DOI: 10.31276/VJSTE.65(4).84-93

Application of deep learning in chest X-ray abnormality detection

Nhan Ngo[1, 2], Toi Vo[1, 2], Lua Ngo[1, 2*]

[1]*School of Biomedical Engineering, International University, Vietnam National University - Ho Chi Minh City, Quarter 6, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam*
[2]*Vietnam National University - Ho Chi Minh City, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam*

Received 19 September 2022; revised 27 November 2022; accepted 2 December 2022

Abstract:

Lung diseases are the most common diseases worldwide, especially in Vietnam. Certain thoracic lung diseases can even lead to dangerous conditions for patients. X-ray are a useful imaging modality for detecting the abnormalities in the chest area. Furthermore, artificial intelligence can improve the detection of abnormalities in X-ray images, reduce misdiagnosis, close the knowledge gap between doctors, and alleviate the pressure on doctors. Therefore, this study aims to apply deep learning techniques to detect abnormalities in chest X-ray images and use data science and statistical approaches to improve the performance of the deep learning model. The data was explored and processed to obtain high quality data with optimal characteristics. We then applied data augmentation and optimization to the RetinaNet model with ResNet101 in a Feature Pyramid Network (FPN) backbone to achieve the best performance. Our model achieved mean average precision of 0.55 at a threshold of 0.5 (mAP@0.5) in a validation set, which included five diseases: aortic enlargement, cardiomegaly, interstitial lung disease, infiltration, and nodule/mass.

Keywords: **deep learning, RetinaNet model, thoracic lung diseases, X-ray.**

Classification numbers: **3.2, 3.6**

1. Introduction

The term "lung disease" refers to a variety of disorders that affect the lungs, such as asthma, chronic obstructive pulmonary disease (COPD), infections like influenza, pneumonia, tuberculosis, lung cancer, and other breathing problems. Some lung diseases can even result in respiratory failure [1]. According to the 2017 Behavioral Risk Factor Surveillance System (BRFSS) survey, approximately 22.5 million (9.1 percent) adults residing in the United States and 7.9 percent of children from twenty-seven states and the District of Columbia reported currently having asthma. About 16.3 million adults (6.6 percent) reported ever being diagnosed with COPD. Close to 33.2 million adults (13.4% reported being diagnosed with chronic lung disease [2].

In most of the cases, doctors rely on X-ray images to diagnose lung diseases. However, they face many daily challenges, with perhaps the greatest difficulty being the large number of chest radiographs to be reviewed. Additionally, the interpretation of X-rays image can lead to medical misdiagnosis, even among the best practicing doctors [3].

Among the available studies on detecting lung diseases in chest X-ray (CXR), one of the most interesting is that of M.T. Islam, et al. (2017) [4]. Their study used ensemble deep learning models to improve the classification of cardiomegaly compared to a single deep learning model. Although the study had an impressive accuracy of 93.0% and an area under the ROC curve (AUC) score [5] of 0.97, it mainly used image classification methods to detect cardiomegaly disease. Therefore, the method in that study is difficult to apply for the detection of many other lung diseases. Another study by H. Huang, et al. (2021) [6] used the Yolo [7] model to detect fourteen lung diseases from the VinBigData dataset [3]. However, the image processing stage in the Huang study only removed images without disease and resized images to 512x512 pixels. The resulting mean average precision at a threshold of 0.5 was not high, with a baseline model of 0.21 and 0.34 after model selection and hyperparameters tuning.

This study applies a state-of-the-art deep learning model to detect lung diseases in five out of fourteen categories in the VinBigData dataset, including aortic enlargement [8], cardiomegaly [9], interstitial lung disease (ILD) [10], infiltration [11] and nodule/mass [12]. Transfer learning techniques were applied by employing a pre-trained RetinaNet [13] model with a ResNet101 [14] backbone, in combination with the smooth L1 loss function [15] for bounding box regression and focal loss (FL) [13] for image classification. Weighted box fusion (WBF) [16] was applied in the first stage to eliminate overlapping ground truth bounding boxes and in last stage to eliminate overlapping predicted bounding boxes. The FPN [13] and FL-based RetinaNet provide an end-to-end approach that achieves high accuracy. The evaluated score from multiple metrics of this solution is promising, accurate, and can be encapsulated via a website to aid doctors in thoracic lung disease detection.

*Corresponding author: Email: ntlua@hcmiu.edu.vn

Crossref Similarity Check Powered by iThenticate

2. Materials and methods

The original dataset, consisting of 18,000 poster-anterior (PA) chest X-ray (CXR) scans in DICOM format with total size is 191.82 GB, was supplied directly by VinBigData [3]. However, this study utilised the dataset indirectly provided by VinBigData, which was converted to JPEG images with a total size of 3.4 GB [17]. This pure Vietnamese dataset includes 15,000 CXR images for training and 3,000 images used to evaluate algorithm models. The data were collected from 108 Military Central Hospital and Hanoi Medical University Hospital, with each image read and diagnosed by professional clinicians.

Of the 18,000 X-ray images provided, 15,000 images were used for the training set, with each image having multiple annotations by doctors via a CSV file. The remaining 3,000 images were used for the test set without any annotations. All images in the training set were labelled for the presence of 14 diseases: aortic enlargement, atelectasis [18], calcification [19], cardiomegaly, consolidation [20], ILD, infiltration, lung opacity [21], nodule/mass, other lesion, pleural effusion [22], pleural thickening [23], pneumothorax [24], and pulmonary fibrosis [25]. The "No finding" class was used to indicate the absence of all the above findings.

The annotations for the training data consist of one row for each object, including the type of disease and offset value of the ground truth bounding box. The following columns are included with each row: image_id: image filename, class_name: the name of the diagnosed disease of the object (or No finding), class_id: the ID of the disease of detected object, rad_id: the ID of doctor who made the observation, and x_min, y_min, x_max, y_max: the offset values of the object's bounding box. If the class_name is "No finding," then the values in the columns (x_min, y_min, x_max, y_max) are not a number (NaN) as shown in Table 1.

Table 1. Sample ground truth from the doctors.

image_id	class_name	class_id	rad_id	x_min	y_min	x_max	y_max
50a418190bc3fb1ef1633bf9678929b3	No finding	14	R11				
21a10246a5ec7af151081d0cd6d65dc9	No finding	14	R17				
9a5094b2563a1ef3ff50dc5c7ff71345	Cardiomegaly	3	R10	230.0	458.0	551.0	610.0
051132a778e61a86eb147c7c6f564dfe	Aortic enlargement	0	R10	421.0	247.0	537.0	339.0
063319de25ce7edb9b1c6b8881290140	No finding	14	R10				

2.1. Data analysis

The original CSV file had 67914 rows with 15,000 images, with each row containing the annotation of each image's ID by an individual doctor. As a result, each image was diagnosed by many doctors. The positive class names are the 14 diseases, which account for 36,096 (53%) annotations and 4,394 (29%) images. The rest of the training data belongs to the negative class, which is "No finding" - as shown in Fig. 1.

