the training set, respectively. $\sigma_{DSC}$, $\sigma_{HD95}$, and $\sigma_{MSD}$ are the standard deviation of DSC, $HD_{95}$ and MSD calculated in the training set, respectively. Contours that fail to meet any of the aforementioned requirements were labeled as low-quality contours. In our dataset, the proportion of low-quality contours in different organs ranges from 12% to 16%. Eight representative contours were shown in figure 2.

### 3.3. Data Preprocessing

To accommodate the variability in CT scan formats from different sources, we performed normalization by converting each scan into uint8 format, where pixel values ranged from 0 to 255. This normalization not only standardized the pixel value range, but also resulted in reduced training time and GPU memory usage. Subsequently, we converted both the manual and automatic generated contour into binary mask. By identifying non-zero pixels in the binary mask, we extracted corresponding regions of interest from the CT scans, which served as training images. To meet the input requirements of the ResNet-152 network [24], we resized the images to dimensions of 224 × 224.

### 3.4. Feature Extraction

We used Resnet-152 as the feature extractor of our model [24]. Its network architecture introduced the concept of residual learning, addressing the issues of vanishing and exploding gradients during the training of deep neural

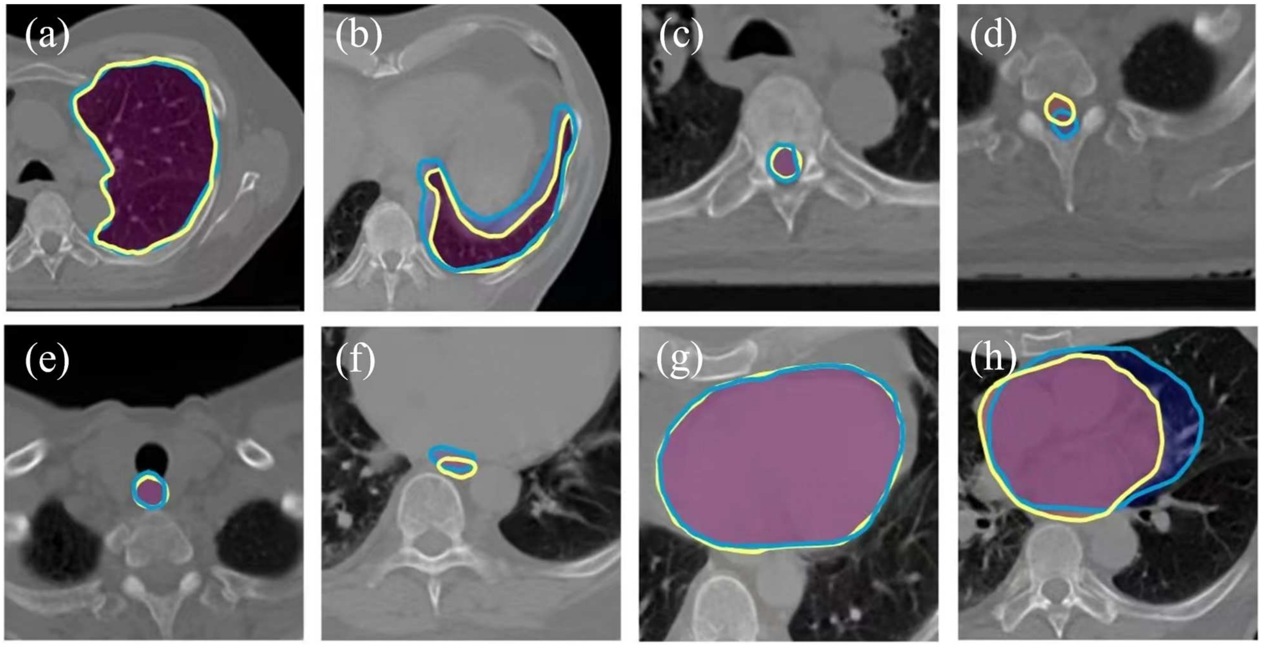

Figure 2. The representative contours created by expert clinicians and deep learning techniques. (a) high quality contour of left lung; (b) low quality contour of left lung; (c) high quality contour of spinal cord; (d) low quality contour of spinal cord; (e) high quality contour of esophagus; (f) low quality contour of esophagus; (g) high quality contour of heart; (h) low quality contour of heart. The yellow lines were contours generated by expert clinicians. The blue lines were contours created by deep learning techniques.

