Performance Comparison and Analysis of Multiple Methods for Lung Cancer Detection Based on LUNA16

Zhongxing Ge *

School of Communications and Information Engineering, Xi'an University of Posts & Telecommunications, Xi'an, Shaanxi, China

* Corresponding Author Email: cpsandwich1@outlook.com

Abstract. Lung cancer is one of the most common and deadly types of cancer worldwide. While computed tomography (CT) of the chest is considered the gold standard for early detection, manual evaluation faces two major challenges: Micronodules are often overlooked, and radiologists suffer from reduced efficiency due to overwhelming daily case numbers and persistent fatigue. To solve these problems, we employ the LUNA16 dataset (including 888 CTs, 1186 annotated nodules and more than 550K candidate samples) to build an integrated experimental platform and systematically compare four kinds of detection methods: traditional manual features, 2D deep learning, 3D deep learning and transformer model. The evaluation metrics include detection accuracy (AUC-ROC curve and recall), ability of micronodule detection, robustness on limited training data (10% / 25% / 50% data amount) and computation efficiency (parameter scale and frame rate). The experimental results show that 3D Swin Transformer T has the best performance. Compared with the traditional method HOG+Random Forest, test AUC-ROC reaches 0.948 and the recall of micro-nodule is 89.7%, which is 34.4% improvement. 3D ResNet-50 has a good robustness in small sample situation (AUC could still be 0.902 when using 10% data for training), and 2D ResNet-50 could get a balance between fast speed and accuracy (85 fps). The experimental results provide support for the selection of model in clinical computer-aided system for lung cancer detection.

Keywords: 3D Swin Transformer; 3D Deep Learning; Micronodule Detection; Lung Cancer Detection; LUNA16.

1. Introduction

Based on the World Health Organization statistics in 2023, lung cancer is the leading cause of cancer death, accounting for more than 2.2 million new cases annually, which represents 18.7% of all cancer deaths [1]. The prognosis of patients depends on the early detection: the 5-year survival probability of patients with stage I after surgery is over 80%; while the survival rate of patients in advanced stages (III-IV) is less than 15%. Chest CT can provide high resolution images for lung tissue, so it is widely used for early detection [2][3]. However, there are two challenges in manual evaluation of images: due to the small size and weak features, micronodules (<5mm) are easily missed when they are highly likely to be confused with vascular or chest wall structures, the misdiagnosis ratio exceeds 25% [4]; radiologists in tertiary hospitals need to evaluate 200-300 CT images (with 50-200 slices per image) every day, which makes the evaluation inefficient. The consistently high workload will lead to tiredness and misjudgment, so it is urgent to evaluate the application possibility of an efficient and accurate CAD method based on computer vision for early detection.

There are four main categories of current computer vision methods for lung cancer detection. These methods have different advantages and disadvantages. Traditional manual feature methods include lung segmentation, handcrafted feature extraction (LBP, HOG) and classification (SVM, Random Forests). Traditional methods are interpretable, but they are easily affected by noise and are not suitable for micronodule detection. 2D deep learning methods take each CT slice as input and use networks like Faster R-CNN [5] and U-Net. Compared with traditional methods, 2D models can achieve satisfactory performance in standard detection tasks. However, CT slices have inter-spatial information, but 2D models lose this information and their adaptability to small nodules is limited. 3D deep learning methods take volumetric data as input and use 3D convolutional kernels to model spatial features in different slices. 3D ResNet and V-Net can be used to model nodule features (The

recall of Setio et al. on LUNA16 is 85% [6]). 3D deep learning methods have large parameter complexity and need a large amount of data. Transformer based approaches take volumetric data as input and use self-attention to model global dependencies. Swin Transformer achieves a good balance between accuracy and efficiency [7]. Different from existing methods, self-attention models' global dependencies and are interpretable. However, there are few comparative evaluations on lung nodule detection tasks and few works on small-sample scenarios. Therefore, existing research still has three deficiencies: a lack of comprehensive comparison on all four approaches using one single dataset, a lack of attention on micronodule detection (<5mm), and a lack of study on method robustness on limited annotated data.