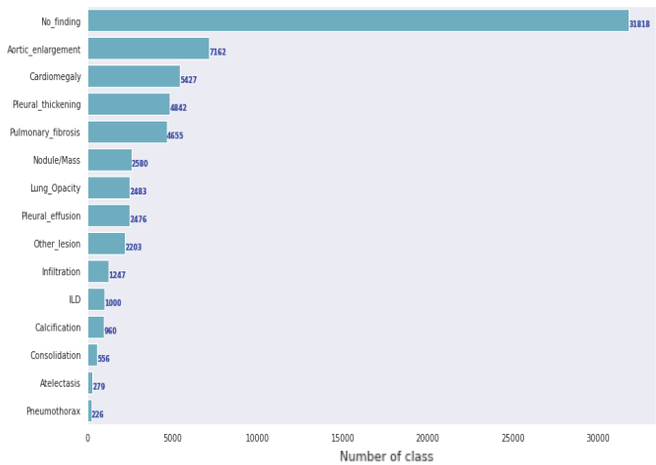


Fig. 1. The number of classes.

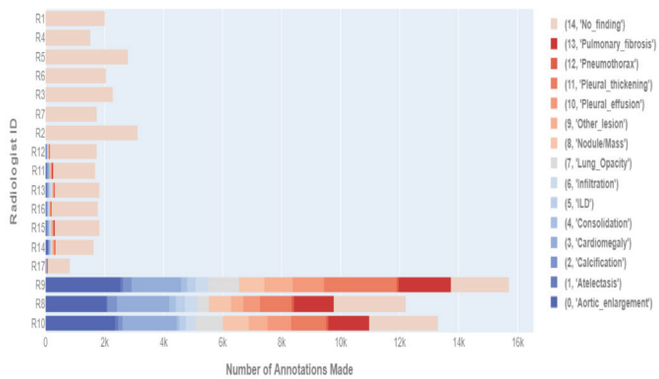


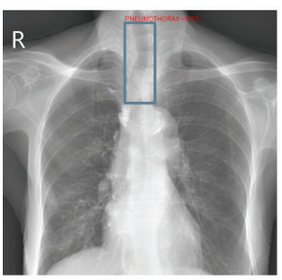
Fig. 2. The number of annotations of each doctor.

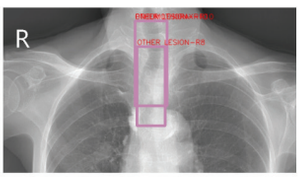
Figure 2 illustrates the relationship between the number of classes and the doctors' contribution by plotting their contributions. The dominance of the "No finding" class is due to the difference in the doctors' contributions. Fourteen out of the seventeen doctors have 80% of the "No finding class" in their contribution, and six of them have 100% of the "No finding" class. Meanwhile, certain doctors mainly annotated all images with positive classes (i.e., those that are not "No finding").

In object detection, the negative class name is not used to train model, so all annotations of the negative class need to be dropped. Based on our analysis of the explored data and data description, we divided the doctors with positive classes into two groups. All annotations of doctor number 8 (R8), doctor number 9 (R9), and doctor number 10 (R10) contain almost all images with positive classes, 4,146 images, which account for 94.35% of the images. All annotations from doctor number 11 (R11) through doctor number 17 (R17) are included in 248 images with positive classes that are not the same images of the first group (of R8, R9, R10).

Based on the heatmap of the bounding boxes of the classes, the diagnostic consistency of three doctors and disease location, the first group is considered a standard dataset for training models. In addition, both groups of doctors have mismatched annotations. Pneumothorax is known as a collapsed lung, but as shown in Fig. 3, the annotations from R10 indicate the pneumothorax is detected in

the middle bone. R10 also labelled that location as "Other lesion," which is why the pneumothorax and "Other lesion" have the same offset values of bounding boxes. This way of annotation has a problem. R10 may have intended to label that location as "Other lesion" but mistakenly labelled it as pneumothorax, or they may have forgotten to remove the pneumothorax annotation.





class_name	rad_id	x_min	y_min	x_max	y_max
Pneumothorax	R10	342	34	424	252
Lung Opacity	R10	104	461	160	532
Other lesion	R10	342	34	424	252
Aortic enlargement	R10	350	257	495	385

Fig. 3. Examples of wrong annotation from doctors.

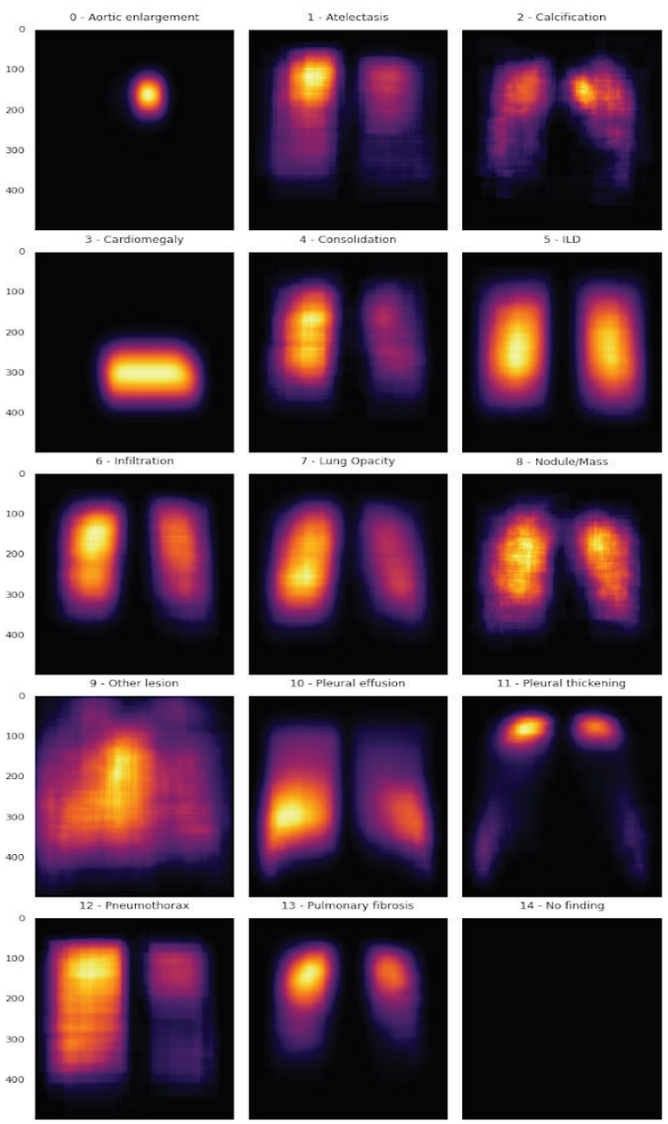
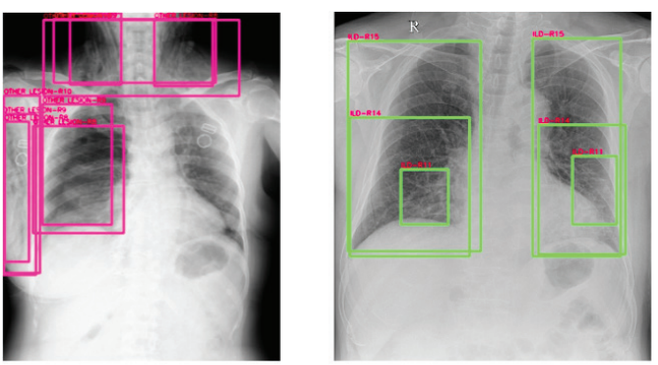


Fig. 4. Heat map for the 14 different diseases and "No finding".