networks through skip connections and identity mappings across layers. ResNet-152 has gained renown for its exceptional depth and high performance, characterized by its 152 layers and proficiency in large-scale image recognition tasks. The key innovation of this network lies in the introduction of residual blocks, enabling the model to learn residual mappings. This meant the network could optimize by learning the residual between the target mapping and the input and our model will have a better performance [24].

### 3.5. One-class Support Vector Machine

Due to the potential presence of various error types within the automatically generated contours and the predominant inclusion of high-quality contours in the training data, a binary classifier could encounter challenges associated with an imbalanced dataset. One-class classification algorithms, referred to as outlier or anomaly detection, is designed to identify instances that differ significantly from the majority class [28]. Its primary objective is to determine whether a given instance belongs to the target class or not, without explicit knowledge or training examples from other classes. In this study, one class support vector machine (OC-SVM) [28] was used. The advantage of OC-SVM is its capacity to detect diverse types of errors, even those that were not encountered during the training process. Furthermore, OC-SVM exhibits superior error detection capabilities compared to a binary classifier. The objective of the OC-SVM found a maximum margin hyperplane in feature space. It solved the following function:

$$\min_{\omega,b,\xi} \quad \frac{1}{2}\|\boldsymbol{\omega}\|_{F_k}^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \qquad (8)$$

where $\boldsymbol{\omega} \in F_k$ , $\rho$ is the distance from origin to hyperplane $\boldsymbol{\omega}$. Nonnegative slack variables $\xi_i$ allow the margin to be soft, but violations $\xi_i$ get penalized. $\|\boldsymbol{\omega}\|_{F_k}^2$ is a regularizer on the hyperplane $\boldsymbol{\omega}$ and $\|\cdot\|_{F_k}^2$ is the norm induced by $\langle\cdot,\cdot\rangle_{F_k}$. In our study, OC-SVM was trained using high-quality contours and subsequently employed to identify low-quality contours. During the training of OC-SVM (Figure 1), we introduced zero-mean Gaussian noises as abnormal samples to the fully connected layer [33].

### 3.6. Evaluation

#### 3.6.1 Prediction Evaluation

We used balanced accuracy, F score, sensitivity, specificity, and AUC to measure the performance of OC-

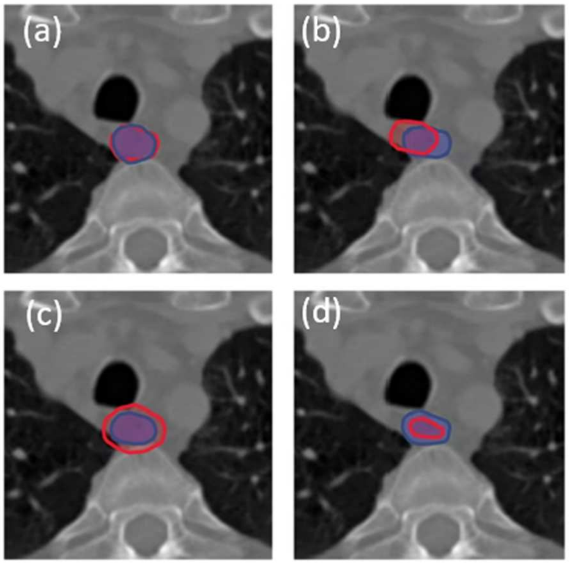

Figure 3. The contours with generated errors. (a) The automatically generated contour of the esophagus was labeled as high-quality, (b) the esophagus contour with translation error, (c) the esophagus contour with an enlargement error. (d) the esophagus contour with a shrinkage error

SVM model. Balanced accuracy is a metric that helps mitigate sample size imbalances. It is defined as follow:

$$BA = \frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right) \tag{9}$$

where TP is the number of the samples with low-quality contour that are accurately identified by the model, FN represents the number of samples with high-quality contour that are inaccurately predicted, TN is the number of samples with high-quality contour that are correctly detected, and FP is the number of samples with contour errors that are incorrectly identified.

