Proceedings of the 42nd Chinese Control Conference
July 24-26, 2023, Tianjin, China

A Lung Lesion Detection Algorithm Based on YOLOv7 and Self-attention Mechanism

Junhua Luo, Shujing Wang*, Qixiang Wang, Shaojun Liu

School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, P. R. China
E-mail: wshujing@shu.edu.cn

Abstract: In recent years, non-invasive medical imaging techniques such as X-ray have become increasingly important in medical diagnosis. However, traditional lesion detection algorithms on chest X-ray using neural deep learning techniques often have low detection rates. To address this issue, we propose a new chest X-ray lesion detection algorithm based on the YOLOv7 object detection algorithms, which incorporates with self-attention mechanism module to enhance the detection ability and reduce computation time. Our algorithm achieved the best results on the VinDr-CXR dataset of 640×640 images, using the YOLOv7-X and Swin Transformer module, with a mAP@0.5 of 41.13% and 17fps detection speed. These results demonstrate that our algorithm achieves high detection accuracy while maintaining fast detection speed, making it a promising approach for chest X-ray lesion detection.

Key Words: Lesion detection, chest X-ray, YOLOv7, Swin Transformer, medical imaging, computer aided diagnosis

1 Introduction

Lung health is closely related to human life. Lung cancer is one of the most influential cancers in the world. According to statistics, the number of lung cancer cases worldwide will exceed 2,000,000 and the incidence rate is about 22.4 per 100,000 in 2020. Among them, more than 1,000,000 people died of lung cancer, which is the leading cause of death from cancer. Furthermore, since 2019 the outbreak of pneumonia caused by COVID-19 has swept the world, with a total of more than 500 million confirmed cases. The study of lung lesions has become a hot spot at present.

Medical imaging is widely used in auxiliary diagnosis when detecting lung diseases, and non-invasive medical imaging technologies such as X-ray, MRI, and CT have attracted more and more attention from researchers. Among them, X-ray imaging technology has the advantages of being simply operational devices, easy to read, basically insensitive to the person being photographed, low cost, and has been widely used in hospitals and medical places. The imaging principle is to use the X-ray to penetrate the object, the attenuation coefficient of the penetrated material and the thickness are different, and the intensity value after penetrating the material is recorded. Then manifested by the grayscale difference in the image. Reading X-ray films has high requirements on the experience and ability of the reviewers, and repeated monotonous reading is likely to cause visual fatigue and lead to misdiagnosed lesions. Therefore, the study of an efficient Computer Aided Diagnosis (CAD) system can reduce the pressure of manual film reading and subjective diagnosis errors for doctors, and assist doctors in accurate and rapid diagnosis of diseases, also reducing the rate of misdiagnosis in clinical diagnosis.

In this paper, we propose a novel CAD algorithm based on YOLOv7 and self-attention mechanisms for detecting lung lesions on chest X-ray images. Our algorithm achieves high detection accuracy and fast detection speed, offering an efficient and reliable solution for assisting physicians in diagnosing lung lesions. With the growing prevalence of lung cancer and the recent COVID-19 pandemic, our algorithm has the potential to make a significant impact in the field of medical imaging and improve patient outcomes.

2 Related Works

With the rapid development of IT technology, research results of medical image analysis using deep learning have been published frequently in recent years. [1] explored the interpretability of deep neural networks in medical imaging technology. [2] Use the YOLOv4 network with the residual structure to distinguish COVID-19 and common pneumonia. [3][4] and [5] respectively used scalable attention residual CNN(SAR-CNN), EfficientDet, YOLOv5, and other algorithms to detect lung lesions using chest X-ray films. [6] summarize commonly used chest X-ray datasets and the literature using these datasets.

It should be noted that many previous studies in the

Code is available on https://github.com/DF4D1999/A-Lung-Lesion-Detection-Algorithm-Based-on-YOLOv7-and-Self-attention-Mechanism.