Figure 4 shows the heat map of all bounding boxes of 14 diseases in this dataset. Only two diseases, aortic enlargement and cardiomegaly, have centralized bounding boxes. The bounding boxes of the other diseases are too scattered. For instance, the ILD bounding boxes are located in two main regions (one on the left and one on the right side of the chest), but within these regions, the diseases can be labelled at different locations. The areas marked with ILD on the right side of Fig. 5, annotated by R11, are smaller than that annotated by R15. The reason for the difference in the bounding boxes for this ILD by different doctors is that some doctors may want to identify the damage clearly and draw the box right at the damaged tissue position, while other doctors draw the box bigger to encompass the entire location. The bounding boxes for pneumothorax, consolidation, and lung opacity are also located in two main regions, the left and right side of the chest, but are concentrated mainly on the right side. Another particular case that shows inconsistency in annotation is "Other lesion." Although it mainly locates in the middle region, it can appear at many different regions, resulting in a high dispersion of the bounding box positions for "Other lesion."

Figure 5 clearly shows the scattered property of the bounding box distributions. The ground truth bounding boxes of "Other lesion," annotated by different doctors, are located in three different main regions, namely, lower right-hand area, right lung location, and the upper part of the neck. The "Other lesion" in the VinBigData dataset refers to a lesion of the thoracic lung that does not belong to the 13 other mentioned diseases (aortic enlargement, atelectasis, calcification, cardiomegaly, consolidation, ILD, infiltration, lung opacity, nodule/mass, pleural effusion, pleural thickening, pneumothorax, and pulmonary fibrosis). Another example that clearly shows inconsistency in annotation is the image on the right side of Fig. 5, which shows ILD disease. It can be observed that the sizes of bounding boxes annotated for ILD very greatly among different doctors.



Other lesion **ILD**

Fig. 5. Examples of inconsistency in doctors' diagnoses.

2.2. Data processing

Because of all of the differences in the size of the bounding boxes mentioned above, in this study, we did not eliminate the ground truth bounding boxes but modified them based on the

Intersection over Union (IoU) of all bounding boxes in the same class. There are several techniques depending on IoU, such as Non-Maximum Suppression (NMS) [26], Soft Non-Maximum Suppression (Soft-NMS) [27], and WBF [16] (Table 2). This study applied all three techniques to compare their results.

Table 2. Performance of NMS, Soft-NMS and WBF IoU techniques.

	Number of boxes
Original boxes	36,096
NMS@0.5	23,940 (↓ 33.7%)
Soft-NMS@0.5	32,273 (↓ 10.6%)
WBF@0.5	23,955 (↓ 33.64%)

Figure 6 explains the difference between NMS and WBF. WBF is based on all the ground truth bounding boxes of three doctors (R8, R9, R10) and generates the new bounding boxes. In another circumstance, the NMS chooses the bounding boxes with highest confidence score. However, in this case, all bounding boxes are the ground truths, and therefore all of the confidence scores of those boxes are equal to 1. Consequently, NMS chooses randomly. In Fig. 6, NMS randomly chose the ground truth bounding box of doctor R2. That is why the performance of WBF is better than that of NMS because it generates new bounding boxes based on all information on the ground truth bounding boxes of the doctors. Fig. 7 shows the different performances of NMS and WBF with an IoU threshold equal to 0.2. NMS eliminates the large "Other lesion" ground truth in the middle, while WBF generates new ground truth based on all ground truth bounding boxes. resulting in new large "Other lesion" ground truth bounding boxes on the left.

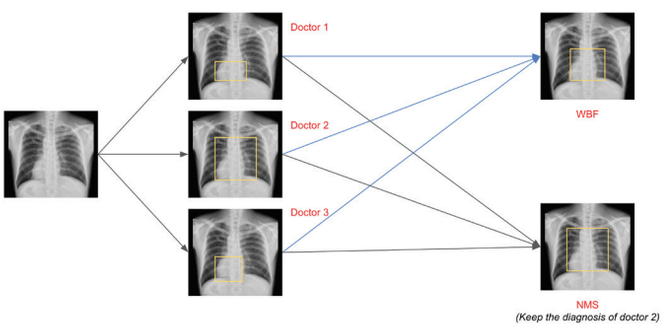


Fig. 6. Comparison of the NMS and WBF techniques.

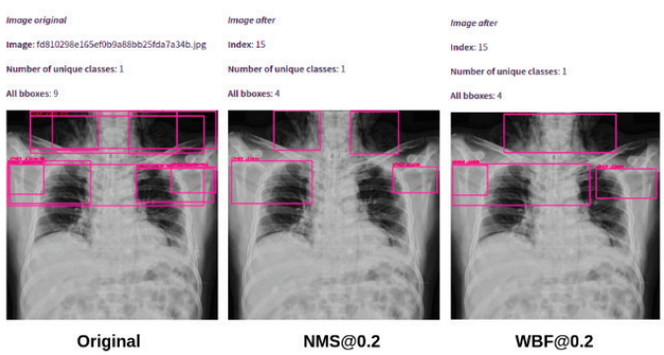


Fig. 7. Performance of NMS and WBF at an IoU threshold of 0.2.

Based on the data analysis section, the original dataset was split into two groups. The first group contains all annotations of R8, R9, and R10, which account for almost all of the images (94.35%) and is called the standard dataset. The second group contains all annotations from R11 through R17, which contains other parts of images and has many errors in the bounding boxes due to inconsistent diagnoses by doctors. After processing this group, it is called the Additional Dataset.

Figure 8 shows that the performance of WBF varies with different IoU thresholds. When the IoU is 0.3, WBF reduces several ground truth bounding boxes, while at an IoU of 0.2, the new ground truth bounding box on the top right contains the wrong location of the calcification disease. That is why we must calculate the IoU of each disease with different diagnoses of doctors and then obtain the average value of those IoUs. This study used the standard dataset to calculate the average value of all diseases except "Other lesion" and pleural thickening because those two diseases have dispersion in the heat map (Fig. 4). Our result of the IoU threshold is 0.4, as shown in Table 3.

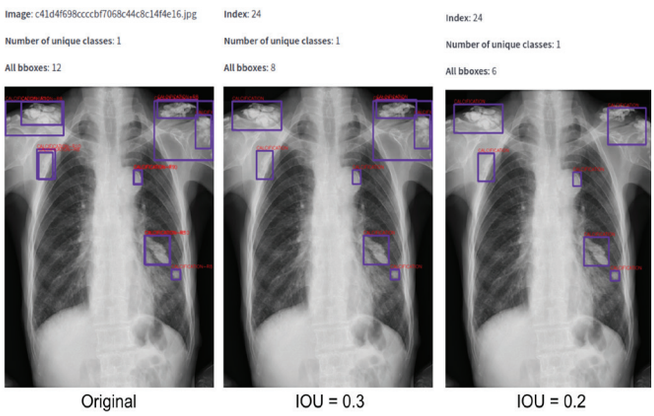


Fig. 8. WBF with different IoU values.

Table 3. Various IoU thresholds of WBF.