F-score measures the accuracy of a prediction model on a dataset. It is the harmonic mean of the precision and recall:

$$F\ score = \frac{2TP}{2TP+FN+FP} \tag{10}$$

Sensitivity is the probability of a low-quality test result, given that the contour is truly low-quality, which evaluates the ability to identify low-quality contours. Specificity is the probability of a high-quality test result, given that the contour is truly high-quality, which reflects the ability to identify high-quality contours. They are defined as follows:

$$Sensitivity = \frac{TP}{TP+FN} \tag{11}$$

$$Specificity = \frac{TN}{TN+FP} \tag{12}$$

### 3.6.2 Extended Evaluation using Generated Errors

To expand the sample size and thoroughly evaluate the performance of OC-SVM, we additionally generated various types of errors derived from the automatically generated contours. The errors encompassed translation and resizing. To simulate a translation error, we started by randomly selecting a direction and subsequently applied a displacement to the contour using the direction and a designated distance. To create an enlargement error, a dilation kernel with a disk of radius 2 was employed to progressively enlarge the contour until the low-quality was achieved. The reduction error was generated using an erosion kernel, following a similar procedure as the enlargement error, but with the aim of reducing the contour size. These generated contours that did not meet the criteria (equation 6-8) mentioned above were labeled as low-quality as shown in Figure 3. For each high-quality contour, we generated three distinct types of low-quality contours, resulting in a total number of the generated low-quality contours of 2634.

In our experiment, an NVIDIA GeForce RTX 3080 was used in the proposed quality assurance model (Figure 1). Due to the limitation of our GPU memory, the batch size was set to 32. We used Adam optimizer in the training of ResNet-152. The total epochs was set to 50. The training process took about 3 hours. Testing involved identifying 2000 contours within 5 minutes. On average, the prediction of quality assurance model for the contour on a single slice took 150 ms.

Table 1. The comparison of CNN model and the proposed model on the test dataset.

| **OAR** | **BA** | | **F score** | | **Sensitivity** | | **Specificity** | | **AUC** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CNN | Proposed | CNN | Proposed | CNN | Proposed | CNN | Proposed | CNN | Proposed |
| **Esophagus** | 0.92 | 0.96 | 0.93 | 0.97 | 0.93 | 0.96 | 0.91 | 0.98 | 0.95 | 0.96 |
| **Heart** | 0.95 | 0.98 | 0.96 | 0.98 | 0.96 | 0.97 | 0.94 | 1.00 | 0.95 | 0.97 |
| **Left lung** | 0.98 | 0.99 | 0.91 | 0.98 | 0.96 | 0.98 | 1.00 | 1.00 | 0.93 | 0.97 |
| **Right lung** | 0.97 | 0.99 | 0.92 | 0.98 | 0.97 | 0.98 | 1.00 | 1.00 | 0.94 | 0.97 |
| **Spinal cord** | 0.91 | 0.96 | 0.92 | 0.96 | 0.93 | 0.94 | 0.89 | 0.98 | 0.91 | 0.95 |

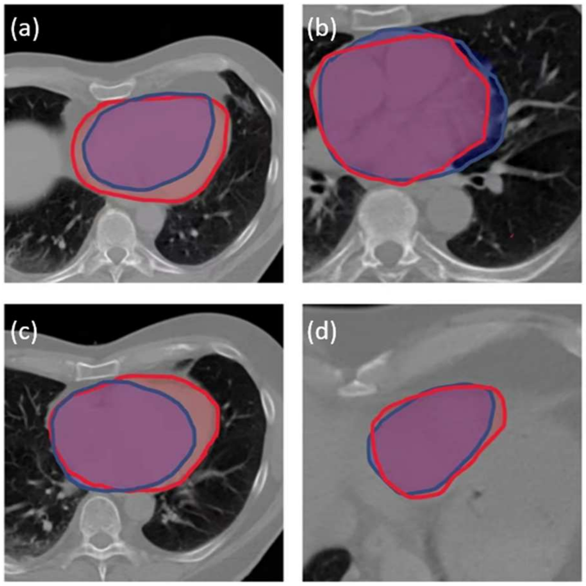

Figure 4. The comparation of the errors detected by our method and CNN method. The contour errors in (a) and (c) were detected by both CNN method and the proposed method. The contour errors in (b) and (d) were only detected only by the proposed method. The red line was gold standard. The blue line indicated the contour generated by deep learning technique.

