Deep Learning based Automated Chest X-ray Abnormalities Detection

Vraj Parikh[1], Jainil Shah[1], Chintan Bhatt[1], Juan M Corchado[2], and Dac-Nhuong Le[3]

[1] U & P U. Patel Department of Computer Engineering, Charotar University of Science and Technology (CHARUSAT), Gujarat, India
chintanbhatt.ce@charusat.ac.in

[2] BISITE Research Group, University of Salamanca, 37007 Salamanca, Spain
corchado@usal.es

[3] Faculty of Information Technology Haiphong University, Haiphong, Vietnam
Nhuongld@dhhp.edu.vn

Abstract. Abnormalities related to the chest are a fairly common occurrence in infants as well as adults. The process of identifying these abnormalities is relatively easy but the task of actually classifying them into specific labels pertaining to specific diseases is a much harder endeavour. The exponential increase in COVID-19 patients is overwhelming healthcare systems across the world. With limited testing kits, it is impossible for every patient with respiratory illness to be tested using conventional techniques. Thus in such dire circumstances, we propose the use of modern deep learning techniques to help in the detection and classification of a number of different thoracic abnormalities from a chest radiograph. The goal is to be able to automatically identify and localize multiple points of interest in a provided chest X-ray and act as a second level of certainty after the radiologists. On our publically available chest radiograph dataset, our methods resulted in a mean average precision of 0.246 for the detection of 14 different thoracic abnormalities.

Keywords: Deep learning· chest x-rays· pulmonary diseases· object detection· chest abnormalities.

1 Introduction

A chest X-ray is an imaging test that utilises low doses of radiation in short blasts to create images of the inside of a patient's chest. They are the most prescribed medical images. Any improvement in the process of reading and reporting on these images can have a meaningful impact on the radiology workflow. The availability of various open-sourced datasets for the same has created growing academic research interest in AI model development for disease detection in chest X-rays. Researchers have been successful in tagging X-ray images with global labels (say ”Consolidation”), but not as successful in mapping it to an area of the lung. The result is that building interpretable AI models for chest X-ray disease detection remains difficult and an open research problem [14].

Notable public datasets of CXR include ChestX-ray8, ChestX-ray1410, Padchest11, CheXpert3 and MIMIC-CXR12. ChestX-ray14 [16] , an extended version of ChestX-ray8, was released by the US National Institutes of Health (NIH), containing over 112,000 CXR scans from more than 30,000 patients. Without being manually annotated, this dataset poses significant issues related to the quality of its labels. CheXNet [13] is one such simple model architecture trained on ChestX-ray14 giving a much better performance. In the past few years because of AI and Deep learning, significant advancement has been made in the medical science which helps doctors to diagnose diseases early and easily which was tedious and time-consuming not so long ago. Hence in our study with the help of the chest radiographs dataset, we provide a method for classifying and localizing various different thoracic abnormalities using various convolutional neural network model architectures.

It can work as a pair of second eyes to assist and validate the base finding that a radiologist might make. The model will be able to automatically identify and localize multiple points of interest in the data provided to it and may be integrated in conjunction with other similar operations to increase the overall efficiency in medical diagnosis and procedures. This project also provides understanding of the social, gender and environmental influences.

2 Related Works

Since the research was a part of a Kaggle Competition there were various other teams that also collaborated on the dataset and produced commendable results on the dataset. The team having notebook name “VinBigData complete pipeline” scored 0.286 and used a 2-class classifier complete pipeline. Different things done by them includes, using “timm” (pytorch-image-models, providing a lot of popular SoTA CNN models with pretrained weights). They also used “Pytorch Ignite” (Training/Evaluation abstraction framework on top of pytorch) and “Pytorch pfn extras” which is used to add more feature-rich functionality on Ignite Useful for logging, printing, evaluating, saving the model, scheduling the learning rate during training.

This has been an area of research interest since the 1960s when the first papers describing an automated abnormality detection system on CXR images were published [10]. In recent years, deep learning has become the technique of choice for image analysis tasks and made a tremendous impact in the field of medical imaging. The CXR research community has benefited from the publication of numerous large labeled databases in recent years, predominantly enabled by the generation of labels through automatic parsing of radiology reports. This trend began in 2017 with the release of 112,000 images from the NIH clinical center. In 2019 alone, more than 755,000 images were released in 3 labeled databases (CheXpert, MIMIC-CXR, PadChest) [3, 9, 10].

