ỨNG DỤNG MÔ HÌNH YOLO ĐỂ CHẨN ĐOÁN CÁC LOẠI BỆNH VỀ PHỔI DỰA TRÊN ẢNH X-QUANG LỒNG NGỰC

Lê Phương Thảo[1,*], **Nguyễn Quốc Dương**[1], **Đỗ Văn Tuấn**[1]

[1]*Học viên cao học, Khoa học dữ liệu ứng dụng, Trường Đại học Quy Nhơn*

[*] *Tác giả liên hệ chính. Email: phthao1275@gmail.com*

TÓM TẮT

Chụp X-Ray phổi là xét nghiệm cận lâm sàng quan trọng nhất trong chẩn đoán các bệnh lý về phổi. Thông qua kỹ thuật này, bác sĩ sẽ nhanh chóng phát hiện các bất thường trên lồng ngực và lên kế hoạch điều trị bệnh hiệu quả nhất; đồng thời theo dõi tình trạng hồi phục của người bệnh nếu đang trong thời gian điều trị. Mục đích của nghiên cứu này là sử dụng mô hình YOLO để phát hiện các bất thường trên ảnh X-Ray lồng ngực nhằm hỗ trợ bác sĩ chẩn đoán bệnh. Mô hình YOLOv5 và YOLOv7 đã được thử nghiệm để tìm ra mô hình có hiệu suất cao hơn. Chúng tôi sử dụng bộ dữ liệu ảnh X-Ray lồng ngực được cung cấp bởi Vingroup Big Data Institute. Kết quả cho thấy mô hình YOLOv7 có độ chính xác cao hơn so với YOLOv5 dựa trên các thước đo đánh giá như mAP@.5, mAP@.5:mAP@.95, precision và recall. Đồng thời, chúng tôi xây dựng một API để thử nghiệm trực quan hơn cho bài toán này.

Từ khóa: *chest x-ray, x-ray, yolov7, yolov5, vinbigdata*

YOLO MODEL APPLICATION TO DIAGNOSE LUNG DISEASES BASED ON CHEST X-RAY IMAGE

Phuong-Thao LE[1,*], Quoc-Duong NGUYEN[1], Van-Tuan DO[1]

[1] *Master's Student, Applied Data Science, Quy Nhon University*

[*] *Corresponding author. Email: phthao1275@gmail.com*

ABSTRACT

Chest X-ray is the most important laboratory test in the diagnosis of lung diseases. Through this technique, the doctor will quickly detect abnormalities on the chest and plan the most effective treatment; At the same time, monitor the patient's recovery status during treatment. The purpose of this study is to use the YOLO model to detect abnormalities on chest X-ray images to assist doctors in the diagnosis. Models YOLOv5 and YOLOv7 were tested to find the model with higher performance. We use the chest X-Ray image dataset provided by Vingroup Big Data Institute. The results show that the YOLOv7 model has higher accuracy than YOLOv5 based on evaluation measures such as mAP@.5, mAP@.5:mAP@.95, precision and recall. At the same time, we built an API for more intuitive testing of this problem.

Key words: *chest x-ray, x-ray, yolov7, yolov5, vinbigdata*

1. Introduction

Artificial Intelligence (AI) is having a huge impact on almost all fields of science, especially the medical field. Today, AI is being applied for the detection, diagnosis and early intervention of human diseases. Some studies have used AI to diagnose diseases such as diabetes[1], corneal pellets[2], diagnose liver cancer[3], predict cerebral palsy[4], ... This is a step forward in the detection and treatment of diseases. AI also makes it easier for healthcare systems to shift focus and resources from cure to prevention.

Recently, lung diseases are increasing and alarming. It can be said that chest X-ray is the most commonly used technique in clinical examination. Chest X-ray is a technique to help check, screen and detect abnormalities in different locations of the lungs. Through the results of chest X-ray, along with other necessary tests, will suggest the diagnostic of lung-related diseases, thereby giving the most appropriate treatment regimen for the patient. However, radiologists face many challenges every day, one of which is taking and diagnosing chest radiographs. There are currently no specifications for their placement on images, which sometimes leads to inexplicable results. In practice, a complex work of reasoning that often requires careful observation and a good knowledge of anatomical, physiological, and pathological principles. These factors increase the difficulty in developing a consistent and automated technique for reading chest X-ray images, while considering all common thoracic diseases[2]. Stemming from these difficulties, a number of studies have used AI to assist physicians in diagnosing coordination-related diseases through chest X-ray images[5–9]. This problem is collectively known as object detection and classification. This is a common problem in computer vision. The goal of object detection is to identify and classify objects that exist in the image.