## 4. Result

4.1. Comparison of CNN model and the proposed model

We evaluated the performance of CNN model [20] and our proposed quality assurance model on the test dataset. As shown in table 1, the proposed model exhibited higher balanced accuracy compared to the CNN model (Esophagus: 0.96 vs. 0.92; heart: 0.98 vs. 0.95; left lung: 0.99 vs. 0.98; right lung: 0.99 vs. 0.98; spinal cord: 0.96 vs. 0.91). The F-scores of the proposed model outperformed those of the CNN model (Esophagus: 0.97 vs. 0.93; heart: 0.98 vs. 0.96; left lung: 0.98 vs. 0.91; right lung: 0.98 vs. 0.92; spinal cord: 0.96 vs. 0.92). The sensitivity of the proposed model was superior to that of the CNN model (Esophagus: 0.96 vs. 0.93; heart: 0.97 vs. 0.96; left lung: 0.98 vs. 0.96; right lung: 0.98 vs. 0.97; spinal cord: 0.94 vs. 0.93). The specificity of the proposed model surpassed that of the CNN model (Esophagus: 0.98 vs. 0.91; heart: 1.00 vs. 0.94; left lung: 1.00 vs. 1.00; right lung: 1.00 vs. 1.00; spinal cord: 0.98 vs. 0.89). The AUC of the proposed model exceeded that of the CNN model (Esophagus: 0.96 vs. 0.95; heart: 0.97 vs. 0.95; left lung: 0.97 vs. 0.93; right lung: 0.97 vs. 0.94; spinal cord: 0.95 vs. 0.91). The proposed model was able to achieve higher detection accuracy compared with CNN model. As shown in figure 4, we showed some errors detected by our method and missed by the traditional CNN method.

4.2. The performance of the proposed method on generated dataset

The balanced accuracy of the predication on different generated errors was shown in Table 3. The proposed method achieved a higher balanced accuracy in identifying reduction errors compared to the CNN method (Esophagus: 0.82 vs 0.46; heart: 0.83 vs 0.67; left lung: 0.85 vs 0.66; right lung: 0.85 vs 0.66; spinal cord: 0.81 vs 0.44). In terms of identifying enlargement errors, the proposed method demonstrated a superior balanced accuracy in comparison to the CNN method (Esophagus: 0.88 vs 0.50; heart: 0.86 vs 0.69; left lung: 0.88 vs 0.70; right lung: 0.88 vs 0.70; spinal cord: 0.83 vs 0.49). The balanced accuracy of the proposed method in identifying translation errors surpassed that of the CNN method (Esophagus: 0.88 vs 0.51; heart: 0.89 vs 0.70; left lung: 0.89 vs 0.70; right lung: 0.89 vs 0.70; spinal cord: 0.87 vs 0.49). As shown in figure 5, the generated errors were detected by our method and missed by the traditional CNN method (Figure 5c & Figure 5d).

## 5. Discussion

In traditional approaches, researchers use binary classifiers to categorize errors[18, 20]. However, this study introduced single classifiers and attention mechanisms, which provide the model with increased versatility and enhanced precision compared to previous research. For binary classifiers, when the test set contains errors significantly different from those in the training set, the model's recognition performance tends to deteriorate. Leveraging the characteristics of a one-class classifier, we