* Corresponding author.

field of chest X-ray image analysis have utilized outdated object detection architectures and have not taken advantage of newer, more advanced models that can provide superior performance for identifying and locating various types of lesions. Furthermore, most of these studies have focused primarily on detecting and classifying cases of pneumonia, with limited research conducted on the classification and localization of a broader range of lung lesions.

3 Proposed Framework

Object detection algorithms based on deep learning techniques are commonly used to locate and classify objects within images. These algorithms can be categorized into two groups: one-step algorithms and two-step algorithms. Two-step algorithms begin by generating candidate regions, which are then analyzed for object detection. Examples of such algorithms include the R-CNN series. In contrast, one-step algorithms directly predict the presence and position of objects within the input image, resulting in faster detection speed. Examples of one-step algorithms include the YOLO series, SSD algorithm, and FCOS algorithm.

3.1 YOLOv7

The YOLO series [7] object detection algorithms generate multiple anchors to an image through the k-means clustering algorithm. The algorithm first uses the convolutional neural network to extract features from the input image and then makes predictions for each anchor. If an object is detected, the location offset between the anchor center and the object is calculated, generating multiple bounding boxes. Then go through the non-maximum suppression processing, calculate the highest confidence score bounding box and output the result.

The overall structure of YOLOv7[8] is like YOLOv5, with "bag-of-freebies". The main improvement is the addition of ELAN (Extended Efficient Layer Aggregation Networks) and Max Pooling modules to enhance the network learning ability, as Fig. 1 shows. Also, the re-parameterization structure [9] is used in the detection head to improve the detection effect on small objects. The loss function has also been improved.

The loss function of YOLOv7 consists of three parts: classification loss, localization loss, and confidence loss, the calculation of the loss function is shown in formula (1).

$$\text{Loss}_{total} = \text{Loss}_{Class} + \text{Loss}_{Local} + \text{Loss}_{Confidence} \quad (1)$$

The confidence loss and classification loss are the binary cross entropy with logits loss, as shown in formula (2). y stands for the algorithm prediction and $\hat{y}$ stands for the ground truth.

$$\text{Loss} = Sigmoid\{-\frac{1}{n}\sum[y\, ln\hat{y} + (1-y)\ln(1-\hat{y})]\} \quad (2)$$

YOLOv7 uses CIoU (Complete-IoU) [10] to calculate localization loss as shown in formula (3) (4) (5).

$$\text{Loss}_{Local} = 1 - \text{IoU} + \frac{\rho^2(b,b^{gt})}{c^2} + \alpha V \quad (3)$$

$$\alpha = \begin{cases} 0, & IoU < 0.5 \\ \frac{V}{(1-IoU)+V}, & IoU \geq 0.5 \end{cases} \quad (4)$$

$$\text{V} = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} + arctan\frac{w}{h})^2 \quad (5)$$

IoU stands for the Intersection over Union ratio between the ground truth bounding box and the predicted bounding box. ρ^2 stand for the Euclidean distance of the center points between two boxes. c stands for the diagonal length of the smallest rectangle containing two boxes. CIoU loss function considers more aspects than normal IoU, including overlapping area, center point distance, and aspect ratio, so the regression of the network could be more accurate and faster.