For the work of object detection in general and abnormal classification on chest X-ray images in particular, YOLO is one of the highly appreciated models. YOLO is a CNN network model used for object detection, recognition and classification. YOLO is created from the combination of convolutional layers and connected layers. In which, the convolutional layers will extract the features of the image, and the full-connected layers will predict that probability and the coordinates of the object. One of the advantages that YOLO brings is that it only uses the entire image information once and predicts the entire object box containing the objects, the model is built in an end-to-end fashion should be trained. purely by gradient descent. Anyone who has ever worked in object recognition has heard of YOLO. The first version[10] came out in 2016 and so far there have been many improved versions of YOLO. In improved versions of YOLO, YOLOv5 and YOLOv7[11] have yielded amazing results and are classed as state-of-the-art models of object detection. From the above reasons, in this paper, we apply the YOLO model to detect abnormalities on chest X-ray images in order to provide doctors with useful information to support the diagnosis. We compare the performance for the YOLOv5 and YOLOv7 models to choose the model with the better results. For the implementation, the dataset “VinBigData Chest X-ray Abnormalities Detection” is provided by Vingroup Big Data Institute and is available at kaggle[1].

2. Related works

Currently, studies on the detection and classification of diseases on chest X-ray images are promoted. In 2017, Pranav Rajpurkar and associates developed an algorithm that can detect pneumonia from chest x-ray films with a level far beyond that of practicing radiologists[12]. Recently, the world has faced the covid-19 pandemic and caused a wide range of respiratory problems, ranging from mild to critical or fatal. Several studies have collected and constructed different datasets on chest X-ray images for the purpose of studying the effects of Covid-19 on the lungs. New Generation Computing magazine has published the XCOVNet[13] model to help classify chest X-ray images for early detection of COVID-19. In particular, the Radiological Society of North America organized the RSNA Pneumonia Detection Challenge to build an algorithm to detect signs of pneumonia on chest X-ray images. Therefore, building AI models to assist specialists in making diagnoses is extremely necessary and has practical significance.

Stemming from that practice, we need a good enough dataset to support building machine learning models. There are many published chest radiograph data sets including ChestX-ray8, ChestX-ray14[14], Padchest[15], CheXpert[16] and MIMIC-CXR[17]. Among of those, ChestX-ray14 an extended version of ChestX-ray8, was released by the US National Institutes of Health (NIH). Most of these datasets have disease type labels but do not specify their location on the X-ray film. This hinders the application of machine learning algorithms to detect and localize breast abnormalities. For that reason, Vingroup Big Data Institute released 18,000 images that were manually annotated by a total of 17 experienced radiologists with 22 local labels of rectangles surrounding abnormalities and 6 global labels of suspect diseases[10]. The data set consists of a training set of 15,000 and a test set of 3,000. Details of the data collection and labeling procedures are detailed in the study[10].

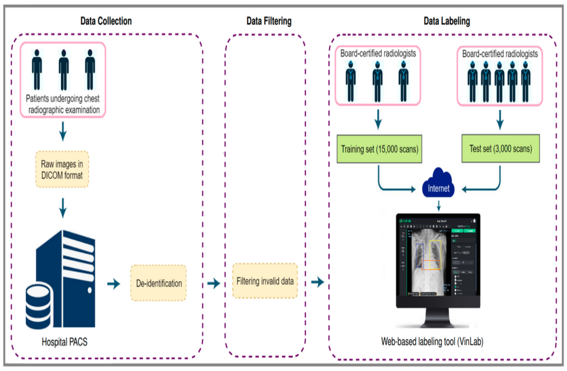
Data Collection
Data Filtering
Data Labeling
Patients undergoing chest radiographic examination
Raw images in DICOM format
Hospital PACS
De-identification
Filtering invalid data
Board-certified radiologists
Board-certified radiologists
Training set (15,000 scans)
Test set (3,000 scans)
Internet
Web-based labeling tool (VinLab)