Table 2. The comparison of the identification performance on generated dataset

| **OARs** | **Reduction** | | **Enlargement** | | **Translation** | |
|---|---|---|---|---|---|---|
| | CNN | Proposed | CNN | Proposed | CNN | Proposed |
| **Esophagus** | 0.46 | 0.82 | 0.50 | 0.88 | 0.51 | 0.88 |
| **Heart** | 0.67 | 0.83 | 0.69 | 0.86 | 0.70 | 0.89 |
| **Left lung** | 0.66 | 0.85 | 0.70 | 0.88 | 0.70 | 0.89 |
| **Right lung** | 0.66 | 0.85 | 0.70 | 0.88 | 0.70 | 0.89 |
| **Spinal cord** | 0.44 | 0.81 | 0.49 | 0.83 | 0.49 | 0.87 |

apply it to the proposed method to enhance the detection capability for errors that have not been encountered before. Apart from the initial training, this method, unlike the approach proposed by Duan et al. [19], doesn't require ground truth for quality assessment.

The use of a one-class classifier effectively addresses the issue of having fewer error samples. However, in past studies, the majority of researchers[15, 19, 20] did not pay much attention to this issue. The scarcity of error samples has been a challenge, resulting in suboptimal performance for previous methods. However, even a simple approach of generating error samples using the method mentioned in this paper and training a binary classifier can significantly enhance the classifier's performance. Henderson et al. [23] proposed a method of voxel-wise to generate errors. They generated signed distance transforms of the ground-truth segmentations and introduced structured noise on a voxel-wise basis. The structured noise was generated by drawing from a normal distribution, which was subsequently convolved with a 7.5 mm Gaussian kernel. The amplitude of the convolution was adjusted to ensure that the resulting structured noise had a standard deviation of 1 mm. The perturbed segmentation was obtained by utilizing the marching cubes algorithm (at level = 0) to generate a triangular mesh manifold. Subsequently, a connected components algorithm was employed to eliminate any disconnected spurious segmentations. However, implementing this method in three dimensions requires using 3D kernels for convolution, and this might lead to high-performance demands for the process. If not artificially generate error samples, we can mitigate this issue by using a weighted cross-entropy loss function or employing cross-validation methods.

## 6. Conclusion

We proposed a new QA model. This model incorporates two significant improvements, the one class classifier and attention mechanism. They can enhance the model's generality and precision. We can detect different errors with the model. Moreover, the model's fast execution speed can significantly reduce the burden on physicians.

**Acknowledgments:** Supported by National Natural Science Foundation of China (No. 12005301), Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515012456), Shenzhen High-level Hospital Construction Fund, Shenzhen Key Medical Discipline Construction Fund (No. SZXK013), Sanming Project of Medicine in Shenzhen (No. SZSM201612063), Shenzhen High-level Hospital Construction Fund, Hospital Research Project (No. SZ2020MS002, E010322028, E010321015).

## References

[1] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, M. B. J. C. m. Cuadra, and p. i. biomedicine, "A review of atlas-based segmentation for magnetic resonance brain images," vol. 104, no. 3, pp. e158-e177, 2011.

[2] M. Bach Cuadra, V. Duay, J.-P. J. H. o. B. I. M. Thiran, and C. Research, "Atlas-based segmentation," pp. 221-244, 2015.

[3] O. Ronnerberger, P. Fischer, and T. Brox, "U-Net: Convolutional Neural Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, 2015.

[4] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 2016, pp. 424-432: Springer.

[5] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, 2016, pp. 565-571: Ieee.

[6] V. Badrinarayanan, A. Kendall, R. J. I. t. o. p. a. Cipolla, and m. intelligence, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," vol. 39, no. 12, pp. 2481-2495, 2017.

[7] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," vol. 36, pp. 61-78, 2017.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. J. I. t. o. p. a. Yuille, and m. intelligence, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," vol. 40, no. 4, pp. 834-848, 2017.

[9] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. J. N. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," vol. 170, pp. 446-455, 2018.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961-2969.

[11] A. A. Taha and A. J. B. m. i. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," vol. 15, no. 1, pp. 1-28, 2015.

[12] T. Wang, Y. Chen, M. Qiao, and H. J. T. I. J. o. A. M. T. Snoussi, "A fast and robust

convolutional neural network-based defect detection model in product quality control," vol. 94, pp. 3465-3471, 2018.

