ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

Empowering Radiologists: ConvNeXt-Large Model on CheXpert Using a Customized Webtool

[1] Zahra Khamesi, [2] Jean-Francois Samson, [3] Ching Y. Suen, [4] Mathieu Mailhot

[1] [3] Concordia University,
[2] [4] Centre Intégré Universitaire de Santé et de Services Sociaux (CIUSSS), du Centre-Sud-de-l'Ile-de-Montréal
Corresponding Author Email: [1] zahra.khamesi.ccsmtl@ssss.gouv.qc.ca, [2] jean-francois.samson.ccsmtl@ssss.gouv.qc.ca, [3] suen@cse.concordia.ca, [4] mathieu.mailhot.ccsmtl@ssss.gouv.qc.ca

***Abstract*— In collaboration with CIUSSS du Centre-Sud-de-l'Ile-de-Montréal, this paper presents RADIA, a project combining several techniques of deep learning algorithm developed for detecting 14 classes in the thoracic chest X-ray, using the combination of 112,120 public images. According to the ML group at Stanford, this concept was built based on previous studies on ChexNet. Our training phase has been extended by integrating Frontal and Lateral views and use of algorithms to improve the images. A model has been implemented based on ConvNeXt-Large , with improvements. We present the result obtain with the metrics AUC, F1, G-mean to better define the behavior with a imbalanced dataset. The challenge does not end with the realization of a software tool based on the deep learning technics but it will continue with the development of a decision helper combining inference model and webtool for the radiologists and other healthcare teams.**

***Index Terms*— Chest X-Ray, Healthcare, Pattern Recognition, Deep Learning.**

I. INTRODUCTION

Chest X-Ray is an image widely used as a non-invasive pre-diagnostic exam for checking the patient's chest condition. This technology allows radiologists to detect visual evidence for diseases more effectively and efficiently, which helps to save numerous lives.

The overlap of chest tissue, the lack of detail in grayscale X-ray images, and common symptoms of abnormalities are challenges to overcome. Deep learning algorithms for pattern recognition extract features from complex chest X-ray images [18], allowing the detection and classification of lung diseases to help overcome these challenges.

In this research project, our team aims to train the model on public chest X-ray images and demonstrate its result on a customized web tool for radiologists called RADIA.

II. RESEARCH BACKGROUND

As a result of the capability to extract useful features automatically, Deep Learning is the best choice for solving medical problems due to its high accuracy and data analysis [2]. With image processing, valuable information about an image is capture and highlight [16], increase the relative sharpness and making the image more computationally efficient. As a way of better understanding of lung disorders, our team attempted to learn experiences from the following studies, which helped us gain valuable insights into lung abnormalities classification with deep learning.

A. CheXNet

The CheXNet [15] detects pneumonia, exceeding the F1 rate of human radiologists. A 121-layer convolutional neural network is trained on the ChestX-ray14 dataset, a public chest X-ray repository containing over 100,000 frontal views with 14 chest diseases, Including Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural-thickening, Cardiomegaly, Nodule, Mass, Hernia, and No finding. Model performance results achieved state-of-the-art.

Medical AI Researcher and Radiologist Lauren Oakden-Rayner, in her blog (https://laurenoakdenrayner.com/2018/01/24/chexnet-an-in-depth-review/), highlights both strengths and limitations of using a single radiologist to label ground truth data. One limitation she notes is the potential for bias in the labeling process. Additionally, Oakden-Rayner points out a metric limitation in CheXNet that can impact the accuracy of the model.

Our team believes that improving the research project during training and for future development is crucial, and the accuracy of the labels is essential to the training process.

B. CheXpert

The CheXpert [7] is a project contest with a public dataset of over 224,000 chest radiographs and presents an approach for automatic detection of seven of the same classes as in ChexNet and seven different classes such as Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other and Support Devices. Labels are automatically associated with a patient file combining frontal and lateral images. Algorithms are validated on a set of 200 manually annotated chest radiographs. The main findings are that different uncertainty approaches are useful for anomaly detection. The best models are evaluated with a test set of 500 images. Models outperform at least two of

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

three radiologists in detecting five clinically relevant anomalies. The authors released the CheXpert dataset to the public as a benchmark for chest radiograph interpretation models.

C. CheXseen

A DenseNet121 model is used to detect diseases that were not labeled and not labeled in different stages by dividing CheXpert dataset into two categories: 'seen diseases' and 'unseen diseases.' Results showed that models might report 'no disease' rather than an 'unknown disease' and highlight unknown disease detection to prevent misclassification. The authors establish that Multi-label models provide useful baselines for detecting unseen diseases [17].