Fig. 1: *The flow of creating VinDr-CXR dataset*

Lung diseases often present as an area on a chest X-ray. There are many machine learning models to help detect pneumonia on X-ray images based on many different object detection models such as CNN, Mask RCNN, Faster R-CNN, YOLOv3 and they both gave pretty good results in some studies on several chest radiograph data sets. In particular, the YOLOv5 version is good and superior to all previous versions. Recently, the emergence of YOLOv7 has become prominent when it defeats all object detection models. In this paper, we reported the performance of the YOLOv5 and YOLOv7 models based on the VinBigdata dataset.

3. Architecture of the YOLO model

The first YOLO version got a lot of attention for outperforming state-of-the-art object detectors such as DPM, Fast R-CNN and Faster R-CNN. Also the simplicity of YOLO was quite refreshing for using only 24 convolutional layers and 2 fully connected layers only, unlike the sophisticated R-CNN based architectures. This first successful initial release of YOLO was a launch pad for development of subsequent YOLO versions, so that a new class of object detectors was born. This class of architecture contains 3 main components: backbone, neck and head. The backbone has the role to extract features from the image, while the neck tries to aggregate the features from different stages from the backbone. In the end the head does the prediction. Usually the backbone has a quite simple build like ResNet-101 and Darknet-53. For feature aggregation SPP and FPN are a popular choice.

There are now countless versions of YOLO developed by different research groups. Of these, YOLOv5 by ultralytics because it quickly exhibits good performance, is written in a popular framework

[1] https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection

like PyTorch, and provides support for logging and debugging tools. Other versions, such as YOLOv4, YOLOX and YOLOR are some interesting candidates, but each has its drawbacks. YOLOv4 has similar performance to YOLOv5, however it has longer training time and very rudimentary logging system (does not support tools like Tensorboard). Furthermore, the YOLOv7 algorithm is making a big splash in the computer vision and machine learning communities. The official YOLOv7 article titled "YOLOv7: Trainable Free Bags Set the Most Advanced Technology for Real-Time Object Detectors" was published in July 2022 by Chien-Yao Wang, Alexey Bochkovskiy and Hong-Yuan Mark Liao. The YOLOv7 research paper became extremely popular in just a few days. The YOLOv7 algorithm surpasses all object detection models and previous YOLO versions in both speed and accuracy. It saves on hardware costs and can be trained much faster on small data sets without any pre-trained weights. We will compare the performance of these two models based on the chest-xray dataset.

4. Evaluation metrics

YOLO detectors generally use a distinctive format that consists of 5 mandatory and 1 optional values: *x, y, w, h, class, conf (optional)*. The first 2 values *(x, y)* are the coordinates of the center of the bounding box normalized with respect to the width and height of the image. The next values *w,h* are the width and height of the bounding box, also normalized with respect to the width and height of the image. The next value *class* is the class id. Finally, the value *conf* is used only for detected bounding box to store the confidence of the detection. If *conf* is not given, then it can be assumed that the bounding box is a ground truth and not a detection. Each version might have a slight variation of this notation format, but conceptually they are all the same.

In this section, we provide some metrics to evaluate the model. The first key concept is the Intersection over Union (IoU), which expresses, how good the overlap between a detection and ground truth bounding box is. The IoU divides the overlap area by the union area, as visualized in Figure 2.

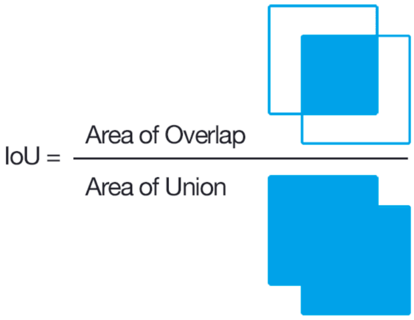

Fig. 2: *IoU*

The second key concept is the classification of detections in true positive, false positive, false negative. This can be done by taking each detection and seeing, if it has a high IoU with a ground truth bounding box. For a high IoU (i.e. $IoU \geq 0.5$) , then the detection is considered a true positive, else it's a false postive. If a ground truth bounding box is not overlapped by any detections, then the respective missing detection is classified as false negative. The metrics of precision and recall can be defined:

$$p(c,i) = \frac{TP}{TP+FP},$$

$$r(c,i) = \frac{TP}{TP+FN},$$

where:

- TP = true positives count,
- FP = false positives count,
- FN = false negatives count,
- c = threshold for the confidence of the detections,
- i = threshold for the IoU values.