[13] D. J. Rhee *et al.*, "Automatic contouring QA method using a deep learning–based autocontouring system," vol. 23, no. 8, p. e13647, 2022.

[14] H. Nourzadeh *et al.*, "Knowledge-based quality control of organ delineations in radiation therapy," vol. 49, no. 3, pp. 1368-1381, 2022.

[15] K. Men, H. Geng, T. Biswas, Z. Liao, and Y. J. F. i. O. Xiao, "Automated quality assurance of OAR contouring for lung cancer based on segmentation with deep active learning," vol. 10, p. 986, 2020.

[16] C. McIntosh, I. Svistoun, and T. G. J. I. t. o. m. i. Purdie, "Groupwise conditional random forests for automatic shape classification and contour quality assessment in radiotherapy planning," vol. 32, no. 6, pp. 1043-1057, 2013.

[17] A. Kis, F. Kovács, and P. Szolgay, "Analogic CNN algorithms for textile quality control based on optical and tactile sensory inputs," in *Proc. of The 8th IEEE International Biannual Workshop on Cellular Neural Networks and their Applications, Budapes*, 2004.

[18] C. B. Hui *et al.*, "Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach," vol. 45, no. 5, pp. 2089-2096, 2018.

[19] J. Duan *et al.*, "Contouring quality assurance methodology based on multiple geometric features against deep learning auto-segmentation," 2023.

[20] X. Chen *et al.*, "CNN-based quality assurance for automatic segmentation of breast cancer in radiotherapy," vol. 10, p. 524, 2020.

[21] M. Altman *et al.*, "A framework for automated contour quality assurance in radiation therapy including adaptive techniques," vol. 60, no. 13, p. 5199, 2015.

[22] Y. Zhang, T. E. Plautz, Y. Hao, C. Kinchen, and X. A. J. M. p. Li, "Texture-based, automatic contour validation for online adaptive replanning: a feasibility study on abdominal organs," vol. 46, no. 9, pp. 4010-4020, 2019.

[23] E. G. Henderson, A. F. Green, M. van Herk, and E. M. Vasquez Osorio, "Automatic identification of segmentation errors for radiotherapy using geometric learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 319-329: Springer.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[26] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. J. I. j. o. c. v. Smeulders, "Selective search for object recognition," vol. 104, pp. 154-171, 2013.

[27] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," in *Artificial Intelligence and Cognitive Science: 20th Irish Conference, AICS 2009, Dublin, Ireland, August 19-21, 2009, Revised Selected Papers 20*, 2010, pp. 188-197: Springer.

[28] D. M. J. Tax, "One-class classification: Concept learning in the absence of counter-examples," 2002.

[29] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. J. N. c. Williamson, "Estimating the support of a high-dimensional distribution," vol. 13, no. 7, pp. 1443-1471, 2001.

[30] L. Ruff *et al.*, "Deep one-class classification," in *International conference on machine learning*, 2018, pp. 4393-4402: PMLR.

[31] P. Oza and V. M. J. I. S. P. L. Patel, "One-class convolutional neural network," vol. 26, no. 2, pp. 277-281, 2018.

[32] J. Yang *et al.*, "Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017," *Medical Physics,* Article vol. 45, no. 10, pp. 4568-4581, Oct 2018.

[33] M. Machtay, M. Matuszak, J. Bradley, V. Hirsh, R. Ten Haken, and D. J. U. Pryma, "NRG ONCOLOGY ECOG-ACRIN RTOG 1106/ACRIN 6697 RANDOMIZED PHASE II TRIAL OF INDIVIDUALIZED ADAPTIVE RADIOTHERAPY USING DURING-TREATMENT FDG-PET/CT AND MODERN TECHNOLOGY IN LOCALLY," 2014.

[34] T. J. B. s. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons," vol. 5, pp. 1-34, 1948.

[35] H. Blumberg, "Hausdorff's Grundzüge der Mengenlehre," 1920.

[36] T. Heimann *et al.*, "Comparison and evaluation of methods for liver segmentation from CT datasets," vol. 28, no. 8, pp. 1251-1265, 2009.