Notable public datasets of CXR include ChestX-ray8, ChestX-ray1410, Padchest11, CheXpert3 and MIMIC-CXR12. The US National Institutes of Health (NIH), containing over 112,000 CXR scans from more than 30,000 patients, released ChestX-ray14, an extended version of ChestX-ray8. Without being manually annotated, this dataset poses significant issues related to the quality of its labels. CheXNet is one such simple model architecture trained on ChestX-ray14 giving a much better performance.

There have been previous reviews on the field of deep learning in medical image analysis and on deep learning or computer-aided diagnosis for CXR [2]. However, recent reviews of deep learning in chest radiography are far from exhaustive in terms of the literature and methodology surveyed the description of the public datasets available, or the discussion of potential and trends in the field. The literature review in this work includes 295 papers, published between 2015 and 2021, and categorized by application.

In the past few years because of AI and Deep learning, significant advancement has been made in the medical science, which helps doctors to diagnose diseases early, and easily which was tedious and time-consuming not so long ago. Hence, in our study with the help of the chest radiographs dataset, we provide a method for classifying and localizing various different thoracic abnormalities using various convolutional neural network model architectures.

3 Dataset Specifics

The primary objective was to automatically localize and classify 14 types of thoracic abnormalities from chest radiographs. We worked with a dataset consisting of 18,000 scans that have been annotated by experienced radiologists. We can train our model with 15,000 independently-labeled images and will be evaluated on a test set of 3,000 images. These annotations were collected via VinBigData's web-based platform, VinLab. Details on building the dataset can be found in their recent paper [12]. We have used a public dataset hosted on Kaggle in the form of a competition. The dataset is available in DICOM format with a total size of 195 GB. DICOM files are implementations of a predetermined medical file standard that packages the original radiograph images along with the relevant metadata might be useful for visualizing and classifying. Any DICOM medical image consists of two parts—a header and the actual image itself. The header consists of data that describes the image, the most important being patient data. The above specified original dataset can be found here [8] and the resized dataset wherein all the original images have been scaled down to a 256 × 256 format can be found here [18]. All images were labeled for the presence of 14 critical radiographic findings as listed below:

1. Aortic enlargement
2. Atelectasis
3. Calcification
4. Cardiomegaly
5. Consolidation
6. ILD
7. Infiltration
8. Lung Opacity
9. Nodule/Mass
10. Other lesion
11. Pleural effusion
12. Pleural thickening
13. Pneumothorax
14. Pulmonary fibrosis
15. "No finding" observation was intended to capture the absence of all findings above

4 Approach

4.1 Exploratory Data Analysis

We started by performing exploratory data analysis on the available dataset to gather some important insights to the data and to summarize and visualize the main characteristics gathered from the dataset in Fig. 1.

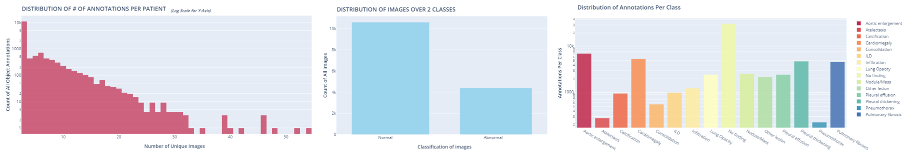

Fig. 1. The distribution of annotations per patient

The `image_id` column contains a Unique IDentifier (UID) that indicates which patient the respective row (object) relates to. As there can be up to three radiologists annotating the same image and potentially multiple objects/bboxes per image, it is possible for a single image UID to occur many times. However, it is stated in the competition data details that there exists only one image for one patient. This means that if a specific `image_id` appears 12 times, that there are 4 objects in the image, and each object was annotated by all three radiologists. The following is a visualization of the total number of objects/annotations per images as per the dataset. We use a log y-scale to essentially normalize the final values for better visualization.

Gained Insights: Each image has at least 3 annotations including the No Finding class i.e. one each from a radiologist. The majority of images (10000 out of 15000) have 3 annotations associated with them. The maximum possible annotations on an image from the dataset are 57. The data columns specify the relevant class ids and class names corresponding to the annotations. There is a possibility to detect 15 classes including the "Normal" class which we just saw is in abundance. To identify the distribution of the different annotated classes over the complete group of training images, we can employ a bar graph. We again use a logarithmic scale on the y-axis.