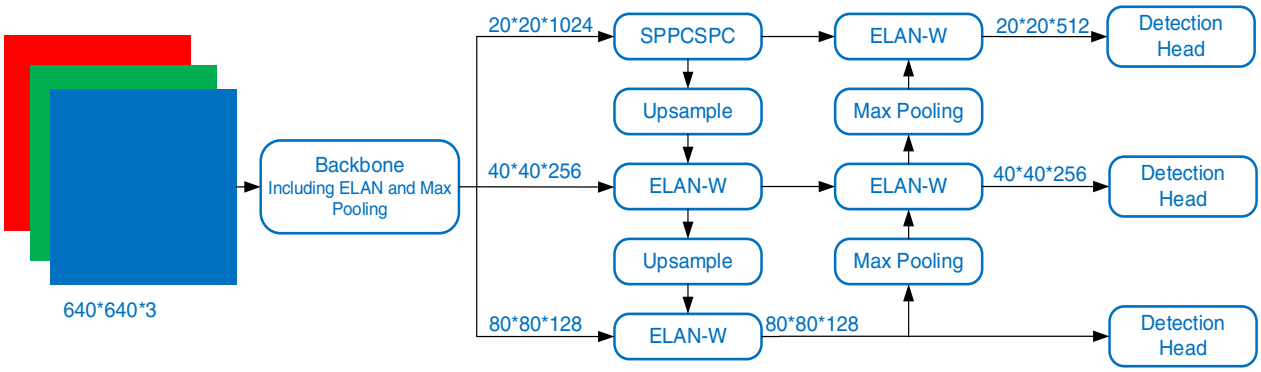

Fig. 1: The overall architecture of YOLOv7

3.2 Swin Transformer

Swin Transformer was proposed by Microsoft [11] and is an improved algorithm based on Vision Transformer (ViT) [12]. In ViT, the image is divided into multiple patches to form a sequence and then inputs the attention module, the computational complexity is the square of the image scale, resulting in an excessively large amount of calculation. Swin Transformer divides the image into multiple patches using a sliding window, first calculates the attention inside a sliding window, then performs cross-window connection to calculate the attention, shown in Fig. 2.

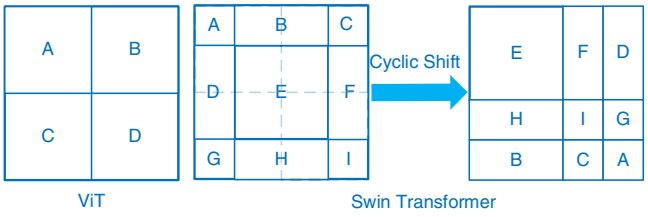

Fig. 2: ViT and Swin Transformer self-attention calculation

Swin Transformer reduces the number of convolutional layers while reducing the number of calculations, with high operational efficiency. Swin Transformer can be flexibly modeled at various scales and has linear computational complexity related to image size. Swin Transformer is therefore compatible with a wide range of computer vision tasks and has achieved good results in object detection and instance segmentation field.

The object detection algorithm can also be used for the detection of chest X-ray images, so this paper proposes a chest X-ray lesion detection algorithm based on YOLO and the self-attention mechanism. The first convolution layer of the ELAN structure in YOLOv7 was replaced by the Swin Transformer module to improve the detection ability of the algorithm and reduce the time required for calculation. The ELAN structure overall structure as Fig. 3 shows.

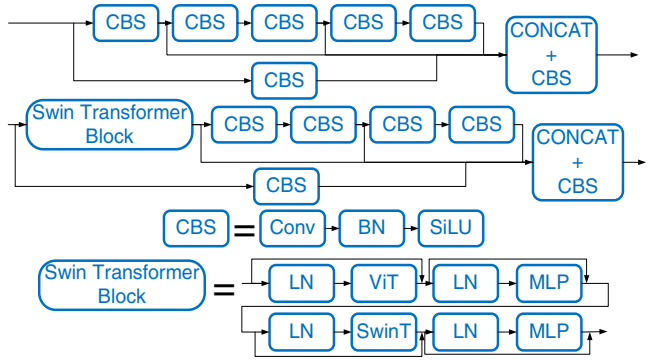

Fig. 3: The architecture of the original ELAN and after embedding the Swin Transformer

The overall procedure from X-ray shooting to the output result is shown in Fig. 4.

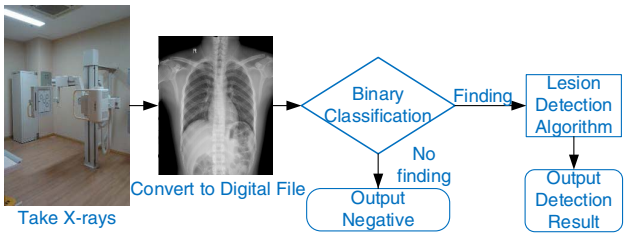

Fig. 4: Overall procedure

4 Experiment

4.1 Dataset

In this paper, we utilize the VinDr-CXR dataset, which was curated by VinBigData, a prominent big data research institute in Vietnam. The dataset comprises 18,000 chest X-ray images collected from two hospitals in Vietnam between 2018 and 2020, with 15,000 images reserved for training and 3,000 for testing [13]. The algorithm under evaluation was trained using the entire set of 15,000 training samples, which can be accessed on Kaggle [14]. It is important to note that each sample in the dataset may contain multiple annotations, as indicated in Table 1.