D. IBM Research

IBM Research [21] conducts a two years study to develop algorithms to detect chest anomalies. Several architectures were experimented with, and the conclusion came that training a single network was the best option. The classification results are inputs for automatically generating a preliminary textual report, helping the radiologist expedite the writing [12] [19].

III. RADIOLOGY WORKLOAD

Most hospital departments widely use chest X-ray images, especially in the emergency room. This modality does not provide a diagnostic, but more likely, a global view of the patient's state before defining a medical specialist decides on a diagnosis. The radiology service typically takes two images of a person, a frontal and a lateral view. Radiologists at their working stations usually view these two views. These images and the radiologist report are stored in an image repository Répertoire d'Imagerie Diagnostique (RID), which is a major participant of this project.

The workload for this type of modality is around 500 images per day for a main Quebec hospital; each radiologist can watch 100 to 200 images per day [5].Because of the significant use of these images, there is a need to develop a software tool to assist the radiologist in making decisions and reducing their workloads.

IV. DATASETS

Data from CheXpert datasets are anonymous and publicly available. This dataset contains various abnormality classes such as Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion, No Finding, Pneumonia, Pneumothorax, Lung Opacity, and Emphysema.

The team used the CheXpert train and validation datasets for train and validate. Tests will run on a private dataset from CIUSSS that radiologists will label in the future to provide ground truth.

One of the main challenges is imbalanced datasets [8] during training and validation Figure 1 and Figure 2. It significantly affects deep learning models' performance, and leading to poor results for minority classes, especially for validation datasets for which some of the classes are not present, like fractures Figure 2. From another hand, the model has bias through majority classes like "Lung Opacity." We address this issue through techniques such as class-weighted calculation [6], means classes is weighted contrary to their amount during training. Consequently, the smaller class is given a higher weight and the larger class a lower weight and label smoothing technique [14]. Addressing imbalanced datasets is crucial in order to have a more accurate result.

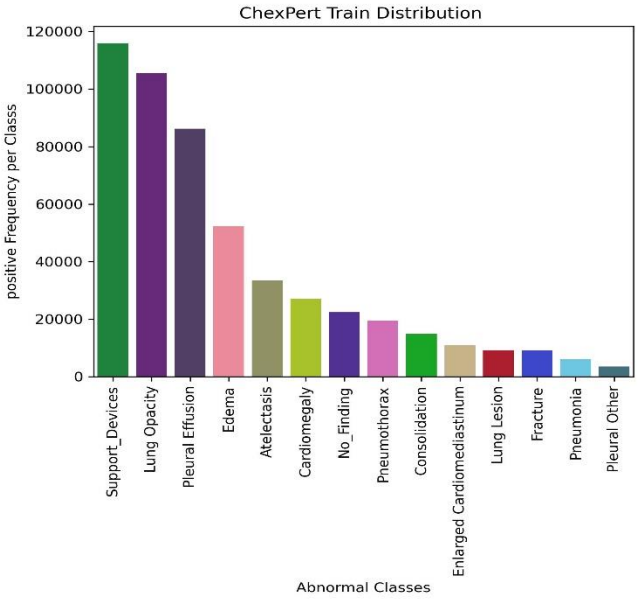

Figure 1: ChexPert Train Positive Frequency per Class

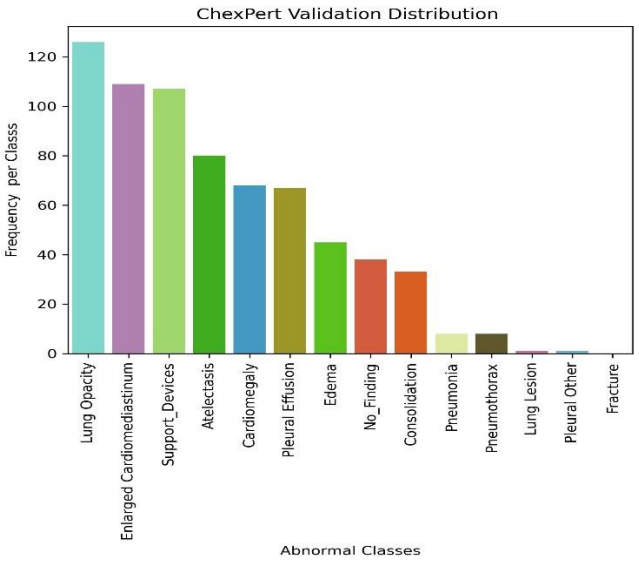

Figure 2: ChexPert Validation Frequency per Class

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

V. CATEGORIES & CLASSES

The anomaly classes are categorized to assist users in localizing anomalies. The results are formatted according to the local medical team's requirements Figure 3, allowing them to group abnormalities based on their inherent characteristics, such as Air, Liquid, etc. This preformatted result will facilitate the preparation of future reports.

