Location Fusion and Data Augmentation for Thoracic Abnormalites Detection in Chest X-Ray Images

Nguyen Thi Van Anh[1], Nguyen Duc Dung[2], Nguyen Thi Phuong Thuy[1]

[1] Institute of Physics, VAST, 18 Hoang Quoc Viet, Nghia Do, Cau Giay, Hanoi, Vietnam
[2] Institute of Information & Technology, VAST, 18 Hoang Quoc Viet, Nghia Do, Cau Giay, Hanoi, Vietnam

Correspondence: Nguyen T. Van Anh, anhntv@iop.vast.vn
Communication: received 05 Dec 2022, revised 16 Dec 2022, accepted 01 Feb 2023
DOI: 10.32913/mic-ict-research.v2023.n1.1172

***Abstract*: The application of deep learning in medical image diagnosis has been widely studied recently. Unlike general objects, thoracic abnormalities in chest X-ray radiographs are much harder to label consistently by domain experts. The problem's difficulty and inconsistency in data labeling lead to the downgraded performance of the robust deep learning models. This paper presents two methods to improve the accuracy of thoracic abnormalities detection in chest X-ray images. The first method is to fuse the locations of the same abnormality marked differently by radiologists. The second method is applying mosaic data augmentation in the training process to enrich the training data. Experiments on the VinDr-CXR chest X-ray data show that combining the two methods helps improve the predictive performance by up to 8% for F1-score and 9% for the mean average precision (MAP) score.**

Keywords: *chest X-ray image, thoracic abnormalities, deep learning, object detection*

I. INTRODUCTION

In the last decade, deep learning has been widely applied in computer-aided medical diagnosis (CAD) systems [1]. The main task of deep learning models is detecting or classifying abnormalities or diseases in medical images such as X-rays, MRI, endoscopy, and ultrasound. CAD systems assist in the diagnosis process for early disease detection or to support telemedicine systems [2–4]. Deep learning techniques have been intensively studied for COVID-19 detection using chest X-ray CT scan images, e.g., [5].

Difference from the detection of radiological experts who use domain knowledge and experience to examine the images, deep learning (DL) methods require many images with carefully annotated labels to guide the training process of AI models. Unfortunately, the most robust deep learning algorithms could not learn by themselves or learn effectively in many domains, and of course, in medical imaging. The standard process in building CAD systems using deep learning methods is to collect medical images and then ask the doctor to annotate the images, usually to mark the abnormal regions and assign the disease label to each abnormal region.

Compared with other domains like image object detection in the wild or street image segmentation, medical image diagnosis using deep learning faces more difficulties. The first challenge is data collection, where the privacy and authority of assessing a large volume of patient data are very sensitive. The second challenge is the manually annotated data. For the natural image, the annotation process could be done efficiently, e.g., using a crowd-sourcing approach. The annotation of medical images requires the time and effort of domain experts with deep knowledge and years of experience. Moreover, the third one, it is difficult to correctly identify abnormalities and disease types using only image observation, e.g., from X-ray or ultrasound images. At least, it is much more complex than identifying a bicycle or a car in a natural scene. Due to the natural difficulty of the problem, the annotation disagreement among experts (doctors) also negatively affects the performance of deep learning models. The label confusion in medical image data includes the abnormality's location and the disease type.

In this paper, we investigate and introduce two data processing techniques to improve deep learning performance on lung lesion detection and classification from X-ray images. Firstly, two box fusion methods have been tried to process the location disagreement among annotations of domain experts. According to our statistics (Table I), up to 39.8% of annotations in the VinDrCXR are inconsistent in lung pathologies or location of the abnormality (as illustrated in Figure 1). Our experiments show that weighted box

fusion (WBF), a technique commonly used in the prediction phase, has an advantage over non-maximum suppression (NMS). WBF helps improve the F1 predictive performance by up to 5% on the VinDr-CXR dataset. To overcome the limitation of labeled data, mosaic image augmentation was applied to the training process. Its combination with the label WBF helped improve lung disease detection by up to 8% in F1 score and 0.9% in MAP, a significant result in a very challenging problem.

The remaining part of this paper is organized as follows. Section II presents the challenges and related works in thoracic anomaly detection from chest X-ray images. Section III introduces two methods for dealing with the differences in data labeling and augmentation techniques for improving the efficiency of the training phase of deep learning. Section IV presents experiments and an analysis of the effectiveness of the two mentioned methods. Some conclusions will be summarised in the last section.