This research aims to systematically analyze the performance of four types of computer vision methods on the LUNA16 dataset and find their strengths and applications in clinical scenarios. The research includes establishing a uniform experimental framework based on standardized pre-processing, training and evaluation procedures, so as to fairly compare the four types of computer vision methods. The two-evaluation metrics, micro-nodule detection ability and robustness in small sample conditions, are selected to complement the deficiency of existing research. Based on the experimental results, the recommendations of model selection will be given based on the balance of accuracy and efficiency, as well as data requirements to support various clinical applications, such as precision diagnosis and real-time screening.

2. Dataset and Methods

2.1. Dataset (LUNA16)

2.1.1 Dataset Overview

Lung Nodule Analysis 2016 (LUNA16) is an authoritative public dataset for lung nodule detection, released by NIH and other institutions together, which is published to facilitate the development and comparison of algorithms for lung nodules [8]. Lung nodule detection algorithm. The dataset is released with 888 chest CT images, organized into 10 subgroups and the total capacity is about 50 GB. All CT data use slice thicknesses of no more than 1 millimeter to ensure the nodule can be clearly distinguished. Two authoritative annotation files are released to support the nodule detection. The first file is annotations.csv, 1186 nodule annotations verified by radiologists. Each data record includes: Series ID of CT image containing this nodule (series_uid), XYZ position of center of nodule, Diameter of nodule (in millimeters). The second released file is candidates.csv. It contains more than 550k candidate nodes. Each data record contains the series ID and the XYZ position of the center of the candidate node. The category label of this node is 1 (true node) or 0 (negative sample). This data set is convenient for positive and negative samples balanced training in model training.

2.1.2 Data Partitioning

In order to make the above results fully reproducible and reliable, we split the LUNA16 dataset into a training set, a validation set and test set with 7: 2: 1 ratio as follows: is used as training set for to learn the model parameters consists of CT scans; for hyperparameter optimisation and early termination evaluation consists of CT scans; for final evaluation. The data partitioning method is a 'stratified sampling method by series ID'. In this way, the number of nodule sizes in each part (divided into three categories: less than 5 mm, 5–10 mm and greater than 10 mm) is consistent with the original data partition. This approach reduces experimental bias as much as possible due to uneven data distribution.

2.1.3 Data Preprocessing

Considering the characteristics of CT images and the model's input requirements, we designed the following data preprocessing workflow. First, HU value normalization is implemented [9]. HU value of CT images varies in the range of [-1000, 4000]. Air, water, and bone have HU values of -1000, 0, and larger than 1000, respectively. Pulmonary tissue features are highlighted by first clipping HU

values to [-1000, 1000], thus eliminating interference from other structures, such as bones. Subsequently, linear normalization maps the above values to [0, 1] based on Eq. (1).

$$HU_{norm}=\frac{HU-(1000)}{1000-(-1000)}=\frac{HU+1000}{2000} \quad (1)$$

Next, nodule region extraction is implemented. According to the nodule center XYZ coordinates in the annotations.csv file, fixed-size voxel blocks are extracted as the model input. Since the dimensional requirements of different networks are different, we designed two different preprocessing pipelines for 3D models and 2D models separately. For instance, for 3D models, including 3D ResNet-50, 3D V-Net, and 3D Swin Transformer-T, the input voxel block size is 32×48×48 (depth × height × width). For instance, for 2D models, including 2D ResNet-50 and 2D U-Net, the input is the nodule center slice plus two layers above and two layers below the nodule center slice, i.e., five layers in total. Each layer is resized to 224×224, and thus the input tensor is 5×224×224.

To solve the problem of overfitting due to the lack of medical data, multiple data augmentation methods were applied in the training process [10]. Firstly, for spatial augmentation, we applied random rotation (±10°), horizontal flipping (probability 0.5), and scaling (0.8~1.2 times). Secondly, for intensity augmentation, we applied HU value jittering (±5%) and Gaussian noise (σ=0.01) to simulate noise in clinical CT images.

2.2. Comparison Method Design

In order to ensure fair and practical comparisons, four categories of models were selected and implemented using PyTorch.