The definition of the categories prepares the integration of algorithms of hierarchical type [13] for applying conditional probability rules.

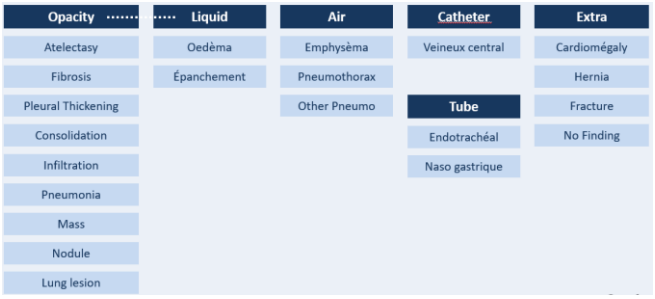

Figure 3: Classes and Anomaly Categories

VI. IMAGE ENHANCEMENT

The thoracic images are mainly gray images, even if they are stored in a DICOM format for supporting colors. DICOM files support supplementary data to identify the image format.

The X-Ray images have a majority of grey levels rather than colors level. We decide to work with one channel of grey image and complete the two other channels by enhancing the contrast with a CLAHE algorithm [11] and applying a log or a Gamma pixel wise correction. The effect foreseen is to highlight the details related to a specific anomaly and even more detach some particular details from the image background Figure 4, in this way, to help model for feature extraction. The training image input is a six channels tensor to store lateral and frontal chest images encoded in grey levels on one channel and the corresponding enhanced images with the CLAHE algorithm.

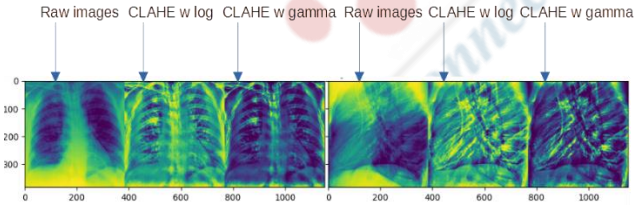

Figure 4: Frontal and Lateral images encoded on 6 channels

VII. IMAGE AUGMENTATION

The classic image augmentation [3] is applied with a probability rate defined at the training start-up: Rotation between 15 to 15 degrees, elastic image conversion with the nearest pixels interpolation, magnifying the image, and affine transformation like translation and grid distortion.

The image augmentation increases the variety of images in the training dataset and is not applied to the validation datasets. Each augmentation is applied with a probability level between 0 and 1 Figure 5 to increase the small variation around the normal image.

VIII. ARCHITECTURE

Figure 6 provides an overview of the various stages for training the model.

The model architecture is built on the top of the public CNN, ConvNeXt Figure 7, combining concepts of Convolutional Neural Networks (CNNs) and Transformers to improve image classification performance. It outperformed the powerful Swin-Transformer model by using various techniques such as adjusting training parameters, data augmentation, and regularization on the ImageNet-1K dataset. ConvNeXt also implemented layer stacking strategies and modified the model's interior, like changing the activation function RELU with GELU and the normalization method to improve performance [10]. It allows choosing a model among Small, Base, Large, and Extra-Large for improving the Deep learning operations.

The PyTorch framework is used for building the classification model; the main parameter values Figure 5 in the model are Mini-batch size 32, Learning rate of 0.0001, Weight decay of 0.05, the Dropout rate of 0.5, Epoch size of 120000 images, optimizer is AdamW, which is effective for deep learning models. During the training, the OneCycle scheduler adjusts the learning rate. Training stops early when no validation loss improvement is measured. Label smoothing is used with a very small value (0.05), and we clip the gradient norm to 1. We use a random sampler to sample data from the training dataset with initial weight from ImageNet-1K. During training, class weights are used to balance the class distribution. The training loss is processed with the BCEWithLogitLoss, and the validation loss using CrossEntropyLoss with a softmax.