II. ABNORMALY DETECTION IN CHEST X-RAY IMAGES

A chest X-ray image, which was born by a small dose of ionizing radiation, is used to help in the diagnosis of some lung diseases such as chest pain, cough, and pneumonia. It is suitable for emergency treatment and diagnosis thanks to its low cost and simplicity. With the development of deep learning, several works have tried to apply general image object detection models to the chest X-ray image domain [6, 7]. Different datasets have been built for training these deep-learning models. For example, the ChesXpert dataset [8] consists of 224,316 chest radiographs. They labeled 14 observations as positive, negative, or uncertain for output. The NIH [9] dataset consists of 108,948 frontal view X-ray images with eight lung pathologies. They used deep learning models to develop the labeler and compared it to the prediction of radiological experts. In addition, the VinDr-CXR [10] dataset was first proposed on Kaggle with 15,000 chest x-ray images of the training folder having 10,606 images of normalities and 4,394 images of lung abnormalities. In 4,394 images of lung abnormalities [10], 17 radiological experts annotated the lung lesion areas with 36.096 rectangular boxes corresponding to the labels of 14 lung pathologies.

Although chest X-ray images are popular, annotating the data is time-consuming and expensive. Many datasets were automated automatically, e.g., using keyword matching or an NLP model [8], [9]. However, the automatic annotation usually needs to be clarified and corrected manually. In the VinDr-CXR dataset, each image was annotated by three radiologists. The multiple annotations of the same image sometimes bring inconsistency in the anomaly or disease label. There are two types of consistency, as shown in Figure 1. The first type is the same region annotated with a different type of disease (Figure 1 (a)) or the same disease but in a different location (Figure 1 (b)). The inconsistency in data annotation causes a negative effect on the performance of deep learning models.

In this work, we focus on solving the problem of location annotation differences and show that a suitable method for fusing them could help improve the accuracy of thoracic abnormalities detection in chest X-ray images.

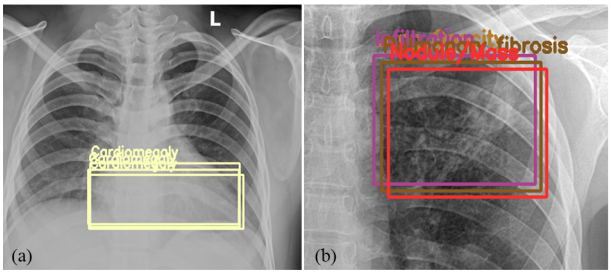


Figure 1. Due to the difficulty of the problem, the same region in an X-ray image could be annotated differently: (a) the same class of disease but different bounding boxes or (b) the different classes and different bounding boxes.

III. LOCATION FUSION AND DATA AUGMENTATION FOR CHEST ABNORMALITIES DETECTION FROM X-RAY IMAGES

1. Box Fusion algorithm

There are several techniques to solve the problem of combining different bounding boxes of the same labels in object detection. For example, we can group the bounding boxes having approximate coordinates, keep one and eliminate the rest, or use a suppression algorithm to suppress the bounding boxes under a predefined threshold of IOU (intersection of a union) score [11]. For the lung X-ray images, the locations of lesion areas are not easy to find, so the bounding box obtained using elimination or suppression is not suitable. Besides, reducing bounding boxes by elimination and suppression negatively impacts the model because it eliminates valuable information in the training phase of deep learning models.

Therefore, we propose to apply the weighted box fusion (WBF) algorithm [12], which is popularly used to ensemble the predicted bounding boxes of different object detection models for better results. In our case, the WBF technique helps produce more accurate bounding boxes and labels for the training data to observe and have a better decision in supervised learning during the training process and then improve the model efficiency. It does not eliminate any bounding box but harmonizes the annotations, so it could help improve the efficiency of the trained model.

The WBF algorithm worked as follows:

At first, the bounding boxes were sorted in descending order of experts (considered as confidence score). Then, a list of possible combinations of bounding boxes was created and checked if any fusions matched the initial sorted ones in cycle iteration. This procedure performed through checking the IOU between the fusion and original bounding boxes. If the IOU is larger than a specific threshold (IOU = 0.55 in their paper), then a formula was used to adjust the coordinates and confidence scores of all the boxes in the fused boxes list. The fusing calculation was described in Figure 2.

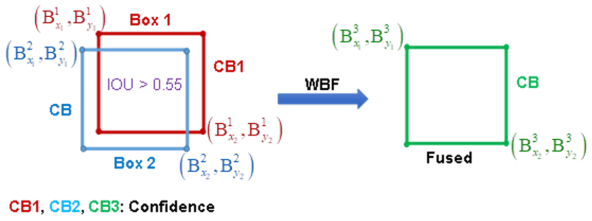


Figure 2. Two boxes were fused into one box by the WBF algorithm.