In the VinDr-CXR dataset, each X-ray sample was evaluated by three physicians to mark the corresponding lesion bounding box. To account for this, the multiple annotations for the same lesion were averaged during data preprocessing, a technique referred to as Weighted Boxes Fusion (WBF) [15]. Fig. 5 illustrates a comparison between the original annotation bounding boxes and those obtained after applying the WBF technique.

Table 1. The Distribution of Labels in the VinDr-CXR Dataset

Lesion Type	Number
No Finding	31818
Aortic Enlargement	7162
Cardiomegaly	5427
Pleural Thickening	4842
Pulmonary Fibrosis	4655
Nodule/Mass	2580
Lung Opacity	2483
Pleural Effusion	2476
Other Lesion	2203
Infiltration	1247
Interstitial Lung Disease (ILD)	1000
Calcification	960
Consolidation	556
Atelectasis	279
Pneumothorax	266

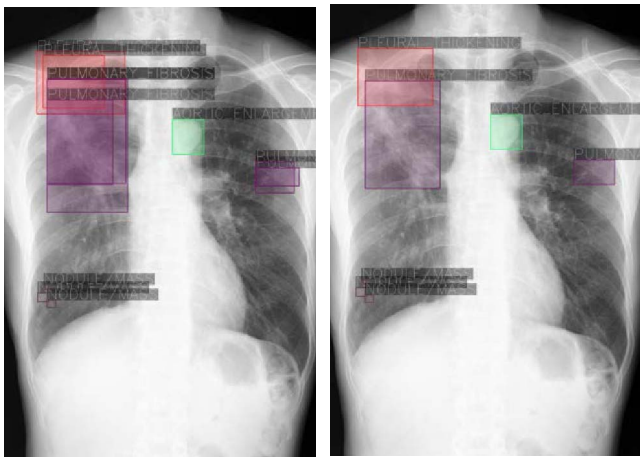

Fig. 5: Original annotation box and after WBF box

Table 1 shows the dataset contains many samples (exactly 31818÷3=10606 samples) with no-finding labels. The usual practice is to add a binary classification network and let the samples identified as not having no-finding labels enter the object detection algorithm [4]. Here we omit the binary classification network and just remove the no-finding samples. Then the remaining 4,394 samples were randomly shuffled and divided into training set: testing set = 3:1 to evaluate the algorithm.

4.2 Device

A high-performance Dell R730XD server is used to train and test the algorithm. The hardware configuration is shown in Table 2. Due to the benefit of huge memory, we use the *--cache-images* option caching data into memory to speed up training (about 10% faster).

Table 2. Hardware Platform Configuration

CPU	Intel Xeon E5-2698v3*2
RAM	DDR4 128GB 2133MHz
GPU	Nvidia Tesla P40 24G
Programing Language	Python 3.9.13
Framework	PyTorch 1.31.0 + YOLOv7 v0.1

4.3 Training configuration

For better algorithm performance, avoiding overfitting and time-saving considerations, we set the training epoch to 300 and use the YOLOv7 default hyperparameter.

We use mAP as the evaluation index to measure the algorithm. Mean average precision at 50% or mAP@0.5 is used to measure the average accuracy when IoU is 50%, and mAP@0.5:0.95 refers to IoUs at 50% to 95%, with a step of 5%, then calculate the mean of average accuracy on those IoUs. The curve of box loss, objectness loss, classification loss, precision, and recall of training set and testing set during the training process is shown in Fig. 6.