Configuration Values	
Mini-batch	32
Learning rate	0.0001
Weight decay	0,05
Dropout rate	0,5
Epoch size	120000
Optimizer	AdamW
Label smoothing	0.05
Clip Gradient norm	1
Training loss	BCEWithLogitLoss
Validation loss	CrossEntropyLoss
Activation function	Softmax
Threshold	Dynamic
Augmentations prob	[0.5,0.0,0.5,0.5,0.5]

Figure 5: Main parameter values in the model

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

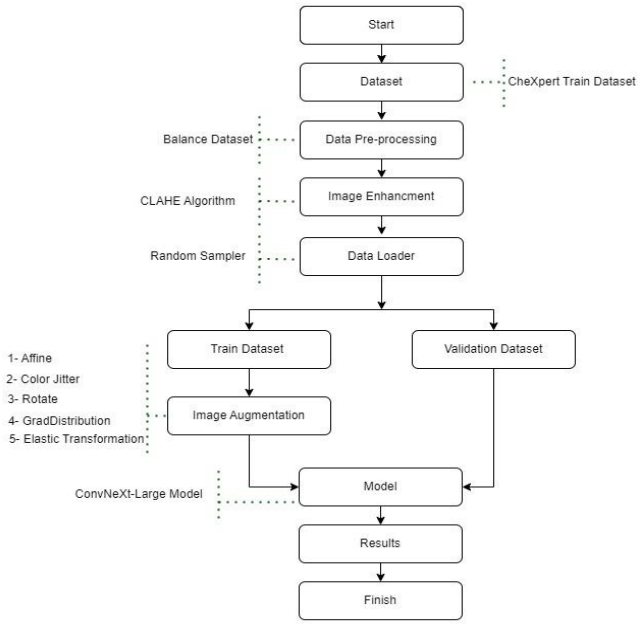

Figure 6: Process Flowchart

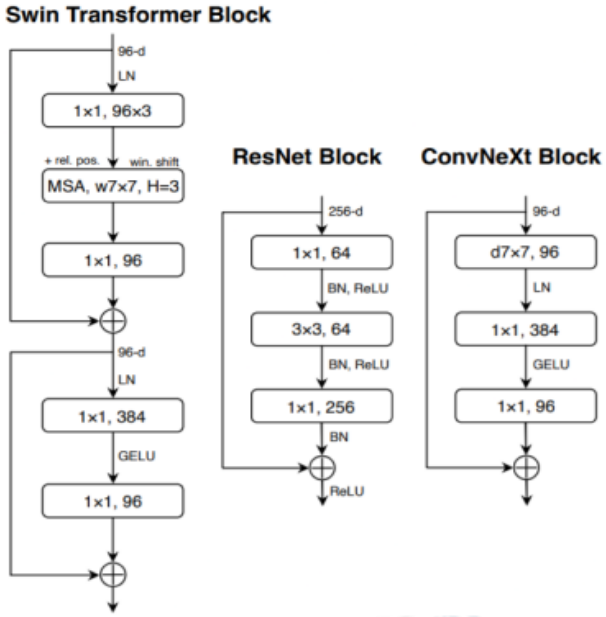

Figure 7: Block designs for a ResNet, a Swin Transformer, and a ConvNeXt [10]

IX. METRICS & THRESOLDS

The metrics processes during the training and validation phases are F1, G-mean [4], AUC (Area Under the ROC Curve), and precision/recall. The most appropriate metrics for imbalanced datasets are G-mean, F1 score, and the AUC because it considers both measurement sensitivity and specificity [1]. It gives equal weight to both and can better evaluate the model performance that needs to classify minority and majority classes. The thresholds are dynamically processed per class to maximize the F1.

X. RESULT

Our team monitoring process during training the model uses an online web server called Weight& Biases (https://wandb.ai/).

We present the AUC results, which is the most significant metric for imbalanced datasets. Our results show that ConvNeXt-Large requires more training data. It is our intention to integrate features such as frontal and lateral image training, as well as more anomalies such as rib fractures, into our product.

Figure 8 provides details on the different results for two experiments to train a classification model based on the CheXpert dataset. Both applied the ConvNeXt-Large backbone, AdamW optimizer, OneCycle learning rate, Class Weighted, and loss function the same. However, we used different values for certain parameters like data augmentation probabilities and the number of samples per epoch. We also used different input values for the CLAHE algorithm and dynamic threshold instead of static to maximize F1 for 'smart-cherry-1132', which details are described in the architecture section. Figure 8 shows improvement after applying the above changes. Furthermore, for classes with insufficient or no data in validation datasets, we have result 0, and based on local radiologist requirements, 'Support Devices' are dropped during training.