The coordinates of the fused box were weighted sums of the coordinates of the boxes that form it and were calculated by the formula:

$$B^3_{x_1} = \frac{B^1_{x_1} \times CB1 + B^2_{x_1} \times CB2}{CB1 + CB2}, \quad (1)$$

$$B^3_{x_2} = \frac{B^1_{x_2} \times CB1 + B^2_{x_2} \times CB2}{CB1 + CB2}, \quad (2)$$

$$B^3_{y_1} = \frac{B^1_{y_1} \times CB1 + B^2_{y_1} \times CB2}{CB1 + CB2}, \quad (3)$$

$$B^3_{y_2} = \frac{B^1_{y_2 1} \times CB1 + B^2_{y_2} \times CB2}{CB1 + CB2}. \quad (4)$$

In which, the $CB1$, $CB2$ and $CB3$ were confidence scores for the corresponding boxes.

The new confidence score for the fused box was simply the average confidence of all boxes that form it and was computed by the formula:

$$CB3 = \frac{CB1 + CB2}{2}. \quad (5)$$

An example of WBF implementation is shown in Figure 3.

Figure 3(a) showed the images had two lesion areas, and their labels had three overlapped bounding boxes that became one box after applying the WBF technique in Figure 3(b). The WBF algorithm performs better than other ensemble algorithms and produce better data for the model, the results will be presented in the experiment part. The

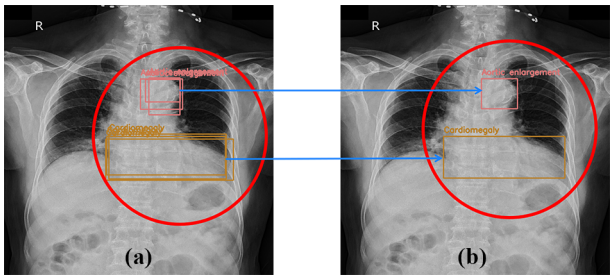


Figure 3. An implementation of the WBF algorithm on an original chest x-ray image in a dataset. (a) the original one presented two lesion areas annotated by overlapped bounding boxes, and (b) the one described two lesion regions without overlapped bounding boxes after applying the WBF algorithm.

main parameter of the WBF technique is the IOU threshold that is inverse proportion to the number of merged boxes, i.e., a large number of boxes are merged with decreasing IOU. So, we have to investigate the performance of the model with different IOU values to find the most suitable IOU in conformity with each dataset.

The comparison to other methods we showed in the experiment section. In addition, the chosen IOU thresholds in the WBF are crucial because the number of bounding boxes also affected the performance of the training model. Therefore, we first investigated some IOU thresholds (0.4, 0.5, 0.6, and 0.7) for the dataset and finally found out that the IOU threshold of 0.5 is the best for the number of bounding boxes (see in the experiment section).

2. Data augmentation

Training deep learning models requires a massive volume of data to ensure the well-generalized ability of the trained model. Ideally, the training data should cover all possible variations in reality. In practice, such a requirement is impossible for both labeled and unlabeled data. One of the most popular techniques for overcoming this limitation is to apply data augmentation techniques in the training process. Data augmentation uses image processing methods to generate labeled data increase in size and diversity. The augmentation methods consist of photometric distortion and geometric distortion. In photometric distortion methods, new images are created by adapting brightness, hue, contrast saturation, and noise. In contrast, in geometric distortion methods, the images are generated by adjusting rotation, flipping, random scaling, and cropping [13]. Besides, more complex image augmentations are used for suitable objects to make the data more diverse as Mixup, CutMix, or Mosaic [14]. The data augmentation not only improves the generalization ability of the trained model but also works as a regularization technique to help reduce overfitting in the training models [15].

Chest X-ray images are in greyscale with two monochromatic color spaces, one is the lowest value of a pixel and interpreted as white, and the other is as black. Moreover, a lung pathology in chest x-ray images was annotated by bounding boxes in different sizes and unclear distribution of its location. In this work, we first adapted the hyper-parameter of hue-saturation-value, horizontal flip, and scale, then continued to tune the mosaic hyper-parameter to achieve the best performance possible. Illustrations for data augmentations used in this study, were shown in Figure 4. In section IV, we conduct different experiments to examine the effectiveness of data augmentation techniques for the detection of lung pathologies in chest X-ray images.

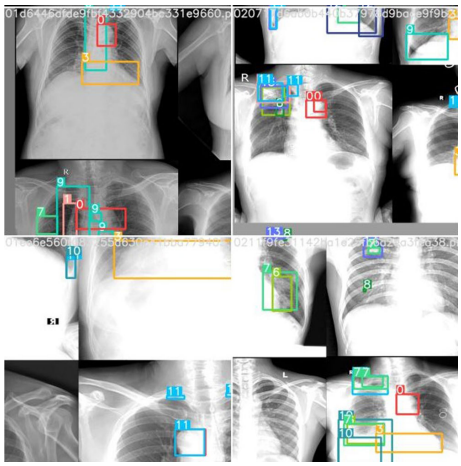


Figure 4. An example of chest x-ray images was applied mosaic technique.