During the training process of the algorithms, the box loss, objectness loss, and classification loss are gradually reduced for both the training and testing sets. By around 200 epochs, the loss values on the testing set become relatively stable, while on the training set, they continue to decline due to overfitting. At the same time, precision and recall values gradually increase, indicating improved accuracy in object detection. After 200 epochs, the changes in each evaluation index tend to stabilize. Fig. 7 presents examples of ground truth annotation boxes and predicted boxes with confidence scores generated by the algorithm for testing samples after 300 epochs of training.

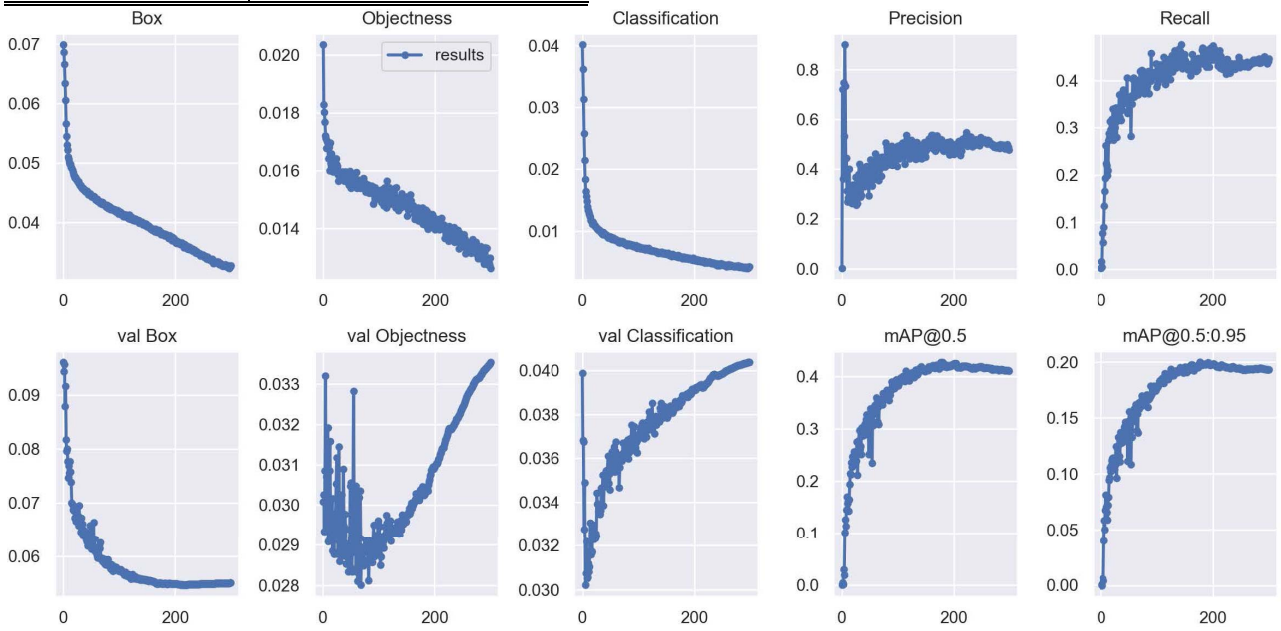

Fig. 6: The curve of box loss, objectness loss, classification loss, precision, and recall