Gained Insights: Pneumothorax and Atelectatsis are the target abnormality classes with the least number of annotations in the training subset. Aortic Enlargement and Cardiomegaly are the target abnormality classes with the most number of annotations in the training subset. The `rad_id` column of the dataframe is determinant of the particular radiologist that made that particular ground truth annotations. We have already derived that each image has been annotated by 3 distinct radiologists each identified by a unique id. Since there are 17 radiologists in total, each has been denoted by an id between R1 and R17. However one great issue with having a large number of radiologists annotate these radiographs independently of each other, the labels and bounding-boxes can greatly differ from one radiologist to the next even for a single image. Furthermore we can deduce the distribution of different radiologists with respect to amount of annotations they made pertaining to a specific target class. We can figure out how the annotations of different classes were divided among the available radiologists using the following visualizations.

Gained Insights: The three radiologists with the highest number of annotations made have also annotated the majority of the “Abnormal” class or the classes from 1 to 14 denoting a specific target abnormality class. Among the other 14 radiologists, a majority of them have only ever annotated images as “No finding” or “Normal” class. This might allow us to estimate that radiologists other than R8, R9, and R10, are much more likely to annotate images as No finding.

4.2 Data Pre-processing

We started our data pre-processing by creating helper functions to read a DICOM file and to convert the raw DICOM data into a human readable format by applying a VOI LUT transformation on the numpy array. The images were also converted into a 3-channel format with normalization of the values between 0 and 1 to assist in the scaling down of the images prior to providing them as inputs to a deep learning neural network model during the training or evaluation of the model.

We also created a few visualizing modules to help in drawing the annotated boxes on the original image. Here we came across our first issue with the overlapping of various different annotations by different radiologists creating an erratic ground truth boxes compilation which would cause an overflow of false information during the training phase thus bringing down the mean average precision.

Overlapping Annotations : One of the major problems as discussed above is the variation in annotations made by different radiologists on the same image thus leading to contradicting and overlapping bounding boxes which will essentially pass wrong ground truth values to the model thus gaining wrong inference (see Fig. 2).

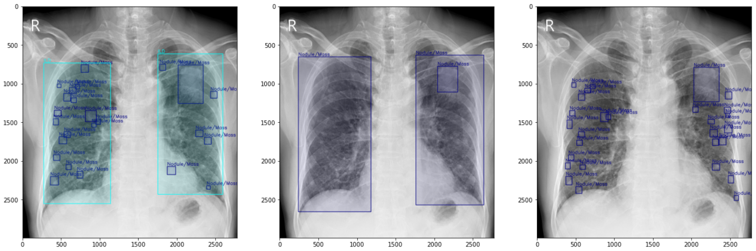

Fig. 2. Bounding boxes grouped by radiologists

One solution to this problem is to employ the IoU function i.e. ratio of intersection over union, over the overlapping bounding boxes (see Fig. 3).

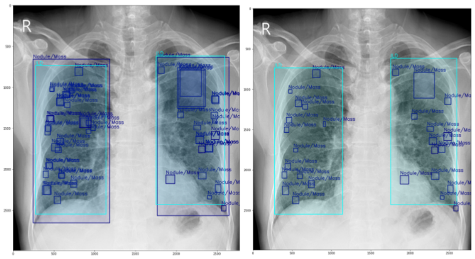

Fig. 3. Original boxes vs. Reduced boxes

Several important parameters for performance evaluation in target detection models are IoU, precision and recall. The IoU serves as a very important function of the performance mean average precision(mAP) calculation of the target detection algorithm. Mean Average Precision is essentially a metric to gauge the performance of object detection tasks, by taking the mean of precision-recall slope calculated over different IOUs between the ground and predicted bounding boxes. We further elaborate on this technique while discussing the results.

The overview of IOU is that when the IoU is calculated for a pair of overlapping bounding boxes and it is greater than a certain threshold, a new bounding box i.e annotation is created from the common area between the two original boxes. This not only reduces the amount of excess bounding boxes in the image but also simplifies the annotations by essentially narrowing down the overall feature detection to a smaller area. The target class of the resulting box is assigned as the target class of the larger box out of the original boxes.

Data Augmentations is a strategy for artificially increasing the quantity and complexity of existing data [4]. We know that training a deep neural network needs a large amount of data to fine-tune the parameters (see Fig. 4).