The c parameter can be adjusted to improve the recall or precision, which makes sometimes the comparison between two models unclear. Setting the same $conf$ in order to compare two models seems a good approach, but this naive approach might not work, if the models have very different distributions of the confidence values e.g. one model is overconfident and the other one is underconfident. In YOLOv5 and YOLOv7, this problem was solved by calculating for each model the precision and recall based on the confidence that maximizes the F1-score:

$$F1(c,i) = 2 * \frac{p(c,i) * r(c,i)}{p(c,i) + r(c,i)}.$$

This way a manipulation of the threshold of the confidence values can be avoided and therefor a fairer comparison between models is guaranteed.

One caveat of this approach is the assumption that precision and recall have the same importance. Another caveat of this approach is that it reduces the performance of the model to a single threshold value, but the mean average precison (mAP) can be used to express a more general performance.

As the name suggests, the mAP is based on the average precision (AP), which is calculated indivudually for each class. The AP for a class C and a IoU threshold i can be calculated as follows:

$$AP_C(i) = \sum_c (r(c_n, i) - r(c_{n-1}, i))p(c_n, i).$$

where c_n are ordered values of confidence thresholds. Finally the mAP is expressed as:

$$mAP(i) = \sum_{C \in classes} AP_C(i).$$

The IoU threshold can be obviously be lowered to increase the mAP, but usually models are compared for at a threshold of 0.5. A final metric is the average mAP, usually denoted as $mAP@[0.5 : 0.95]$. This is just the average mAP for IoU thresholds ranged in 0.5 to 0.95 with a 0.05 step size.

5. Experimental Results

5.1. Data Processing

The dataset used consisted of 18000 thoracic x-ray image files, divided into 2 sets of training and testing with the corresponding number of images 15000 and 3000. Each image is read by leading radiologists in Vietnam. For training data, each scan was independently labeled by three doctors. Meanwhile, for the evaluation data, each scan was labeled by a panel of 5 doctors. We used a dataset that was converted from DICOM to jpg format and kept the original size of the image[2]. The dataset includes the following labels: 0 - Aortic enlargement; 1 – Atelectasis; 2 – Calcification; 3 – Cardiomegaly; 4 – Consolidation; 5 – ILD; 6 – Infiltration; 7 - Lung Opacity; 8 – Node/Mass; 9 - Other lesion; 10 - Pleural effusion; 11 - Pleural thickening; 12 – Pneumothorax; 13 - Pulmonary fibrosis; 14 - "No finding". Labels correspond to lung diseases, details of which can be found at[10]. Figure 3 shows us that the number of each label in the dataset has a very large imbalance. Figure 4 illustrates some of its items.