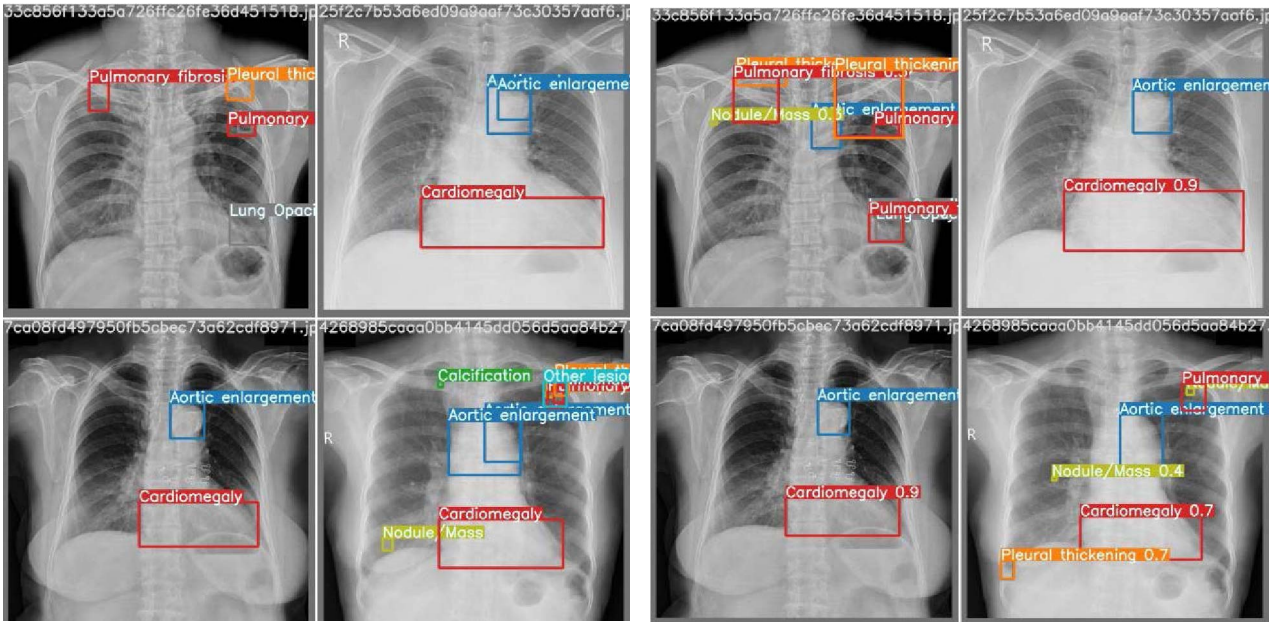

Fig. 7: Some testing samples' grand truth annotation box and algorithm predict box with the confidence score

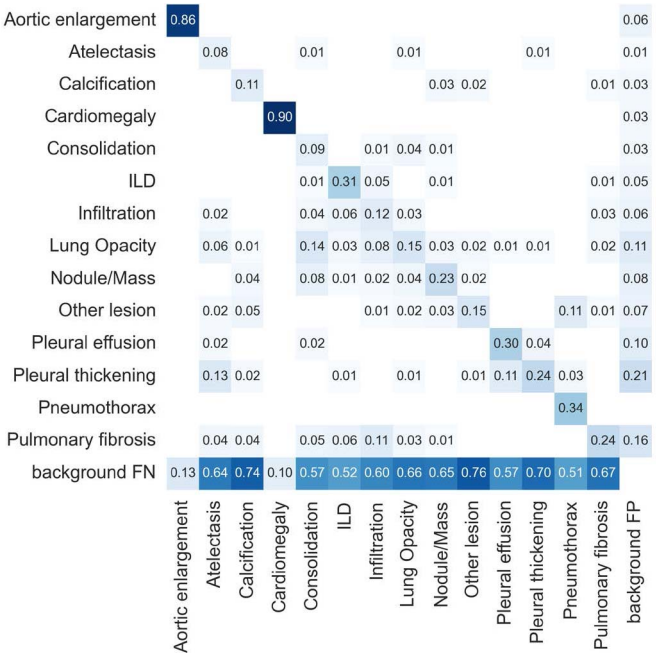

Fig. 8: The confusion matrix of the test set

The confusion matrix of the testing set is shown in Fig. 8. The column is the algorithms predict, the line is ground true.

Background FP (the last line) stands for negative but is predicted into positive by the algorithm, also known as negative is falsely checked. Background FN (the last column) stands for positive but is predicted as negative, indicated the non-background object is missed.

Table 3 compares the number of parameters, mAP@.5, mAP@.5:.95:.05 and the inference speed at batch size = 1 with other algorithms on VinDr-CXR dataset. The proposed algorithm in this paper has reached a higher level of object detection and maintains fast detection speed.