3. Object detection model

There are two branches of deep learning models for general object detection tasks. The first one is two-stage detectors which conduct the first stage of region proposal generation, followed by the second stage of object classification and bounding box regression such as R-CNN, Fast R-CNN, and Faster R-CNN [1]. This method generally shows a high accuracy but has a disadvantage of a slow detection speed and lower efficiency. The second one is one-stage detectors, which extract object classification and bounding box regression simultaneously without a region proposal stage as the Yolov1, Yolov2, and SSD models [1]. This method generally has fast detection speed and high efficiency but low accuracy. Different improvements have been proposed to improve the basic models like Yolov3, Yolov4 [1] and EfficientDet [17] to achieve better and better model performance. The ResNet-50 was reported as the best model for evaluating multi-label chest X-ray image classification [9]. In the Yolov3 architecture, the feature extractor (Darknet53) consists of a convolution neural network and five residual blocks that give more powerful feature extraction for the phase of multi-labels image classification. The head of the detector took inspiration from Feature Pyramid Network [18] architecture which designed three scales for the prediction in three different resolutions. Additionally, Yolov3 uses many other helpful layers, such as con-cat, rout, and skip, to pass related feature maps to the head of the detector. It was an actual powerful model with a total of 106 layers. The feature maps are stabilized by a Spatial Pyramid Pooling (SPP) module [16] that was set up in the Yolov3 architect before sending to the head detector. The structure of the Yolov3 model with the SPP module [16] shows in Figure 5.

The model was fast in the extreme and saw the entire image during training and test time, so it implicitly encoded contextual information about classes and their appearance. The algorithm outperformed other top detection methods and resulted in big public datasets such as Coco and Pascal VOC [19]. It presented an impressive performance as a premier algorithm in terms of speed and accuracy for detecting and determining object coordinates compared to the other models at that time. For the above reasons, we use Yolov3 with SPP module [16] (Yolov3-spp) as a baseline for the experiments on the chest X-ray dataset. The obtained results in this experiment may be the same or better in other object detection models depending on the thorough investigation.

IV. EXPERIMENT

In this section, we evaluate the effectiveness of the WBF algorithm in fusing locations of lungs abnormally marked by domain experts on the same image data and the effectiveness of the data augmentation technique in this domain. We used the Yolov3-spp model described above and conducted experiments on the VinDr-CXR chest X-ray dataset. The dataset is divided into three sets with ratios of 80%, 10%, and 10% for training, validation, and testing, respectively. In the following sections, we describe the experimental setup and the results in detail.

1. Dataset

In the public dataset, the VinDr-CXR [10] comprises 4.394 images of lung abnormalities annotated by 17 radiological experts. Fourteen annotated lung diseases are Aortic enlargement, Cardiomegaly, Atelectasis, Pneumothorax, Consolidation, Infiltration, Lung Opacity, Pleural effusion, Interstitial lung disease, Pulmonary fibrosis, Pleural thickening, Nodule/Mass, Calcification, Other. The remaining images of lung normalities can be used in training on other deep-learning models. The information of bounding boxes

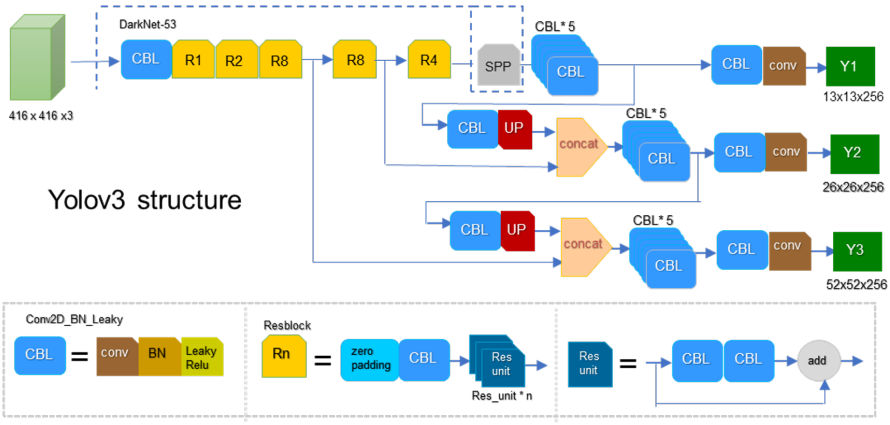


Figure 5. The structure of Yolov3 with a spatial pyramid pooling (SPP) module [16].

TABLE I
THE BOUNDING BOXES ON THE LESION AREAS OF 14 LUNG PATHOLOGIES IN THE VINDR-CXR DATASET.