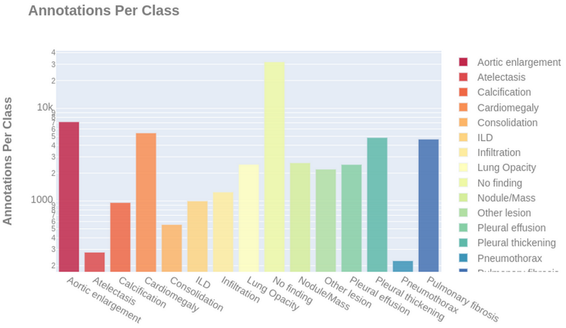

Fig. 3: *Annotations per class*

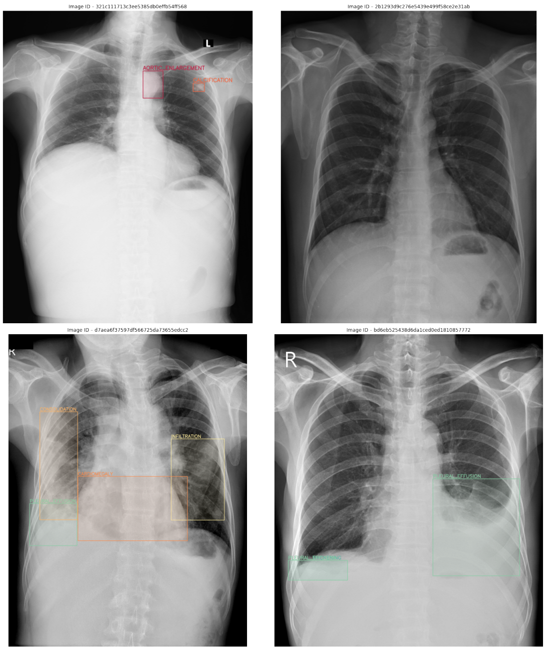
Image ID - 321c111713c3ee5385db0effb54ff568
L
AORTIC_ENLARGEMENT
CALCIFICATION
Image ID - 2b1293d9c276e5439e499f58ce2e31ab
Image ID - d7aea6f37597df566725da73655edcc2
R
CONSOLIDATION
INFILTRATION
CARDIOMEGALY
PLEURAL_EFFUSION
Image ID - bd6eb525438d6da1ced0ed1810857772
R
PLEURAL_EFFUSION
PLEURAL_THICKENING

Fig. 4: *Some illustrations of the data*

As we introduced in section 2, each image is independently labeled by 3 doctors, so there can be duplicate predictions on the same area in the image. This will appear to have two predictions with the same label with 2 overlapping prediction boxes with $IOU \geq 5$. Both of these boxes are correct predictions because they were annotated independently by 2 radiologists. The model returns 2 output boxes corresponding to the labels. In this case, we use NMS - Non-maximum Suppression to remove 1 cell in the output even though it is the correct result. When using YOLO, and for train data, merging the boxes removes the doctor's attributes and takes the average of the duplicate boxes. The decision whether to merge the boxes will be based on the IoU - Intersection over Union index with a good threshold i.e. $IOU \geq 5$. Figure 5 shows the bounding boxes before and after pooling. Then we normalize the box

[2] https://www.kaggle.com/datasets/awsaf49/vinbigdata-original-image-dataset

coordinates from 0 to 1.

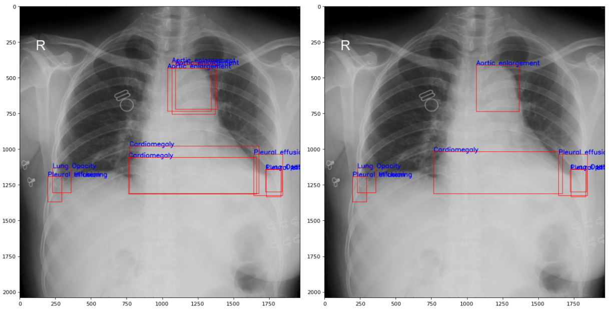

Fig. 5: *Take the average of the coordinates using IOU*

5.2. Model Training

Two models were trained with the same number of epochs of 200 and run separately to monitor the effects of the models under the same conditions: google colab pro with GPU A100-SXM4-40GB. They have been tested on identical test/test datasets. We use the mechanism to stop training early when the last 100 epochs have not improved results. The YOLOv5 model stopped training on its own at epoch 171 after 2,938 hours of training, and the YOLOv7 model was trained with 200 epochs after 4,853 hours. From Figure 6 (left), it seems that the best results observed at epoch 70 for YOLOv5.

Based on the training results of the YOLOv7 model in Figure 6 (right), we see that the loss function of val objectness has not converged well. This may be because the learning rate is still quite large. In the future, we will reduce the learning rate for the YOLOv7 model to better converge the model and avoid overfiting.

Table 1 summarizes the results of the two models. We see that the mAP@.5, mAP@.5:mAP@.95, Precision and Recall measures of the YOLOv7 model are all higher than those of the YOLOv5 model. This may initially indicate that the YOLOv7 model exhibits better performance than the YOLOv5 model. In terms of training interval for each epoch, the YOLOv5 model gives faster training results than YOLOv7.