2.2.1 Conventional manual feature methods

We used LBP + SVM and HOG + Random Forest in our study. For LBP + SVM, we used circular LBP operators (radius=3, neighborhood=8) to extract pixel-level features on CT images, and then counted the histogram (dimension=10) of these features and input them into the SVM with RBF kernel (C=10, γ=0.1) to classify. The 5-fold cross-validation was used to optimize the SVM parameters. For HOG + Random Forest, we used the HOG feature descriptor to extract features from the 2D slice computed in central CT layers [12]. HOG used a window of 16×16, a block of 8×8, and a cell of 4×4 with 9 orientations, which generated a 3780-dimensional vector as input for the Random Forest classifier with 100 trees (max depth=10). These models are computationally efficient and interpretable, but limited in detecting micro-nodules.

2.2.2 2D deep learning model

The 2D deep learning models incorporated ResNet-50 and U-Net. We modified the 2D ResNet-50 model to accept the input of 5 CT slices (input channels=5). In addition, modifications were made to the model's output to align it with the final binary classification output. The weights of the initial layers were fixed, while the remaining layers underwent a fine-tuning process (learning rate = 1e-4, batch size = 32). The four-level encoder-decoder architecture with skip connections in the 2D U-Net model [11] was employed. Dice loss and cross-entropy loss were used in the training process for segmentation and classification, respectively. The batch size of this model was 16. These models are capable of reaching a moderate level of parameter complexity and fast inference, which are suitable for real-time applications.

2.2.3 3D deep learning model

In 3D ResNet-50, we replace the 2D convolutions with 3D kernels (input channel = 1)[12], while keeping the residual connections and train the model from fundamentals (batch size = 16, learning rate = 1×10^{-4}). While in 3D V- Net [13], we used a five-level 3D encoder-decoder with residual blocks, Dice loss for segmentation, batch size = 8 and mixed precision training, which can distinguish nodules from vascular structures smoothly. Obviously, these two models directly capture volumetric

features, which improve the performance of micro-nodule detection and achieve better performance with a small sample.

2.2.4 Transformer approach

ViT-B/16 and 3D Swin Transformer-T were adopted. ViT-B/16 took 224×224×3 images from central slices [14], 12-layer multi-head attention, fine-tuned pre-trained weights (learning rate 5e-5, batch size 32). 3D Swin Transformer-T took 32×48×48×1 voxels input, hierarchical attention (channel 96→768) [7], mixed-precision training (batch size 8, learning rate 1e-4), balance 3D spatial features and global dependencies. 3D Swin Transformer-T has 28.1M parameters, while achieving higher efficiency and accuracy than ViT-B/16.

3. Experiments and Results Analysis

3.1. Experimental Setup

3.1.1 Hardware environment

The hardware configuration of the experimental environment was as follows: GPU: NVIDIA GeForce RTX 4060; CPU: Intel i9-13900H; Memory: 16GB DDR (3200MHz); Storage: 1TB SSD.

3.1.2 Software environment

The software environment of experimental environment was as follows: Windows 10 operating system; PyTorch 2.0.1+cu118 and TorchVision 0.15.2 as deep learning framework; SimpleITK 2.2.1 as medical image processing package, used to read CT. mhd/.raw files; Data processing used NumPy 1.24.3, Pandas 2.0.3 and Scikit-learn 1.3.0, mainly used in metric calculation; visualization used Matplotlib 3.7.1, used to draw ROC curve and line chart.

3.1.3 Training strategy

In order to make the test results as fair as possible, all the deep learning models (except for the traditional method) adopt the same training strategy. The specific strategy is as follows: first, the AdamW optimizer is selected, with the weight decay coefficient set to 1e-4 in order to reduce the risk of overfitting. Next, a learning rate scheduling mechanism is implemented: the initial learning rate is set to 1e-4, and cosine annealing is used to reduce the learning rate to 0.5 every 10 epochs, facilitating convergence in later phases. Third, the batch size for 2D models is set to 32, while 3D models and Transformer models use a batch size of 16 due to GPU memory limitations. The training process consisted of 50 epochs and used an early stopping strategy: training was stopped when the AUC-ROC value of the validation set decreased for five consecutive epochs. Regarding loss functions, the cross-entropy loss function was used for binary classification tasks, while a weighted combination of Dice loss and cross-entropy loss (weighted 1:1) was used for segmentation tasks.