Class names	Number of bounding boxes	Number of overlapped bounding boxes of the different labels	Number of overlapped bounding boxes of the same labels
Peumothorax	226	16	95
Atelectasis	279	194	47
Consolidation	556	372	121
Calcification	960	340	215
ILD	1000	300	272
Infiltration	1247	703	293
Other lesion	2203	209	367
Pleural effusion	2476	1300	708
Lung Opacity	2483	713	466
Nodule/Mass	2580	259	713
Pulmonary fibrosis	4655	1302	1291
Pleural thickening	4842	1488	788
Cardiomegaly	5427	2	3031
Aortic enlargement	7162	23	3785
Total	36,096	7221	12,192

and labels are provided by a metafile enclosed to the dataset and shown in Table I.

Table I showed that the inconsistency in the annotation of the lesion areas on chest X-ray images was at a high rate. Some diseases, such as Pleural thickening, Pulmonary fibrosis, Pleural effusion, or lung pathologies in minority groups such as Atelectasis, Consolidation, and Calcification, had a rate of labeling overlap on other diseases up to 50%. The inconsistency in annotation among radiologists presented on not only the chest X-ray images with the bounding boxes of many lung diseases in the same lesion area but also many ones of the same lung disease at the approximate locations (Figures 6(a) and 6(b)). In addition, the bounding boxes of the same lung disease label had different sizes and were distributed at different locations inn the image (Figure 6(c)).

These are challenges for deep learning models, affecting the model performance as well as the accuracy of the model. We shall apply the Weighted Box Fusion technique described in Section III to remove overlapping bounding boxes on the same label at their approximate locations. In addition, the number of images, the various sizes, and the distribution of bounding boxes on the same lung disease also affect the training model because it causes over-fitting [15].

2. Performance Measurement and Evaluation

We evaluate the model performance through the MAP (Mean Average Precision), which was the average of AP (Average Precision) over the classes. AP was calculated for each class by the formula:

$$AP = \sum_{k=0}^{k=n-1} [Recall(k) - Recall(k+1)] * Precision(k), \tag{6}$$

where, $Recall(n) = 0$; $Precision(n) = 1$ and n = number of thresholds. Then, the equation of MAP is:

$$MAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k, \tag{7}$$

where, AP_k = the AP of class k; n = number of classes. In addition, Precision and Recall measure how accurate were your predictions and how good you found all the positives, respectively. They are defined by the formula:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}, \tag{8}$$

where, TP (true positive) metric presented the times of a positive sample that were correctly classified. FP (False Positive) metric indicated the times of a negative sample were incorrectly classified as positive ones. FN (False

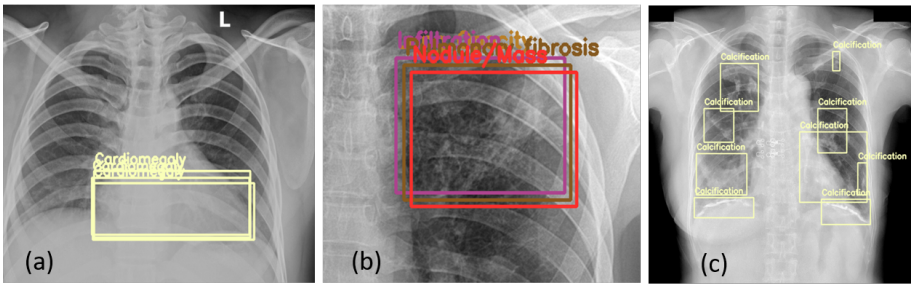


Figure 6. Chest x-ray images described the complexity of radiologists' annotations on lesion areas in the VinDr-CXR dataset.

Negative) metric gave the times of a positive sample were incorrectly classified as negative. The model was highly evaluated if both precision and recall were high. So, a metric called $F1$ score was used to measure the balance of precision and recall, which was calculated by the equation:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}. \quad (9)$$

The value of F1 was high meant that the precision and recall were high. Otherwise, the precision and recall were imbalanced.

3. Experiment results

We first applied the WBF algorithm to merge the bounding boxes of the same label in the chest X-ray dataset to create a new dataset. The chosen threshold of IOU in this algorithm was also crucial because it decided the reduction of the number of bounding boxes in the dataset after merging. The model performance was negatively affected if reducing the bounding boxes so many. Therefore, we investigated several IOU thresholds (0.4, 0.5, 0.6, and 0.7) and then implemented them in the model to find the best threshold for the algorithm. The results showed in Figure 7.

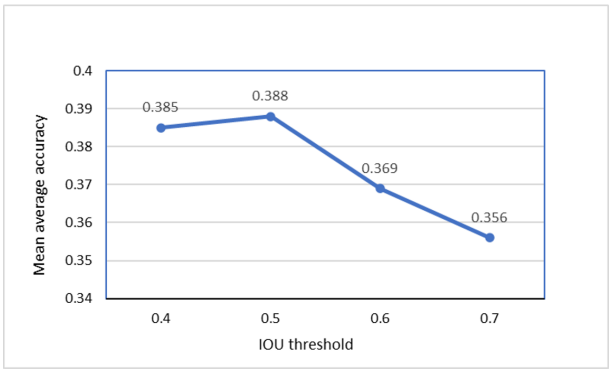


Figure 7. The performance of training model in some IOU thresholds of the WBF algorithm.