Class name	IOU in ground truth	IOU threshold
Aortic enlargement	0.688	0.3
Atelectasis	0.48	0.5
Calcification	0.55	0.5
Cardiomegaly	0.73	0.3
Consolidation	0.6	0.4
ILD	0.59	0.4
Infiltration	0.57	0.4
Lung Opacity	0.52	0.5
Nodule/Mass	0.64	0.4
Other lesion	0.5	0.5
Pleural effusion	0.5	0.5
Pleural effusion	0.47	0.5
Pneumothorax	0.68	0.5
Pulmonary fibrosis	0.5	0.5
Mean with except "Other lesion" and pleural thickening =0.4		

To process all annotations from R11 to R17, min-max normalisation is first applied for the offset value of the ground truth bounding boxes. Then, the area of the ground truth bounding boxes is calculated after each disease normalization. The values are then averaged for each disease saved in the standard dataset. Table 4 compares the diagnoses of each doctor in the Additional Dataset. In Table 5, the standard dataset with doctor R11 is compared to the Additional Dataset to obtain suitable ground truth bounding boxes for R11.

Table 4. Mean bounding boxes area norm in the standard dataset.

Rad_id	Class name	Num bounding boxes	Mean bounding boxes area norm
R8_R9_R10	Aortic enlargement	6956	0.0145
	Atelectasis	230	0.04917
	Calcification	758	0.00764
	Cardiomegaly	5268	0.05997
	Consolidation	520	0.03731
	ILD	812	0.07503
	Infiltration	1189	0.03886
	Lung opacity	2382	0.02751
	Nodule/mass	2468	0.00427
	Other lesion	2035	0.0248
	Pleural effusion	2395	0.02423
	Pleural effusion	4741	0.00698
	Pneumothorax	219	0.09259
	Pulmonary fibrosis	4489	0.01511
Total annotations: 34462/36096 Total images: 4146/4394			

Table 5. Standard dataset and R11 annotations.

Rad_id	Class name	Residual bounding boxes area norm (R11-R8_R9_R10)	Ratio bounding boxes area norm (R11/R8_R9_R10)
R8_R9_R10 and R11	Aortic enlargement	0.03701	3.55
	Atelectasis	0.0093	1.19
	Calcification	0.00049	1.06
	Cardiomegaly	0.04316	1.72
	Consolidation	0.01147	1.31
	ILD	-0.04563	0.39
	Infiltration	-0.03886	0
	Lung opacity	-0.01554	0.43511
	Nodule/mass	0.00597	2.39813
	Other lesion	0.01215	1.48992
	Pleural effusion	0.00826	1.34
	Pleural effusion	0.00946	2.36
	Pneumothorax	-0.08059	0.13
	Pulmonary fibrosis	0.0033	1.2184
Total annotations: 34462/36096 Total images: 4146/4394			

After comparing all ground truth bounding boxes of doctors R11-R17 to obtain the novel Additional Dataset, it was added into the standard dataset to generate the Final data. Fig. 9 summarises the above data processing steps.

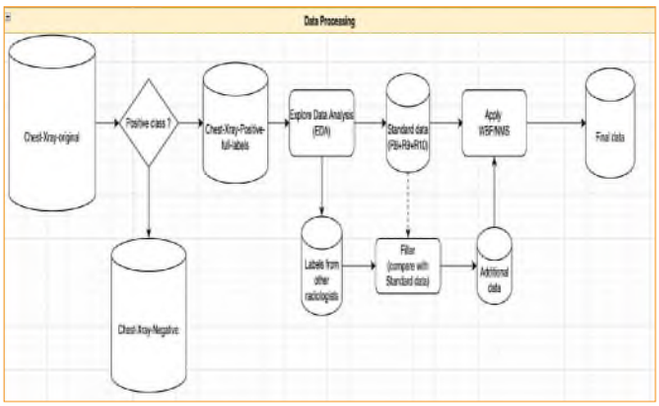


Fig. 9. Data processing steps.

After generating the Final data, several image augmentation techniques were applied such as padding, horizontal flip, and image rotation. Then, the data was randomly split with 80% used for the training dataset and 20% for validation dataset, as seen in Fig. 10.

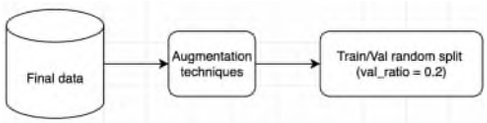


Fig. 10. Augmentation and Train-Val split.

Figure 11 summarises the process from analysing the data to splitting the training data and deploying it into production.

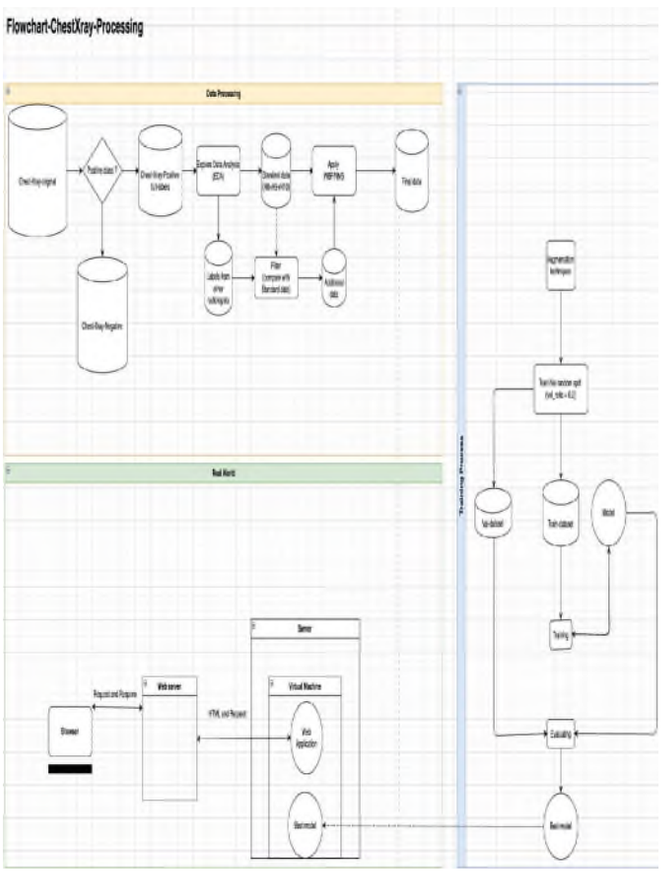


Fig. 11. Flowchart of the study.

2.3. Bounding box suppression techniques

In object detection models, having many overlapped bounding boxes can lead high recall values. Thus, the NMS technique is applied to eliminate redundant overlapping bounding boxes. NMS greedily selects high scoring detections and deletes nearby, less confident neighbours since they are likely to cover the same object. This algorithm is simple, fast, and surprisingly competitive compared to proposed alternatives [26]. The NMS algorithm is described in Algorithm 1.

Algorithm 1: Non-maximum suppression

input: $B = \{b_1, b_2, \dots, b_N\}, S = \{s_1, s_2, \dots, s_N\}, N_t$
B: the list of bounding boxes of single class in individual image
S: the list of confidence scores corresponding to bounding boxes
N_t: the NMS threshold
begin
$D \leftarrow \{\}$
while $B \neq empty$ **do**
$m \leftarrow argmax\ S$
$M \leftarrow b_m$
$D \leftarrow D \cup M$
$B \leftarrow B - M$
for b_i *in* B **do**
if $iou(M, b_i) \geq N_t$ **then**
$B \leftarrow B - b_i$
$S \leftarrow S - s_i$
end
end
end
return D, S
end

Soft-NMS is a variant of NMS. Instead of eliminating the redundant overlapping bounding boxes, Soft-NMS penalizes the redundant bounding boxes by reducing their confidence scores. The Soft-NMS algorithm is shown in Algorithm 2.