Tab. 1: *The result of the experiments*

Result	mAP@.5	mAP@.5:mAP@.95	Precision	Recall	Training time per epoch (mins)
YOLOv5	0.377	0.182	0.443	0.397	0.9
YOLOv7	0.421	0.208	0.466	0.461	1.24

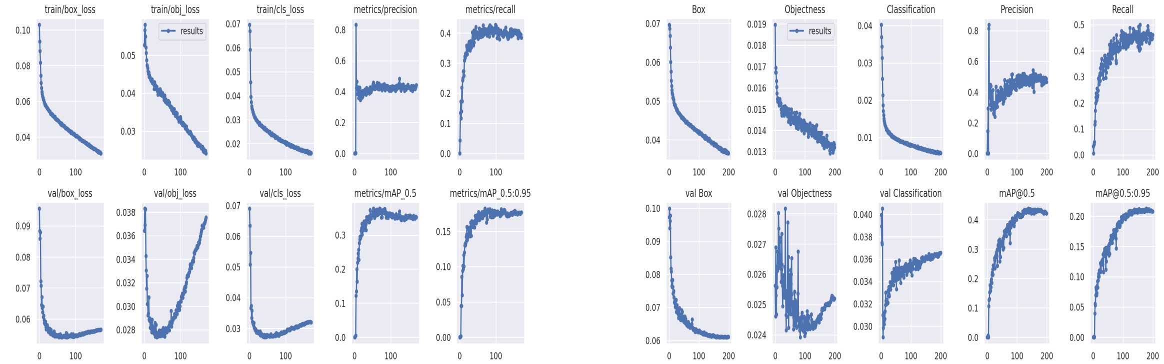

Fig. 6: *Training results of YOLOv5 and YOLOv7 models*

Based on results of training, we choose the YOLOv7 model and build an API interface written in flask library to demo the results of predicting a new image. Our entire implementation is shown in Figure 7. The API will take as input a chest X-ray image and return abnormal detection results including disease label, bounding box position and diagnostic confidence.

The code captures the entire process of data analysis, data processing, model training, and API implementation from the results of the best model found at `https://github.com/nguyenquocduongqnu/Chest-X-Ray-Diagnosis`.

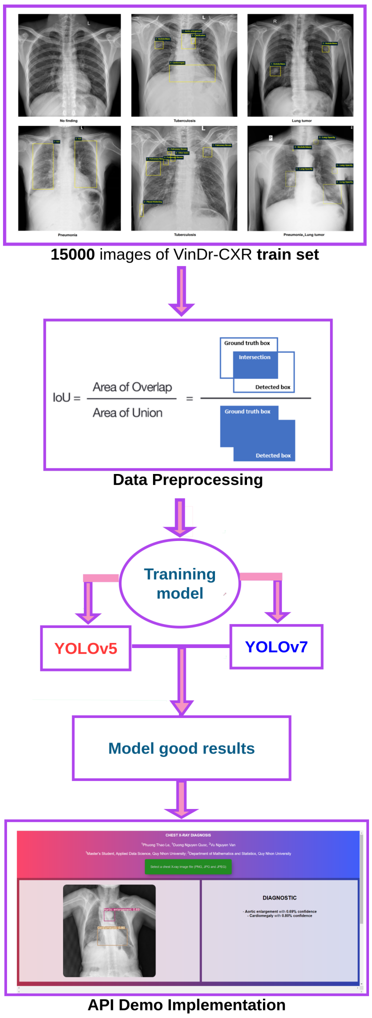

Fig. 7: *Workflow*

6. Conclusion

In this study, the YOLOv7 model gave better results than YOLOv5 based on performance evaluation measures. We have implemented an API application to test the results of the problem. However, the mAP@.5 on both models is still not high. In the future, we need to adjust the learning rate for the YOLOv7 model lower so that the model can learn better, consider and handle the data imbalance for better diagnostic results of the model.

Acknowledgment

Phuong-Thao LE was funded by Vingroup JSC and supported by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Institute of Big Data, code *VINIF.2021.ThS.QN.05*. The authors thank Dr. Van-Vu NGUYEN for his helpful remark as well as many valuable comments related to our results.