TABLE II
THE ORIGINAL DATASET AND THE NEW DATASET AFTER APPLYING THE WBF ALGORITHM FOR EXPERIMENT ON THE MODEL.

Dataset	Train		Validation		Test	
	Images	Bounding boxes	Images	Bounding boxes	Images	Bounding boxes
Before WBF	3861	28857	477	3731	496	3948
After WBF at IoU threshold of 0.5	3861	19236	478	2461	495	2647

In Figure 7, the IOU threshold of 0.5 in the WBF algorithm in the dataset brought the best MAP for the training model. It also implied that the merged bounding boxes at the IOU threshold of 0.5 would produce the most accurate boxes for the model. Then, we set up two datasets for experiments on the Yolov3-spp model. One was a dataset with original bounding boxes, and the other was a new dataset with bounding boxes that were applied by the weighted box fusion technique at the IOU threshold of 0.5. The two datasets are shown in Table II.

Besides, we first adapted hyperparameters of hue-saturate-value, scale, and fliplr augmentations and obtained their values of 0.5 appropriate for this dataset, we also used transfer learning from the Microsoft Common Objects in Context [20]. The experiment results are shown in Table III.

Table III showed that the performance of Yolov3-spp using the WBF dataset improved better than that using the original dataset. For the train set, the MAP increased from 0.320 to 0.388, and F1 scores increased from 0.36 to 0.43. The MAP and F1 scores on WBF data are also better than those on the original data. These expected results depended on the metrics of precision and recall, which were calculated on the confusion matrix of three parameters (TP, FB, FN) that presented the correct or incorrect classification of the model to lung lesion classes in the dataset. So, it was undoubted that the WBF algorithm has efficiency on the chest x-ray dataset and the model reduced its confusion during the training and testing. Therefore, the

TABLE III
THE EXPERIMENT RESULTS OF TWO MODELS, ONE IS ON THE ORIGINAL DATASET AND OTHER IS ON THE NEW DATASET WHICH WAS APPLIED BY THE WBF ALGORITHM AT THE IOU THRESHOLD OF 0.5.

Models	Validation				Test			
	P	R	MAP	F1 score	P	R	MAP	F1 score
Yolov3-spp with original dataset, basic augmentations	0.43	0.31	0.320	0.36	0.37	0.32	0.290	0.34
Yolov3-spp with WBF dataset, basic augmentations	0.50	0.40	**0.388**	**0.43**	0.35	0.45	**0.351**	**0.39**

TABLE IV
PERFORMANCE COMPARISON BETWEEN THE NMS AND WBF ALGORITHMS.

Models	Validation				Test			
	P	R	MAP	F1 score	P	R	MAP	F1 score
NMS	0.45	0.39	0.357	0.41	0.35	0.42	0.331	0.38
WBF	0.50	0.40	**0.388**	0.43	0.35	0.45	**0.351**	0.39

model performance is increasingly improved.

Moreover, we also used the non-maximum suppress (NMS) algorithm [20] to compare to the weighted box fusion (WBF) one [12] in the processing of overlap bounding boxes of the same label problem in the dataset. In practice, the NMS algorithm was often used for the final prediction to output bounding boxes of an object detection model [12]. At first, we carried out to implement the NMS and WBF techniques to create two separate datasets at the IOU threshold of 0.5. The implementation of two algorithms was described in Figure 8.

In Figure 8(b), a chest x-ray image presented three bounding boxes of Cardiomegaly labels occurring at an area and the same to Aortic enlargement lesion. The coordinates of their bounding boxes were described under the image. In Figure 8(a), a bounding box coordinate of Cardiomegaly and one of Aortic enlargement are left after applying the NMS algorithm, while a bounding box coordinate of Cardiomegaly and one of Aortic enlargement are averaged from the original ones in Figure 8(c). To find out the effect of two techniques on the dataset, we set up two new datasets at the IOU threshold of 0.5 for NMS and WBF, respectively, which were then experimented on the Yolov3-spp model. The results are shown in Table IV.