3.2. Experimental Results and Analysis

3.2.1 Overall performance comparison

Table 1. OversAll Performance Comparison of LUNA16 Test Set

Method type	model	Acc(%)	Prec(%)	Rec(%)	F1(%)	AUC-ROC	Params (M)	FPS
Traditional handmade features	LBP+SVM	78.2	72.5	68.3	70.4	0.765	0.1	120
	HOG + random forest	80.5	75.1	71.2	73.1	0.792	0.2	105
2D Deep Learning	2D ResNet-50	88.6	84.3	82.5	83.4	0.898	25.6	85
	2D U-Net	89.3	85.7	83.8	84.7	0.905	19.8	60
3D Deep Learning	3D ResNet-50	92.1	89.5	88.7	89.1	0.932	31.2	45
	3D V-Net	93.5	91.2	90.3	90.7	0.941	28.7	38
Transformer-based	ViT-B/16	90.2	87.1	86.4	86.7	0.918	86.8	30
	3D Swin Transformer-T	94.8	92.6	91.8	92.2	0.948	28.1	32

Table 1 shows the overall performance of eight models on the LUNA16 test dataset. In terms of test results, we can find that there is a large performance difference among the four kinds of methods: Transformer and 3D Deep Learning methods are superior to the 2D Deep Learning method, and the traditional manual characterization method has a relatively poor performance. In the case of the traditional method, HOG and Random Forest achieve a little better performance than LBP and SVM, and the AUC-ROC value reaches 0.792; while its recall rate of only 71.2% cannot meet the clinical requirement of false negative rate control. However, these methods still have certain advantages in computational efficiency: these methods achieve a frame rate above 100 FPS with low memory requirements, and can be used as a preliminary screening tool in a clinical environment with limited conditions.

In the the case of a two-dimensional deep learning method, compared with 2D ResNet-50 (0.898), the 2D U-Net model with segmentation branches achieves a slightly better result, whose AUC-ROC value reaches 0.905. However, its inference speed of only 60 FPS is slower than 2D ResNet-50 with 85 FPS. In general, the main drawback of the two-dimensional method is that there is no information on the Z-axis, which leads to a gap in modeling node features between two adjacent layers.

Among the three-dimensional deep models, 3D V-Net achieved better performance than 3D ResNet-50 in terms of AUC-ROC value (0.941) and recovery rate (90.3%), which might be mainly attributed to the better localization ability of its segmentation branch for the nodal regions. All three-dimensional models have moderate parameters (from 28.7 to 31.2 million) and are advantageous in terms of accuracy and operation efficiency.

Among all transformer-based approaches, 3D Swin Transformer-T achieves the best performance, with an AUC-ROC value of 0.948 and a recall rate of 91.8%. This is because 3D Swin Transformer-T uses a three-dimensional window attention mechanism to obtain global features and avoid the limitations of the local receptive field existing in three-dimensional convolutional neural networks. ViT-B/16 has more parameters (86.8 million) and a much lower inference speed (1/30 fps).

3.2.2 Performance comparison on micro-nodule detection.

Table 2. Recall and AUC-ROC comparison for different sized nodules.

model	Micronodules (<5mm) Rec (%)	Micro nodule AUC-ROC	Medium nodule (5-10 mm) Rec (%)	Large nodules (>10mm) Rec (%)
LBP+SVM	48.2	0.652	76.8	89.5
HOG + random forest	51.3	0.683	79.2	91.2
2D ResNet-50	76.3	0.856	88.9	94.7
2D U-Net	78.5	0.872	89.6	95.3
3D ResNet-50	85.2	0.915	92.1	96.8
3D V-Net	87.6	0.928	93.4	97.5
ViT-B/16	82.4	0.897	90.8	95.9
3D Swin Transformer-T	89.7	0.936	94.2	98.1

The early-stage lung cancer contains most of the micronodules (diameter <5 millimeters). The detectability of micronodules decides the application value of computer computer-assisted diagnosis system. So, we also compared the performance of different method for different size nodules. We can find that three-dimensional has more advantages in the micronodule detection.