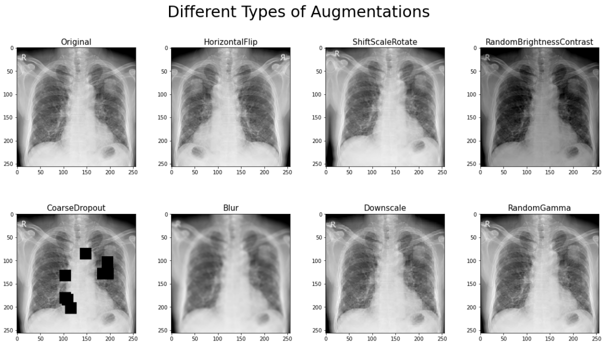

Fig. 4. Data augmentations over dataset

We applied data augmentation techniques on our training dataset by adding modifications to our images by making minor changes, such as flipping, rescaling, rotation, blurring and brightness and contrast. It will increase our training data size and our model will consider each of these small changes as a distinct image, and it will enable our model to learn better and perform well on unseen data.

4.3 Proposed CNN Architectures

ResNet50/ResNet101 : ResNet, short for Residual Network is a specific type of neural network that was introduced in [6]. It was built on the concept of residual functions.

ResNet50 is a 50-layer Residual Network with 26M parameters. The network can take the input image having height, width as multiples of 32 and 3 as channel width. Every ResNet architecture performs the initial convolution and max-pooling using 7 × 7 and 3 × 3 kernel sizes respectively. Afterward, Stage 1 of the network starts and it has 3 Residual blocks containing 3 layers each. The size of kernels used to perform the convolution operation in all 3 layers of the block of stage 1 are 64, 64 and 128 respectively. For deeper networks like ResNet50, ResNet152, etc., bottleneck design is used. For each residual function F, 3 layers are stacked one over the other. The three layers are 1×1, 3×3, 1×1 convolutions. The 1×1 convolution layers are responsible for reducing and then restoring the dimensions. The 3×3 layer is left as a bottleneck with smaller input/output dimensions. Finally, the network has an Average Pooling layer followed by a fully connected layer having 1000 neurons (ImageNet class output) (see Fig. 5).

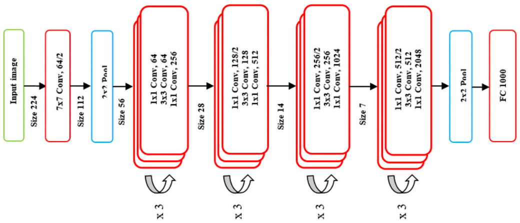

Fig. 5. ResNet50 model architecture

We employed the Detectron2 framework [17] to use a pre-trained ResNet50 model in our experiment and fine-tuned it. The Detectron2 framework contains pytorch high-quality implementations of various modern architectures and their variants. It also has a highly function API which allows both abstraction and customization to create efficient models. We format out augmented data in dataset dictionaries and start the training process. The hyperparameters tuned included the number of ROIs detected in an image, the number of iterations made by the model and the learning rate.

YoloV5 : Unlike the models in the R-CNN family, YoloV5 is a single stage detector model architecture [15]. What this actually means is that instead of proposing a bunch of regions of interest, the model directly runs detection directly over a dense sampling of possible locations. It reframes object detection as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities. This enables the model to decrease its training time and perform inference faster thus becoming more simpler with a slight tradeoff over accuracy and performance. The base model is similar to GoogLeNet with inception module replaced by 1×1 and 3×3 conv layers. The final prediction of shape $S \times S \times (5B + K)$ is produced by two fully connected layers over the whole conv feature map (see Fig. 6).

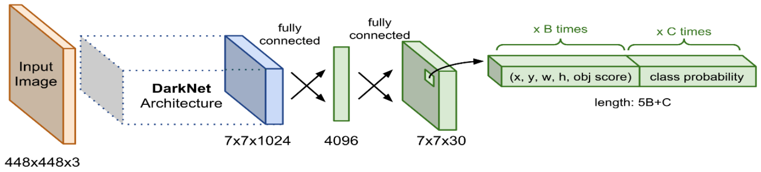

Fig. 6. YoloV5 model architecture

We used the large YoloV5 model architecture after pre-processing the data to obtain annotations in COCO format and dividing the original dataset in specific directories of images and their specific labels in text files with the same name. We hyper-tuned the parameters in the config file and ran the model for 40 epochs with a batch-size of 4.