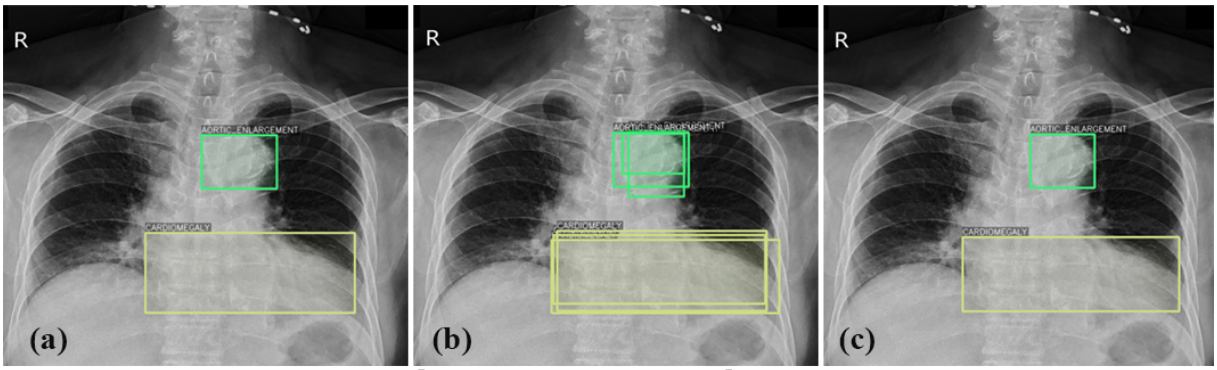
Table IV shows that the WBF technique's model performance was better than the one on NMS. It implied that the merging bounding boxes of WBF gave more accurate bounding boxes than the suppressing ones of the NMS technique.

On the other hand, data augmentation is always the first step to promote the diversity of images for the models. However, it is not a crucial factor to give better solutions because the wrong application of data augmentation may bring negative results. So, it is necessary to analyze the characteristics of the datasets thoroughly to use suitable augmentation methods for the dataset. In this chest x-ray dataset, the analysis showed us that more than half of the images in the dataset have a distribution of lesion areas in low and sparseness. The number of images, which had a sparse distribution of lesion labels, were described in Table V. Mosaic augmentation is quite suitable to apply for this kind of image because the technique combines more than one image into an image to create a broad view of images and diversity of data [14]. The mosaic images brought an extensive view of a scene instead of a single view. Hence, they gave comprehensive, diverse observations in a manner of an image created by the union of four images randomly. The hyperparameter of mosaic augmentation was specified by adjusting several parameters of probability in the range of 0.1:1. The parameter investigation gave us the best value of 0.7 in combination with the WBF dataset, which meant that the mosaic technique used 70% of images in the dataset. The results of training and testing showed in Table VI.

TABLE V
THE NUMBER OF IMAGES HAD A SPARSE DISTRIBUTION OF LESION LABELS.

The number of images	207	632	722	580
The number of lesion areas on each image	1	2	3	4

In Table VI, the dataset created from the WBF technique and included the mosaic augmentation was then trained and tested on the Yolov3-spp model. The MAP and $F1$ scores in the train set and test set gave us better results than those in the original dataset. The last model performance achieved the MAP of 0.418 on the train set and 0.385 on the test set when F1 scores obtained 0.45 for training and 0.42 for



[398.8, 489.8, 844.0, 659.9]
[518.3, 281.0, 678.2, 395.6]

[386.6, 508.5, 871.1, 664.3]
[390.2, 498.3, 841.6, 644.4]
[398.8, 489.8, 844.0, 659.9]
[537.9, 282.3, 669.2, 367.6]
[550.1, 276.2, 668.7, 416.8]
[518.3, 281.0, 678.2, 395.6]

[391.9, 498.9, 852.3, 656.2]
[535.4, 279.8, 672.1, 393.4]

Figure 8. Illustration of the differences between WBF and NMS: (a) The NMS keeps one bounding box that has the largest intersection with the remaining boxes; (b) The original abnormality locations marked by three doctors; (c) The WBF harmonizes the locations by averaging all the bounding boxes;

TABLE VI
THE EXPERIMENT RESULTS OF THE THREE MODELS.

Models	Validation				Test			
	P	R	MAP	F1 score	P	R	MAP	F1 score
Yolov3-spp with original dataset, basic augmentations	0.43	0.31	0.320	0.36	0.37	0.32	0.290	0.34
Yolov3-spp with WBF dataset, basic augmentations	0.50	0.40	0.388	0.43	0.35	0.45	0.351	0.39
Yolov3spp with WBF dataset, basic augmentations, and mosaic augmentation	0.49	0.45	**0.418**	**0.45**	0.39	0.46	**0.385**	**0.42**

testing. The results are high for this challenging dataset, especially since the values between precision and recall in training are high and not too far from each other, and as a result, $F1$ scores were high, too. This meant the imbalanced classes were significantly mitigated.

V. CONCLUSION

In this paper, we investigated the data limitation and annotation inconsistency problems in detecting abnormalities from chest X-ray images. Based on intensive experiments on a publicly available dataset, we show that the WBF algorithm effectively solves the overlapping bounding boxes of the same disease label due to the inconsistency among radiological experts. WBF is also more effective than other methods like NMS or without processing the differences in disease locations.