Table 3. Compare with Other Algorithms

Algorithm	Params (M)	mAP@.5	mAP@.5:.95:.05	Inference Speed (ms)
SAR-CNN [3]	56.48	0.1575(512 × 512)	-	384.6
FasterRCNN+YOLOv5+EfficientDet-D7 [4]	-	0.292(mAP@.4)	-	250
YOLOv5s [5]	7.2	-	0.162	-
EfficientDet+YOLOv5s [16]	-	0.200(mAP@.4)	0.1431	-
RRCNet [17]	41.28	0.1617(mAP@.4)	-	142.7
YOLOv5+Faster RCNN+FPN+2 class filter [18]	-	0.272(mAP@.4)	-	-
Ours (Vanilla YOLOv7)	36.6	0.4027	0.1883	30.2
Ours (YOLOv7 + Swin Transformer)	43.3	0.3862	0.179	39.9
Ours (YOLOv7-X)	70.9	0.3914	0.1857	48.5
Ours (YOLOv7-X + Swin Transformer)	77.6	**0.4113**	**0.193**	58.7

5 Conclusion

The algorithm proposed in this paper, which employs YOLOv7 and a self-attention mechanism for detecting lung lesions in chest X-rays, has demonstrated remarkable effectiveness. Our approach achieved the highest mean average precision (mAP) at a threshold of 0.5, with a value of 41.14%, and the highest mAP over a range of thresholds from 0.5 to 0.95 with an interval of 0.05, with a value of 19.3% when utilizing the YOLOv7-X and Swin Transformer module. Furthermore, the proposed algorithm was able to maintain a high level of accuracy while achieving a detection speed of up to 17fps. These results demonstrate that our algorithm can provide a valuable medical reference for doctors in their clinical decision-making.

References

[1] Van der Velden, Bas HM, et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis." *Medical Image Analysis* (2022): 102470.

[2] Kumar, Akhil. "RYOLO v4-tiny: A deep learning based detector for detection of COVID and Non-COVID Pneumonia in CT scans and X-RAY images." *Optik* 268 (2022): 169786.

[3] Lin, Cong, et al. "Lesion detection of chest X-Ray based on scalable attention residual CNN." *Mathematical Biosciences and Engineering* 20.2 (2023): 1730-1749.

[4] Pham, Van-Tien, et al. "Chest x-ray abnormalities localization via ensemble of deep convolutional neural networks." *2021 International Conference on Advanced Technologies for Communications (ATC)*. IEEE, 2021.

[5] Fan, WenZe, et al. "Research on Abnormal Target Detection Method in Chest Radiograph Based on YOLO v5 Algorithm." *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*. IEEE, 2021.

[6] Çallı, Erdi, et al. "Deep learning for chest X-ray analysis: A survey." *Medical Image Analysis* 72 (2021): 102125.

[7] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[8] Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." *arXiv preprint arXiv: 2207.02696* (2022).

[9] Ding, Xiaohan, et al. "Repvgg: Making vgg-style convnets great again." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[10] Zheng, Zhaohui, et al. "Enhancing geometric factors in model learning and inference for object detection and instance segmentation." *IEEE Transactions on Cybernetics* 52.8 (2021): 8574-8586.

[11] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[12] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *International Conference on Learning Representations*.

[13] Nguyen, Ha Q., et al. "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations." *Scientific Data* 9.1 (2022): 429.

[14] DungNB, Ha Q. Nguyen, Julia Elliott, KeepLearning, NguyenThanhNhan, Phil Culliton. (2020). VinBigData Chest X-ray Abnormalities Detection. Kaggle. https://kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection

[15] Solovyev, Roman, Weimin Wang, and Tatiana Gabruseva. "Weighted boxes fusion: Ensembling boxes from different object detection models." *Image and Vision Computing* 107 (2021): 104117.

[16] Le, Khiem H., et al. "Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis." *IEEE Access* (2023).

[17] Yang, Xinquan, et al. "A Coarse Feature Reuse Deep Neural Network for CXR Lesion Detection." *2021 11th International Conference on Information Technology in Medicine and Education (ITME)*. IEEE, 2021.

[18] Rastogi, Akarsh, et al. "Real time Chest X-ray Pathology detection and localization framework with Convolutional Neural Networks and Ensembling." *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. IEEE, 2022.