4.4 2-Class Classifier Pipeline

In order to increase the overall accuracy of the pipeline, we add a 2-class classifier on top of the original model. This classifier can classify an image into two classes i.e. Normal or the "No Finding" target class or Abnormal which is all of the other target classes. Two thresholds are chosen arbitrarily each representing a lower and a higher limit. The final predictions are modified as follows:

- If the prediction on Normal is greater than the maximum threshold, the final image is marked with a "No Finding" prediction and all the previous detected predictions are dropped.
- If the prediction is between the limits, an extra prediction of "No Finding" class is appended to the existing detected predictions.
- If the prediction is less than the minimum threshold, no changes are made to the existing predictions.

EfficientNet is developed as a new baseline network by performing a neural architecture search using the AutoML MNAS framework, which optimizes both accuracy and efficiency (FLOPS) [5]. The resulting architecture uses mobile inverted bottleneck convolution (MBConv), similar to MobileNetV2 and MnasNet, but is slightly larger due to an increased FLOP budget (see Fig. 7).

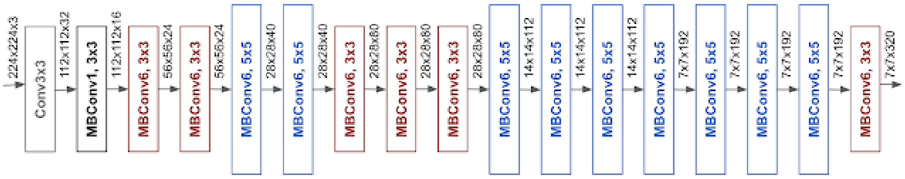

Fig. 7. EfficientNet-B0 Model Architecture

We developed a custom model with a pre-trained Tensorflow and Keras implementation of EfficientNet as the base model with customized dense layers with ReLU and Sigmoid as the activation methods along with a dropout.

ResNet-18 A variation of the Resnet model family includes 4 convolutional layers in each module (excluding the 1 × 1 convolutional layer). Together with the first 7 × 7 convolutional layer and the final fully-connected layer, there are 18 layers in total. Therefore, this model is commonly known as ResNet-18 [7, 1, 11]. We use a pytorch implementation of the ResNet-18 model and fine-tuned the model according to the target data to avoid overfitting.

5 Results

The paper uses mean average precision (mAP) as the evaluation metric. It acquires a higher preference since this is an object detection task primarily utilizing the PASCAL Visual Object Classes (VOC) as bounding box annotations. To obtain the mAP for a set of object predictions, we calculate the ratio of the precision and recall as average precision i.e. the area under the precision-recall curve, and then take a mean of the various iterations of AP thus producing the name mAP.

We experimented with various combinations of model architectures with various data pipelines inclusive of data augmentations and reduced bounding boxes to find the best suitable model for the current task. Our hardware configurations includes 32GB RAM DDR4 2400MHz, CPU Intel i7 9750 H 2.6 GHz and Nvidia GeForce RTX 2070 with Max-Q design 8 Gb. During the training, a validation set was used to validate the model predictions before gathering actual predictions on the unlabeled test dataset. After the various combinations, the final submissions on the test data as validated by Kaggle itself can be used to compare the different architectures as follows:

Table 1. Proposed training pipeline implementations

Model	Data Augmentation	Annotations	2-Class Classifier	mAP
ResNet50 - FPN	No	Original	-	0.136
ResNet50 - FPN	Yes	Original	-	0.044
ResNet50 - FPN	No	Reduced	-	0.092
ResNet101 - FPN	Yes	Original	-	0.067
ResNet50 - FPN	Yes	Reduced	-	0.040
YoloV5	No	Original	-	0.110
YoloV5	Yes	Original	-	0.110
YoloV5	No	Original	EfficientNet-B0	0.198
YoloV5	No	Original	Resnet-18	0.079
ResNet50 - FPN	No	Original	EfficientNet-B0	0.246
ResNet50 - FPN	No	Original	Resnet-18	0.198

(HorizontalFlip : “p”: 0.4; ShiftScaleRotate: “scale_limit”: 0.20, “rotate_limit”: 20, “p”: 0.4).