Overall, the quantity and quality of data used in deep learning models are still major and popular challenges, especially in the chest X-ray image domain. It requires non-stop improvement in deep learning methods and the artificial intelligence field in the future during data analysis and pre-processing play a key role therein. In the future, we will continue to find more solutions to solve the remaining challenges in this field.

ACKNOWLEDGMENT

This work is funded by the Vietnam Academy of Science and Technology under grant number NCVCC02.03/22-22.

REFERENCES

[1] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128 837–128 868, 2019.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017.

[3] S. Yao, Y. Chen, X. Tian, and R. Jiang, "Pneumonia detection using an improved algorithm based on faster r-CNN," *Computational and Mathematical Methods in Medicine*, vol. 2021, pp. 1–13, Apr. 2021.

[4] T. Liu, Q. Guo, C. Lian, X. Ren, S. Liang, J. Yu, L. Niu, W. Sun, and D. Shen, "Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks," *Medical Image Analysis*, vol. 58, p. 101555, Dec. 2019.

[5] D. M. Ibrahim, N. M. Elshennawy, and A. M. Sarhan, "Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases," *Computers in Biology and Medicine*, vol. 132, p. 104348, May 2021.
[6] Y. Tang, Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B. A. Redd, C. J. Brandon, Z. Lu, M. Han, J. Xiao, and R. M. Summers, "Automated abnormality classification of chest radiographs using deep convolutional neural networks," *NPJ Digital Medicine*, vol. 3, 2020.
[7] A. Majkowska, S. Mittal, D. F. Steiner, J. J. Reicher, S. M. McKinney, G. E. Duggan, K. Eswaran, P.-H. C. Chen, Y. Liu, S. R. Kalidindi, A. Ding, G. S. Corrado, D. Tse, and S. Shetty, "Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation." *Radiology*, p. 191293, 2019.
[8] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *CoRR*, vol. abs/1901.07031, 2019.
[9] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray: Hospital-scale chest x-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases," in *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*. Springer International Publishing, 2019, pp. 369–392.
[10] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. T. Tong, D. H. Dinh, C. D. Do, L. T. Doan, C. N. Nguyen, B. T. Nguyen, Q. V. Nguyen, A. D. Hoang, H. N. Phan, A. T. Nguyen, P. H. Ho, D. T. Ngo, N. T. Nguyen, N. T. Nguyen, M. Dao, and V. Vu, "Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations," *Scientific Data*, vol. 429, 2022.
[11] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox," in *Proceedings of the 24th ACM international conference on Multimedia*. ACM, Oct. 2016.
[12] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 107, p. 104117, Mar. 2021.
[13] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020.
[14] A. Alexey Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020.
[15] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, Jul. 2019.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition." Springer International Publishing, 2014, pp. 346–361.
[17] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2020.
[18] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jul. 2017.
[19] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767
[20] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," *CoRR*, vol. abs/1705.02950, 2017. [Online]. Available: http://arxiv.org/abs/1705.02950

Nguyen Thi Van Anh received the B.Sc. degree in Information Technology from the Hanoi University of Technology, Vietnam, and completed an M. Sc degree from Posts and Telecommunications Institute of Technology. She currently works as a researcher at Insitute of Physics, Vietnam Academy of Science and Technology.
Her research interests are data mining, deep learning and network technologies.
Email: anhntv@iop.vast.vn

Nguyen Duc Dung Assoc. Prof. Nguyen Duc Dung graduated with a Bachelor's degree in Mathematics in 1993 from the Hanoi University of Education, and a Master and Doctorate from the Japan Advanced Institute of Science and Technology (JAIST), Japan, in 2003 and 2006. He was a research engineer at KDDI R&D Laboratories Inc., Japan, from 2007 to 2009 and is currently Deputy Director of the Institute of Information Technology, Vietnam Academy of Science and Technology. His research interests include machine learning, pattern recognition, and data mining. Assoc. Prof. Dung was awarded the Innovation Badge of the Youth Union in 1998 for the successful development of Vietnam's first Vietnamese character recognition software, and the Technical Achievement Award in 2008 for his contributions at KDDI R&D Laboratories.
Email: nddung@ioit.ac.vn.

Nguyen Thi Phuong Thuy obtained my B.Sc degree in Physics Education from Hanoi National University of Education. She went on to complete M.Sc degree in Theoretical and Mathematical Physics from Institute of Physics, VAST. She is currently a PhD student at Graduate University Science and Technology. She did her internship in Department of Radiology and Diagnostics Imaging at University of Alberta from May 2017 to Feb 2018. She presently works as a researcher at Insitute of Physics,Vietnam Academy of Science and Technology.
Her primary research interest is modeling for liquid crystal structure to extract the formation, thermodynamic characteristics and to investigate the phase transitions.
Email: ntpthuy@iop.vast.vn.