Algorithm 2: Soft non-maximum suppression

input: $B = \{b_1, b_2, \dots, b_N\}, S = \{s_1, s_2, \dots, s_N\}, N_t$
B: the list of bounding boxes of single class in individual image
S: the list of confidence scores corresponding to bounding boxes
N_t: the NMS threshold
begin
$D \leftarrow \{\}$
while $B \neq empty$ **do**
$m \leftarrow argmax\ S$
$M \leftarrow b_m$
$D \leftarrow D \cup M$
$B \leftarrow B - M$
for b_i *in* B **do**
$s_i = \begin{cases} s_i & if\ iou(M, b_i) \leq N_t \\ s_i(1 - iou(M, b_i)) & if\ iou(M, b_i) \geq N_t \end{cases}$
end
end
return D, S
end

The WBF method is used to combine predictions of object detection models. Unlike the NMS and soft-NMS methods, which simply remove part of the predictions, the proposed WBF method uses the confidence scores of all the proposed bounding boxes to construct average boxes [16]. Algorithm 3 describes the WBF process.

Algorithm 3: Weighted boxes fusion

input: $B = \{b_1, b_2, \dots, b_N\}, S = \{s_1, s_2, \dots, s_N\}, N_t$
B: the list of bounding boxes of single class in individual image
S: the list of confidence scores corresponding to bounding boxes
N_t: the NMS threshold
begin
$WB \leftarrow \{\}$
$WS \leftarrow \{\}$
$FB \leftarrow \{\}$
$FS \leftarrow \{\}$
while $B \neq empty$ **do**
$m \leftarrow argmax\ S$
$M \leftarrow b_m$
$WB \leftarrow b_m$
$WS \leftarrow s_m$
$B \leftarrow B - M$
for b_i *in* B **do**
if $iou(M, b_i) \geq N_t$ **then**
$WB \leftarrow b_i$
$WS \leftarrow s_i$
else
$FB \leftarrow b_i$
$FS \leftarrow s_i$
end
end
$FB \leftarrow f_b(WB)$
$FS \leftarrow f_s(WS)$
end
end

In this algorithm, $WB=\{b_1, b_2 \dots b_K\}$ is the list of overlapping bounding boxes with an IoU greater than the threshold (N_t). $WS=\{s_1, s_2 \dots s_K\}$ is the list of confidence scores corresponding to WB with each $b_1=\{x_1^{min}, x_1^{max}, y_1^{min}, y_1^{max}\}$, and s_1 is confidence score corresponding to b_1. The weighted boxes function f_b and confidence score rescale function f_s are given in detail below:

Function f_b:

$$x^{min} = \frac{\sum_{i=1}^{K} s_i \times x_i^{min}}{\sum_{i=1}^{K} s_i} \qquad x^{max} = \frac{\sum_{i=1}^{K} s_i \times x_i^{max}}{\sum_{i=1}^{K} s_i}$$

$$y^{min} = \frac{\sum_{i=1}^{K} s_i \times y_i^{min}}{\sum_{i=1}^{K} c_i} \qquad y^{max} = \frac{\sum_{i=1}^{K} s_i \times y_i^{max}}{\sum_{i=1}^{K} c_i}$$

Function f_s:

$$s = \frac{\sum_{i=1}^{K} s_i}{N}$$

K is the number of overlapped bounding boxes.

2.4. Methods

2.4.1. Transfer learning with RetinaNet

Within deep learning, transfer learning is a technique whereby a neural network model is first trained on a problem similar to the problem that is being solved. One or more layers from the trained model are then used in a new model trained on the problem of interest [28].

Based on the advantages of transfer learning and the goal of this study, a popular model called RetinaNet was employed, which was presented by T.Y. Lin, et al. (2017)

[13] from Facebook AI study. RetinaNet is a one-stage object detection model that utilizes a FL function to address class imbalance during training. FL applies a modulating term to the CE loss in to focus learning on hard negative examples [13]. RetinaNet's network architecture uses an FPN in the backbone, with the bottom-up being the ResNet architecture [14].

2.4.2. Loss functions

Loss functions are one most important aspects of a deep learning model. They compute the distance (or error) between the actual values and predictions of the models to optimize parameter values in the models. In this study, two loss functions are applied, namely, FL for classification task and Smooth L1 Loss [15] for regression tasks.

FL is designed to address one-stage object detection scenarios in which an extreme imbalance between foreground and background classes during training exists [13]. The formula for FL is:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} \times \log(p_t) \quad (1)$$

where p is the model's estimated probability for the class with a label of y=1 in binary classification: $p_t = \begin{cases} p & if\ y = 1 \\ 1-p & otherwise \end{cases}$, α_t is the balance variant of FL to address the imbalance problem, and γ is the tuneable focusing parameter. When γ=0, FL is equivalent to cross-entropy (CE) [29] and as γ increases, the effect of the modulating factor is likewise increased [13].

Smooth L1 loss for object detection was originally proposed in Fast R-CNN [15]. Smooth L1 loss is used to make bounding box regression more robust by replacing the excessively strict L2 loss. The formula of Smooth L1 Loss is:

$$L_{loc}(t^u, v) = \sum_{i \in \{x,y,w,h\}} smooth_{L1}(t_i^u - v_i) \quad (2)$$

where u is ground-truth of class u; v is ground-truth offset values of the bounding box, and

$$smooth_{L1}(t_i^u - v_i) = \begin{cases} 0.5(t_i^u - v_i)^2 & if\ |t_i^u - v_i| < 1 \\ |t_i^u - v_i| - 0.5 & otherwise \end{cases} \quad (3)$$

2.4.3. Metrics evaluation

In object detection, some applicable evaluation metrics are Intersection over Union (IoU), common statistic metrics such as True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision, Recall, confidence score, and mean Average Precision (mAP). Min-max normalisation [30] is used to standardise data in the data processing.

IoU is a metric evaluation that describes the extent of overlap of two bounding boxes, and a simple implementation of IoU is described in Fig. 12.

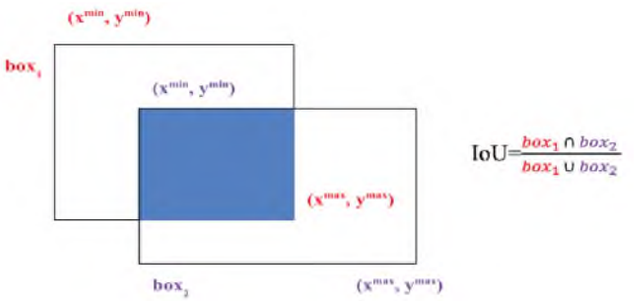


Fig. 12. Intersection over union formula.