As shown in Table 2, compared with the traditional method, the traditional method can only reach a recall of less than 55% for the micronodule due to the difficulty of detecting the tiny structure difference of the micronodule by artificial feature. The two-dimensional method achieved the recall of 76.3%~78.5% for the micronodule, which is lower than the three-dimensional method, but obviously higher than traditional method. The three-dimensional method achieved a recall of 85.2%~89.7% for the micronodule. It should be noted that, compared with HOG+Random Forest, 3D Swin Transformer-T achieved a recall of 89.7% and AUC-ROC value of 0.936, which improved 38.4% in error rate compared with HOG+Random Forest model. This demonstrates the model's ability to effectively capture local texture and global position information of micronodules, thereby reducing the diagnostic error rate.

For large nodules (diameter > 10 mm), all models achieved recall rates above 89%. This is due to the more pronounced characteristics of large nodules. Therefore, the ability to detect micronodules proves to be the key indicator for distinguishing between model performances.

3.2.3 Analysis of small-sample robustness

Table 3. AUC-ROC and performance degradation of models at 10% training set.

model	10% training set AUC-ROC	100% training set AUC-ROC	decline(%)
LBP+SVM	0.621	0.765	18.8
HOG + random forest	0.645	0.792	18.6
2D ResNet-50	0.821	0.898	8.6
2D U-Net	0.837	0.905	7.5
3D ResNet-50	0.902	0.932	3.2
3D V-Net	0.915	0.941	2.8
ViT-B/16	0.835	0.918	9.0
3D Swin Transformer-T	0.908	0.948	4.2

CT data is scarce, and radiologists need more than 30 minutes to comment on each case, requiring models with strong small-sample processing capabilities. Table 1 illustrates the trend of AUC-ROC curves for different training set sizes, while Table 3 shows model performance using 10% of the training data.

The analysis shows that traditional methods are the least robust with small samples and experience a drop in AUC-ROC curves of more than 18% when the training data is reduced to 10%. This is because manual features are sensitive to data distribution, with a reduction in data volume significantly weakening the generalization capabilities of the features. 3D deep learning approaches show greater robustness. Even with only 10% of the training set, 3D V-Net achieves an AUC-ROC value of 0.915, which corresponds to a decrease of only 2.8%. This is because the superior ability of 3D features to capture volumetric structures in CT images, whose informational content far exceeds that of 2D features, and thus their dependence on data amount is reduced.

When there is a small sample, the three-dimensional Swin Transformer-T (which has only one-third of the parameters of 3D ViT-B/16 because of the pre-trained weights for medical images) greatly reduces the amount of data required, and therefore outperforms ViT-B/16 (the two amounts of AUC values are 0.908 and 0.835, respectively). The performance of the two-dimensional approach is between that of traditional methods and three-dimensional approaches. Compared with the two-dimensional ResNet-50, the two-dimensional U-Net is more robust because of the location information provided by the segmentation branch.

3.2.4 Computational efficiency comparison

Table 4. Computational Efficiency Comparison of Models

model	Params (M)	FPS	Single-sample inference time(ms)	memory usage (GB)
LBP+SVM	0.1	120	8.3	0.1(CPU)
HOG + random forest	0.2	105	9.5	0.1(CPU)
2D ResNet-50	25.6	85	11.8	1.2
2D U-Net	19.8	60	16.7	1.5
3D ResNet-50	31.2	45	22.2	3.8
3D V-Net	28.7	38	26.3	4.2
ViT-B/16	86.8	30	33.3	5.5
3D Swin Transformer-T	28.1	32	31.2	4.0

In terms of clinical application scenarios such as real-time screening and cloud application, computational efficiency is very important. For these three aspects, this study compared the computational efficiency of different models from three aspects: number of parameters, inference speed and GPU occupation rate.

According to the results of Table 4, the computational efficiency of traditional methods is the highest. Traditional methods can achieve more than 100 frames per second (FPS) without multiple CPUs and very little GPU memory. However, their accuracy cannot meet the accuracy requirements of clinical applications. The study shows that two-dimensional methods are better than three-dimensional models and Transformer models. Specifically, 2D ResNet-50 achieves 85 FPS with 1.2 GB of GPU memory. Its computational efficiency is very high, and it can be applied in scenarios that require high real-time performance, such as emergency CT. 3D Swin Transformer-T achieves the best balance between accuracy and efficiency. In terms of number of parameters, 3D ResNet-50 has 28.1 million, while ViT-B/16 has 86.8 million. 3D Swin Transformer-T has fewer parameters. With 32 FPS and 4.0 GB of GPU memory, this model can run on consumer GPUs. However, due to the

higher computing power of 3D V-Net, the frame rate of 3D V-Net is 38 FPS, which is lower than that of 3D ResNet-50 (45 FPS), and the memory consumption is 4.2 GB, which is higher than that of 3D ResNet-50 (4.0 GB). These characteristics make it suitable for scenarios where high accuracy of segmentation is required, such as measurement of nodule volume.