As seen in the Table 1, the ResNet-50 model architecture performs the best on the original dataset without implementing any data augmentations and accompanied by the EfficientNet-B0 as a 2-classifier pipeline. Having trained for a considerably high number of iterations, the training time is comparatively greater than other architectures and will only increase with the increase in iterations over the dataset or epochs. After a certain threshold, overfitting might become a considerable problem that will have to be dealt with manually to preserve the accuracy offered by the model.

6 Conclusion

Deep learning based on neural networks offers a new way to process and analyze medical imaging results. Radiographs, as an important health tool for diagnosis, is an essential component of clinical assessment and treatment planning. On specific tasks, computer-aided systems based on deep learning have shown good success in particular activities, such as assisting clinicians in reducing diagnosis time, improving accurate diagnosis rate, and designing treatment plans. We have presented some initial results on detecting various chest abnormalities from chest radiographs using a custom deep-learning model. As stated earlier, due the high ambiguity of inter-applicable labels pertaining to chest abnormalities, the results still look quite promising despite the large size of the publicly available dataset. We plan to try out various other approaches to smooth out the errors in our current ways and to effectively increase the overall accuracy of our predictions. With the assistance of the continued advancement of deep learning algorithms, the growth of high performance parallel computing technologies, the growing sample set of medical image labels and the improvement of medical image quality, we believe that deep learning based medical image processing will be very promising in the future.

References

1. Al-Waisy, A., Mohammed, M.A., Al-Fahdawi, S., Maashi, M., Garcia-Zapirain, B., Abdulkareem, K.H., Mostafa, S., Kumar, N.M., Le, D.N.: Covid-deepnet: hybrid multimodal deep learning system for improving covid-19 pneumonia detection in chest x-ray images. Computers, Materials and Continua **67**(2) (2021)

2. Anis, S., Lai, K.W., Chuah, J.H., Ali, S.M., Mohafez, H., Hadizadeh, M., Yan, D., Ong, Z.C.: An overview of deep learning approaches in chest radiograph. IEEE Access **8**, 182347–182354 (2020)
3. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
4. C. Khoshgoftaar, T.: A survey on image data augmentation for deep learning. (2019). https://doi.org/https://doi.org/10.1186/s40537-019-0197-0
5. Freeman, I., Roese-Koerner, L., Kummert, A.: Effnet: An efficient structure for convolutional neural networks (2018)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
7. Hore, S., Chakraborty, S., Chatterjee, S., Dey, N., Ashour, A.S., Van Chung, L., Le, D.N.: An integrated interactive technique for image segmentation using stack based seeded region growing and thresholding. International Journal of Electrical & Computer Engineering (2088-8708) **6**(6) (2016)
8. Institute, V.B.D.: Vinbigdata chest x-ray abnormalities detection (2020), https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalitiesdetection
9. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr database. PhysioNet10 **13026**, C2JT1Q (2019)
10. Litjens, G., Ciompi, F., Wolterink, J.M., de Vos, B.D., Leiner, T., Teuwen, J., Išgum, I.: State-of-the-art deep learning in cardiovascular image analysis. JACC: Cardiovascular imaging **12**(8 Part 1), 1549–1565 (2019)
11. Mohammed, M.A., Abdulkareem, K.H., Garcia-Zapirain, B., Mostafa, S.A., Maashi, M.S., Al-Waisy, A.S., Subhi, M.A., Mutlag, A.A., Le, D.N.: A comprehensive investigation of machine learning feature extraction and classification methods for automated diagnosis of covid-19 based on x-ray images. Computers, Materials and Continua **66**(3) (2020)
12. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T.T., Dinh, D.H., Do, C.D., Doan, L.T., Nguyen, C.N., Nguyen, B.T., Nguyen, Q.V., Hoang, A.D., Phan, H.N., Nguyen, A.T., Ho, P.H., Ngo, D.T., Nguyen, N.T., Nguyen, N.T., Dao, M., Vu, V.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations (2021)
13. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning (2017)
14. Razzak, M.I., Naz, S., Zaib, A.: Deep Learning for Medical Image Processing: Overview, Challenges and the Future, pp. 323–350. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-65981-7_12, https://doi.org/10.1007/978-3-319-65981-7_12
15. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection (2016)
16. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.: Chestx-ray14: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases (09 2017)
17. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
18. xhlulu: Vinbigdata chest x-ray resized png (256x256) (2020), https://www.kaggle.com/xhlulu/vinbigdata-chest-xray-resized-png-256x256