TP is the total number of correct positive predictions of the model, TN is the total number of correct negative predictions of model, FP is the total number of incorrect positive predictions of model, and FN is the total number of incorrect negative predictions of model. However, in an object detection problem, TP is the number of predicted bounding boxes of the model with an IoU of the ground truth bounding box greater than the IoU threshold (in most cases the IoU threshold is 0.5). FP is the number of predicted bounding boxes of the model with an IoU lower than the IoU threshold, FN is number of models that cannot predict an object in the image, and TN is the number of models that correctly predict background in the image but in the object detection problem.

The precision in object detection is the ratio of correctly predicted bounding boxes of the model and all bounding boxes predicted by the model. Meanwhile, recall is the ratio of the correct bounding boxes predicted by the model and all ground truth bounding boxes.

$$precision = \frac{correct\ bounding\ boxes\ prediction}{all\ bounding\ boxes\ prediction} \quad (4)$$

$$recall = \frac{correct\ bounding\ boxes\ prediction}{all\ ground\ truth\ bound\ boxes} \quad (5)$$

The confidence score describes how the model predicts this bounding box for each class. A high confidence score means that the bounding boxes are accurately predicting the location and classification of the objects. The formula for confidence score is show below:

$$s = P(object) \times IoU(object, groundtruth) \quad (6)$$

Figure 13 illustrates how the confidence score is calculated in object detection.

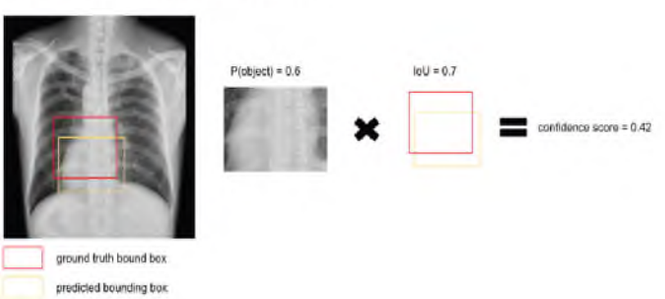


Fig. 13. Confidence score calculation method.

Average Precision (AP) is the average precision corresponding to the recall when changing the confidence score threshold. The confidence score threshold is the threshold used to control the prediction of model. For example, if the predicted bounding boxes have a confidence score lower than the confidence score threshold, then the predicted bounding boxes cannot be displayed as a prediction of the model, thus precision and recall are affected. Therefore, AP is the average precision when changing the confidence score from 1.0 to 0.

The mAP is the average AP of all the classes. The formula for mAP is

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{7}$$

where N is total number of classes in the dataset, i is the contidence score from 1.0 to 0. The mAP value is affected by the IoU threshold because when changing the IoU threshold, it changes the value of TP and FP. Thus, mAP@(IoU threshold) is used to denote the mAP value corresponding to IoU threshold value. Common types of mAP are mAP@0.5 mean mAP with IoU threshold is 0.5 and mAP@[0.5:0.95] means that the average of mAP with IoU threshold from 0.5 to 0.95 with increase step is 0.5.

The purpose of normalisation is to have every data point on the same scale, and min-max normalisation is one of the most common methods of data normalization. The min-max normalisation procedure scales the minimum value to 0 and the maximum value to 1. Every other value is transformed into a decimal value between 0 and 1. In this study, min-max normalisation were applied to the offset values of the ground truth bounding boxes instead of the X-ray images. The min-max normalization method for offset values of ground truth bounding boxes is given as follows:

$$x_i^{\min}{}_{norm} = \frac{x_i^{\min}}{image_{width_i}} \tag{8}$$

$$y_i^{\min\_norm} = \frac{y_i^{\min}}{image\_height_i} \tag{9}$$

$$x_i^{\max}{}_{norm} = \frac{x_i^{\max}}{image_{width_i}} \tag{10}$$

$$y_i^{\max\_norm} = \frac{y_i^{\max}}{image\_height_i} \tag{11}$$

where i is the i^{th} image, and the area of ground truth bounding box after normalization is:

$$area_i^{norm} = \left(x_i^{\max\_norm} - x_i^{\min\_norm}\right) \times \left(y_i^{\max}{}_{norm} - y_i^{\min}{}_{norm}\right) \tag{12}$$

3. Implementation of model

3.1. Model overview

As shown in Fig. 14, the training process of our system employed the RetinaNet model. Furthermore, based on experimental results, ResNet101 was chosen as the best backbone for RetinaNet in this study.

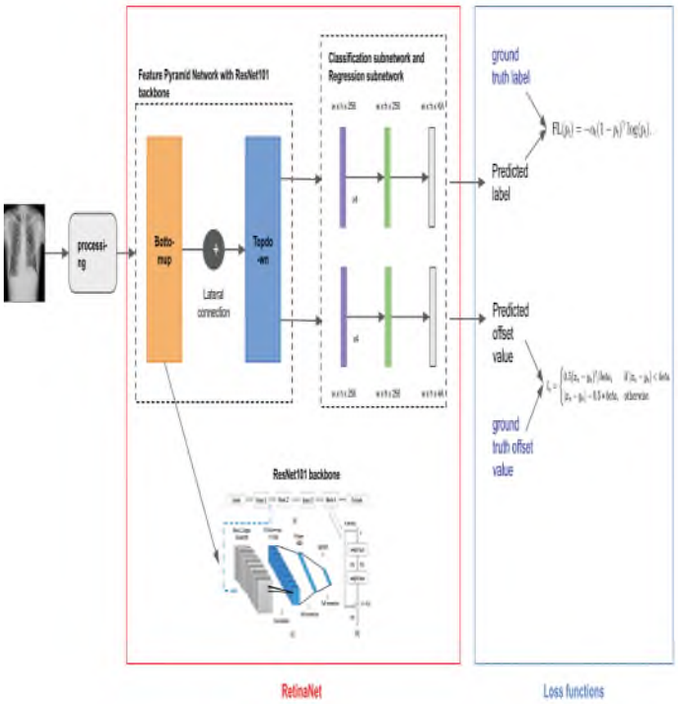


Fig. 14. Model architecture.

3.2. Implementation

3.2.1. FPN with ResNet101 backbone

Figure 15 illustrates the working mechanism of FPN with the ResNet101 backbone. As seen on the left-hand side of Fig. 15, the bottom-up pathway contains a five-stage feature map (C1-C5), which has different scales and channels. The right-hand side shows the top-down structure of FPN, which contains five components (P2-P6), with each component having a lateral connection with the corresponding stage of the feature map of the bottom-up pathway. Components P2-P6 have 256 channels, which are the output of the FPN. The purpose of FPN is to extract the most informative data from the image.

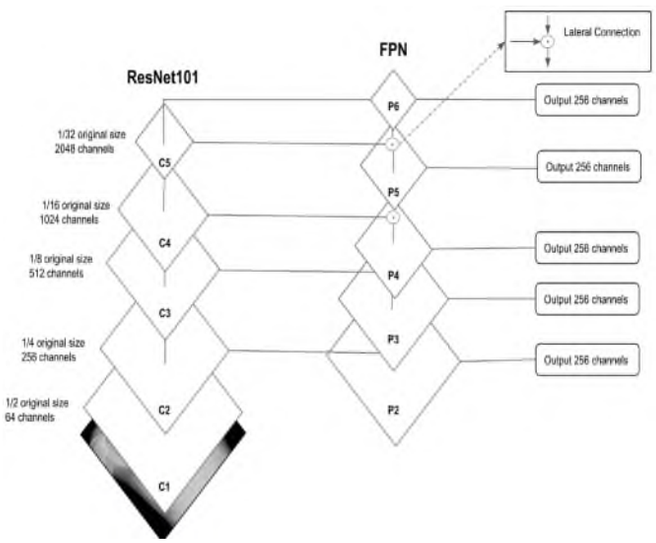


Fig. 15. The FPN with ResNet101 backbone.