References

1. N. P. Tigga, S. Garg, Prediction of Type 2 Diabetes using Machine Learning Classification Methods, *Procedia Computer Science*, **2020**, *167*, 706–716, international Conference on Computational Intelligence and Data Science, URL https://www.sciencedirect.com/science/article/pii/S1877050920308024.

2. C. B. Kim, G. W. Armstrong, Characterizing Infectious Keratitis Using Artificial Intelligence, *International Ophthalmology Clinics*, **2022**, *62*(2), 41–53.

3. P. S. Maclin, J. Dempsey, A neural network to diagnose liver cancer, *IEEE International Conference on Neural Networks*, **1993**, *3*, 1492–1497.

4. H. H. Ramadhan, Cerebral Palsy Prediction using CNN Depending on MRI Images of the Brain, *Journal of Optoelectronics Laser*, **2022**, *41*(8).

5. K. Kallianan, J. Mongan, S. A. and, How far have we come? Artificial intelligence for chest radiograph interpretation, *Clin Radiol*, **2019**, *74*(5), pp. 338–345.

6. A. Bhandary, Deep-learning framework to detect lung abnormality – a study with chest X-Ray and lung CT scan images, *Pattern Recogn Lett*, **2020**, *129*, pp. 271–278.

7. Nasrullah, J. Sang, M. S. Alam, H. Xiang, Automated detection and classification for early stage lung cancer on CT images using deep learning, In: *Pattern Recognition and Tracking XXX*, Vol. 10995 (edited by M. S. Alam), International Society for Optics and Photonics, SPIE, p. 109950S, URL https://doi.org/10.1117/12.2520333.

8. H. Behzadi-khormouji, H. Rostami, S. Salehi, T. Derakhshande-Rishehri, M. Masoumi, S. Salemi, A. Keshavarz, A. Gholamrezanezhad, M. Assadi, A. Batouli, Deep learning, reusable and problem-based architectures for detection of consolidation on chest X-ray images, *Computer Methods and Programs in Biomedicine*, **2020**, *185*, 105162, URL https://www.sciencedirect.com/science/article/pii/S0169260719306960.

9. V. Chouhan, S. K. Singh, A. Khamparia, D. Gupta, P. Tiwari, C. Moreira,

R. Damaševičius, V. H. C. de Albuquerque, A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images, *Applied Sciences*, **2020**, *10*(2), URL `https://www.mdpi.com/2076-3417/10/2/559`.

10. H. Q. Nguyen, et al., VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations., *Scientific data*, **2022**, *9*, 429.

11. C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, *https://arxiv.org/pdf/2207.02696.pdf*, **2022**, .

12. P. Rajpurkar, et al., CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, *CoRR*, **2017**, *abs/1711.05225*, `1711.05225`, URL `http://arxiv.org/abs/1711.05225`.

13. V. Madaan, A. Roy, C. Gupta, P. Agrawal, A. Sharma, C. Bologa, R. Prodan, XCOVNet: Chest X-ray Image Classification for COVID-19 Early Detection Using Convolutional Neural Networks, *New Gener. Comput.*, **2021**, *39*(3-4), 271–278.

14. Wang, Xiaosong and Peng, Yifan and Lu, Le and Lu, Zhiyong and Bagheri, Mohammadhadi and Summers, Ronald M., *ChestX-ray: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases*, chapter Large-Scale Data Mining and Data Synthesis, Springer International Publishing, Cham, pp. 369–392, URL `https://doi.org/10.1007/978-3-030-13969-8_18`.

15. A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, PadChest: A large chest x-ray image dataset with multi-label annotated reports, *Medical Image Analysis*, **2020**, *66*, 101797, URL `https://www.sciencedirect.com/science/article/pii/S1361841520301614`.

16. J. Irvin, et al., CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, In: *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, AAAI Press, pp. 590–597.

17. A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. ying Deng, R. G. Mark, S. Horng, MIMIC-CXR: A large publicly available database of labeled chest radiographs, *CoRR*, **2019**, *abs/1901.07042*, URL `http://arxiv.org/abs/1901.07042`.