4. Conclusions

The test results demonstrate the mechanism and clinical application scenarios of the four methods for lung cancer detection. Traditional methods of manual feature extraction are characterized by superior computational efficiency and interpretability, making them particularly suitable for primary care settings with limited resources. However, this part still needs manual verification. Two-dimensional deep learning methods offer a compromise between parameter amount of moderate complexity and fast inference and can make up for the lack of Z-axis data to some extent by introducing multi-layer scan information. This method is suitable for real-time screening or emergency CT application scenarios, and it is very effective in detecting nodules larger than 10 millimeters in diameter. Three-dimensional deep learning models directly model three-dimensional characteristics and are characterized by excellent performance in detection of micro-nodules (< 5 millimeters), while demonstrating robustness under reduced sampling conditions. Therefore, this method is suitable for routine lung cancer screening and quantitative analysis (e.g., nodule volume measurement). The 3D Swin Transformer integrates three-dimensional spatial modeling and a global attention mechanism and can capture local and global contextual information simultaneously. Therefore, this model is the ideal choice for accurate detection or secondary confirmation of suspicious nodules with a good balance of accuracy and efficiency.

Key findings on the LUNA16 dataset show that 3D Swin Transformer-T achieves the highest overall performance. 3D deep learning models demonstrate strong robustness with limited annotated data. Concurrently, 2D ResNet-50 balances accuracy and speed (85 FPS), and traditional methods are limited to preliminary screening because of low micro-nodule detection rates (<55%).

Future work is expected to extend this work to multiple modal fusion (e.g., CT + PET-CT), semi-supervised learning to utilize unlabelled data, model lightweighting to be deployed on edge and further prospective clinical trials to explore the practical value of the CAD system on its workflow and diagnostic accuracy. This study provides an experimental basis for selecting an appropriate model in clinical CAD system and guides the practical application of computer vision in early lung cancer screening.

References

[1] Sung H, Ferlay J, Siegel R L, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 2021, 71(3): 209-249.

[2] Lyu J, Ling S H. Using multi-level convolutional neural network for classification of lung nodules on CT images. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2018: 686-689.

[3] Wan C, Ma L, Liu X, Fei B. Computer-aided classification of lung nodules on CT images with expert knowledge. In: Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling, Vol. 11598, 2021 Feb: 673-678. SPIE.

[4] Arenberg D. Micronodules detected on lung cancer screening CT scans. Journal of Thoracic Oncology, 2019, 14(9): 1501-1503.

[5] Ma W B, Yang Y, Fang W C. An effective tuberculosis detection system based on improved faster R-CNN with ROI align method. In: 2023 IEEE Biomedical Circuits and Systems Conference (BioCAS), 2023 Oct: 1-5. IEEE.

[6] Setio A A A, Traverso A, De Bel T, Berens M S, Van Den Bogaard C, Cerello P, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medical Image Analysis, 2017, 42: 1-13.

[7] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.

[8] Jaeger S, Candemir S, Antani S, et al. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quantitative Imaging in Medicine and Surgery, 2014, 4(6): 475.

[9] Dragon J M, Guha S, Salvatore M M. Hounsfield units: future applications in clinical practice, radiomics, and artificial intelligence. Clinical Imaging, 2024, 110: 110141. DOI: 10.1016/J.CLINIMAG.2024.110141.

[10] Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning. Journal of Big Data, 2019, 6(1): 1-48.

[11] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015 Oct: 234-241. Cham: Springer International Publishing.

[12] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[13] Milletari F, Navab N, Ahmadi S A. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), 2016 Oct: 565-571. IEEE.

[14] Dosovitskiy A. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint, 2020. arXiv:2010.11929.