Table 6. Performance of RetinaNet with ResNet50 backbone of the 14 diseases.

Class	mAP@[0.5:0.95]
Aortic enlargement	0.5457
Atelectasis	0.0456
Calcification	0.0023
Cardiomegaly	0.5170
Consolidation	0.0311
ILD	0.0508
Infiltration	0.0568
Lung Opacity	0.0398
Nodule/Mass	0.0454
Other lesion	0.0015
Pleural effusion	0.0926
Pleural thickening	0.0323
Pneumothorax	0.0012
Pulmonary fibrosis	0.0322
mAP@[0.5:0.95]: 0.1067 mAP@0.5: 0.2174	

3.2.2. Classification and regression subnetworks

The outputs of RetinaNet are the predictive class labels and the offset value of bounding boxes corresponding to the predictive class labels. Therefore, RetinaNet has two subnetworks for the output: one of them treats the region of interest as image classification and the other treats bounding boxes regression.

4. Results and discussion

4.1. Experiment with all diseases

RetinaNet with ResNet50 and an FPN backbone (3x) gave the best results in the validation set without augmentation after 2500 epochs by applying Detectron2 from the Facebook AI study [31]. Table 6 shows the performance of the model with mAP@[0.5:0.95].

4.2. Experiment with 5 diseases

Based on the results from Table 6 and the problems in data processing, this study selected 5 of the 14 classes that were highly trainable: aortic enlargement, cardiomegaly, ILD, infiltration and nodule/mass. Because of the dataset imbalance,

Table 7. Performance of RetinaNet with different backbones and augmentation methods.

Backbone of FPN	Augmentation	mAP@[0.5:0.95]	mAP@0.5
ResNet50	None	0.2627	0.4547
ResNet50	Horizontal flip + Rotation	0.3052	0.535
ResNet101	Horizontal flip + Rotation	0.3193	0.5500

image augmentation methods like horizontal flip, rotation, and oversampling were deployed for the dataset. Table 7 shows the results of the model in the validation set. RetinaNet with ResNet101 and an FPN backbone gave the best result.

Figure 16 shows the loss functions of ResNet 101 with the FPN backbone during the training procedure. The loss function generally decreased; however, the decrease was not stable because the ground truths labelled by doctors were inconsistent, causing difficulty during the training process. The top right subfigure of Fig. 16 gives the learning rate, and this study applied the learning rate schedule technique [32] with a Cosine warm-up function.

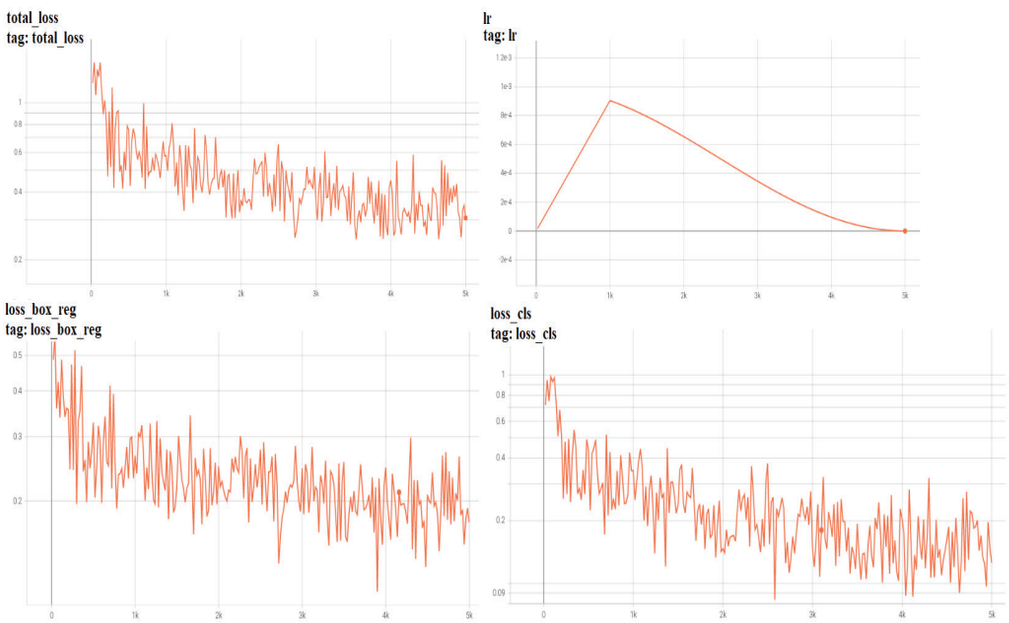


Fig. 16. Loss function and learning rate values during the training process.

4.3. Discussion

Table 8 compares the mAP of the 14 lung disease predictions from the proposed method and other publications.

Table 8. Comparison of the proposed method and other studies.

Study	Score
Our method	0.22 (mAP@0.5)
H. Huang, et al. (2021) [6]	0.21 (mAP@0.5)
Leader in Kaggle competition	0.31 (mAP@0.4)

In H. Huang, et al.'s study (2021) [6], the same dataset collected by VinBigData [3] was used, but with full quality DICOM-format images, while this study utilized lower quality (3x lower) JPEG-format images [17]. In the processing stage of H. Huang's study, the "No finding" class was removed, and images were rescaled to 512x512 pixels as input. The H. Huang, et al'.s study (2021) [6] used the Yolo-v5 [33] model as a baseline. Although the Yolo model is a popular model for object detection, it struggles with the various sizes of abnormalities in each image and new or unusual aspect ratios of images in this type of medical problem. Furthermore, Yolo struggles with imbalanced datasets and small labels inside large labels. The VinBigData Kaggle Competition (2022) [34] published a leaderboard with the winners of the competition, but unfortunately they did not propose any method or publication to explain CXR abnormality detection solution.

Although the proposed study is not among the top performers, the suggested method still has several advantages and proposes new ways to explore and process datasets to generate new datasets with more consistent and homogeneous properties. Firstly, the proposed method analyses the dataset through data science and statistical approaches. Secondly, it generates new ground truths by applying WBF and comparing the labels of doctor to reduce inconsistency in the dataset. Finally, it applies several optimization techniques such as learning rate schedule and different loss functions to boost the model performance.

5. Conclusions

This study has proposed a new method to detect abnormalities in CXR images using the RetinaNet model and applying WBF and statistical analysis to process data for improving the performance of the deep learning model. The performance of the proposed study achieved 0.22 mAP@0.5 for 14 diseases and 0.55 for 5 potential diseases (aortic enlargement, cardiomegaly, ILD, infiltration and nodule/mass). Overall, the achieved results indicate that the proposed method has promising potential for the development of processing-based approaches for the automatic detection of abnormalities in CXR images.

CRediT author statement

Nhan Ngo: Conceptualisation, Methodology, Software, Resources, Writing, Reviewing, Editing; Toi Vo: Reviewing, Editing; Lua Ngo: Supervision, Reviewing, Editing.

ACKNOWLEDGEMENTS

This research is funded by the Vietnam National University - Ho Chi Minh City (VNU-HCM) under grant number NCM2020-28-01.

COMPETING INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this article.

REFERENCES

[1] U.S. National Library of Medicine (2021), "Respiratory failure", https://medlineplus.gov/lungdiseases.html, accessed 15 June 2021.

[2] American Lung Association (2022), "Estimated prevalence and incidence of lung disease", https://www.lung.org/research/trends-in-lung-disease/prevalence-incidence-lung-disease, accessed 15 June 2022.

[3] H.Q. Nguyen, K. Lam, L.T. Le, et al. (2022), "VinDr-CXR: An open dataset of CXRs with radiologist 's annotations", *Sci. Data*, **9(1)**, DOI: 10.1038/s41597-022-01498-w.

[4] M.T. Islam, M.A. Aowal, A.T. Minhaz, et al. (2017), "Abnormality detection and localization in chest x-rays using deep convolutional neural networks", *arXiv Preprint*, DOI: 10.48550/arXiv.1705.09850.

[5] A.P. Bradley (1997), "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition*, **30(7)**, pp.1145-1159, DOI: 10.1016/S0031-3203(96)00142-2.

[6] H. Huang, Y. Long, Y. Wei (2021), *CXR Abnormalities Detection*, Stanford University, 6pp.

[7] J. Redmon , S. Divvala, R. Girshick, et al. (2016), "You only look once: Unified, real-time object detection", *arXiv Preprint*, DOI: 10.48550/arXiv.1506.02640.

[8] Frankel Cardiovascular Center (2022), "Enlarged aorta", https://www.umcvc.org/conditions-treatments/enlarged-aorta, accessed 4 May 2022.

[9] Mayo Clinic (2022), "Enlarged heart", https://www.mayoclinic.org/diseases-conditions/enlarged-heart/symptoms-causes/syc-20355436, accessed 4 May 2022.

[10] Mayo Clinic (2017), "Interstitial lung disease", https://www.mayoclinic.org/diseases-conditions/interstitial-lung-disease/symptoms-causes/syc-20353108, accessed 21 July 2021.

[11] B. Vahid, P.E. Marik (2008), "Infiltrative lung diseases: Complications of novel antineoplastic agents in patients with hematological malignancies", *Can. Respir J.*, **15(4)**, pp.211-216, DOI: 10.1155/2008/305234.

[12] MediLexicon International (2022a), "What is a lung nodule?", *Medical News Today*, https://www.medicalnewstoday.com/articles/317531, accessed 24 May 2022.

[13] T.Y. Lin, P. Goyal, R. Girshick, et al. (2017), "Focal loss for dense object detection", *Proceedings of The IEEE International Conference on Computer Vision*, pp.2980-2988.

[14] K. He, X. Zhang, S. Ren, et al. (2016), "Deep residual learning for image recognition", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, DOI: 10.1109/CVPR.2016.90.

[15] R. Girshick (2015), "Fast r-cnn", *Proceedings of The IEEE International Conference on Computer Vision*, pp.1440-1448.

[16] R. Solovyev, W. Wang, T. Gabruseva (2021), "Weighted boxes fusion: Ensembling boxes from different object detection models", *Image and Vision Computing*, **107**, DOI: 10.1016/j.imavis.2021.104117.

[17] Raddar (2020), "VinBigData competition JPG data 3X downsampled", *Kaggle*, https://www.kaggle.com/raddar/vinbigdata-competition-jpg-data-3x-downsampled, accessed 31 December 2020.

[18] Johns Hopkins Medicine (2021), "Atelectasis", https://www.hopkinsmedicine.org/health/conditions-and-diseases/atelectasis, accessed 8 August 2021.

[19] MediLexicon International (2022b), "What is calcification, what does it mean, and is it serious?", *Medical News Today*, https://www.medicalnewstoday.com/articles/calcification, accessed 8 May 2022.

[20] R. Kampalath (2022), "What causes lung consolidation?", *Verywell Health*, https://www.verywellhealth.com/lung-consolidatio-5221270#:~:text=Consolidation%20occurs%20when%20the%20normal,X%2Dray%20or%20CT%20scan, accessed 12 May 2022.

[21] C. Schaefer-Prokop (2008), "Pulmonary opacity, extensive pattern", *Encyclopedia of Diagnostic Imaging*, pp.1555-1556, DOI: 10.1007/978-3-540-35280-8_2074.

[22] B. Jany, T. Welte (2019), "Pleural effusion in adults-etiology, diagnosis, and treatment", *Dtsch Arztebl Int.*, **116(21)**, pp.377-386, DOI: 10.3238/arztebl.2019.0377.

[23] L. Molinari (2022), "Pleural thickening", *Mesothelioma*, https://www.mesothelioma.com/asbestos-cancer/pleural-thickening/, accessed 1 March 2022.

[24] P. Zarogoulidis, I. Kioumis, G. Pitsiou, et al. (2014), "Pneumothorax: From definition to diagnosis and treatment", *J. Thorac. Dis.*, **6**, Suppl. 4, pp.S372-S376, DOI: 10.3978/j.issn.2072-1439.2014.09.24

[25] Mayo Clinic (2018), "Pulmonary fibrosis", https://www.mayoclinic.org/diseases-conditions/pulmonary-fibrosis/symptoms-causes/syc-20353690, accessed 6 March 2018.

[26] J. Hosang, R. Benenson, B. Schiele (2017), "Learning non-maximum suppression", *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pp.4507-4515, DOI: 10.1109/CVPR.2017.685.

[27] N. Bodla, B. Singh, R. Chellappa, et al. (2017), "Soft-nms improving object detection with one line of code", *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pp.5561-5569, DOI: 10.1109/ICCV.2017.593.

[28] J. Brownlee (2020), "Transfer learning in Keras with computer vision mode", *Machine Learning Mastery*, https://machinelearningmastery.com/how-to-use-transfer-learning-when-developing-convolutional-neural-network-models/, accessed 18 August 2020.

[29] D.R. Cox (1958), "The regression analysis of binary sequences", *Journal of The Royal Statistical Society, Series B (Methodological)*, **20(2)**, pp.215-242, DOI: 10.1111/j.2517-6161.1958.tb00292.x.

[30] S.G.K. Patro, K.K. Sahu (2015), "Normalization: A preprocessing stage", *IARJSET*, DOI: 10.17148/IARJSET.2015.2305.

[31] Y. Wu, A. Kirillov, F. Massa, et al. (2019), "Detectron 2", https://github.com/facebookresearch/detectron2, accessed 8 August 2021.

[32] S.L. Smith, P.J. Kindermans, C. Ying, et al. (2018), "Don't decay the learning rate, increase the batch size", *ICLR 2018*, DOI: 10.48550/arXiv.1711.00489.

[33] R. Couturier, H.N. Noura, O. Salman, et al. (2021), "A deep learning object detection method for an efficient clusters initialization", *arXiv Preprint*, DOI: 10.48550/arXiv.2104.13634.

[34] Vingroup Big Data Institute (2022), "Leader board VinBigData chest X-ray abnormalities detection", *Kaggle*, https://www.kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection/leaderboard, accessed 8 August 2022.