In Figure 9, we also provide results that showed significant improvements in F1 scores. We observed much greater improvement for the second model, which showed higher F1 scores for all classes.

Figure 10 contains results for G-Mean applied to the second model, "smart-cherry-1132", and compares the AUC and F1 metrics results demonstrating an average level of performance in all classes.

AUC		
Classes / Weight&Biases Name	passionate-chocolate-1021	smart-cherry-1132
Atelectasis	0,715	**0,736**
Cardiomegaly	0,679	**0,750**
Consolidation	**0,732**	0,672
Edema	0,777	**0,837**
Enlarged Cardiomediastinum	0,545	**0,731**
Lung Opacity	**0,805**	0,768
No Finding	0,699	**0,929**
Pleural Effusion	0,831	**0,850**
Pneumonia	0,705	**0,778**
Pneumothorax	0,697	**0,842**
Support Devices	0	0
Fracture	0	0
Lung Lesion	0	0
Pleural Other	0	0

Figure 8: AUC Results

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

F1		
Classes / Weight&Biases Name	passionate-chocolate-1021	smart-cherry-1132
Atelectasis	0,193	**0,651**
Cardiomegaly	0,189	**0,511**
Edema	0,301	**0,582**
Lung Opacity	0,517	**0,811**

Figure 9: F1 Results

G-Mean	
Classes / Weight&Biases Name	smart-cherry-1132
Atelectasis	0,640
Cardiomegaly	0,623
Edema	0,706
Lung Opacity	0,621
No Finding	0,515
Pleural Effusion	0,763

Figure 10: G-mean Results

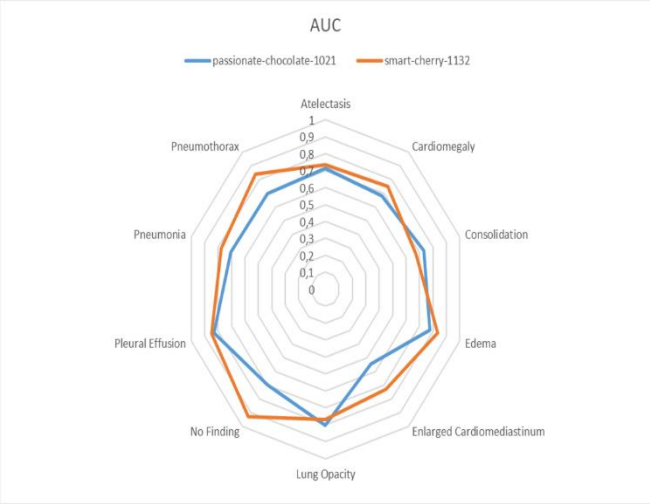

Figure 11: AUC Radar Chart

XI. WEBTOOL

Combining the ConvNeXt-Large model on CheXpert with an adaptable web tool to transform the way radiologists work. The web tool provides an intuitive and user-friendly interface that allows radiologists to receive automated analysis results in real-time, which can improve the efficiency and accuracy of the analysis process. Additionally, the web tool is customizable to suit the specific needs and workflows of different healthcare institutions, including integrating existing EMR systems and incorporating new deep-learning models or algorithms. In order to help medical teams in their daily decisions, image classifications can provide a solution. In this project, we aim to design a tool that functions as a safety net to prevent humans from neglecting the potential anomalies during the expertise of chest X-Ray images.

To present the results from the model, we develop a Web User Interface using the Blazor framework from Microsoft Windows to implement RADIA access on smartphones. The goal is not to replace humans with machines but to provide another perspective on chest X-rays for healthcare teams. The tool purpose is to be used as a decision support during times of high workload. Additionally, radiologists can customize the different options based on their needs during their detection to ensure patients receive faster and better care, especially in emergency rooms.

RADIA aims to help radiologists make the right decision by improving the efficiency and accuracy of the analysis process. However, it is not intended to reduce detecting time by over-reliance on it.

RADIA web tool is divided into the image viewer and the confidence probability panel Figure 12. In the confidence probability panel, our team tries to present the possibility of disease based on the local medical team's requirements, which display according to the main categories and subsets of anomalies. In front of each disease name, a colorful bar indicates whether anomaly is present or absent based on a constant threshold.

The implementation of an inference model shows that the processing of the model is very fast, less than 500 milliseconds for an image resized to 384 x 384.

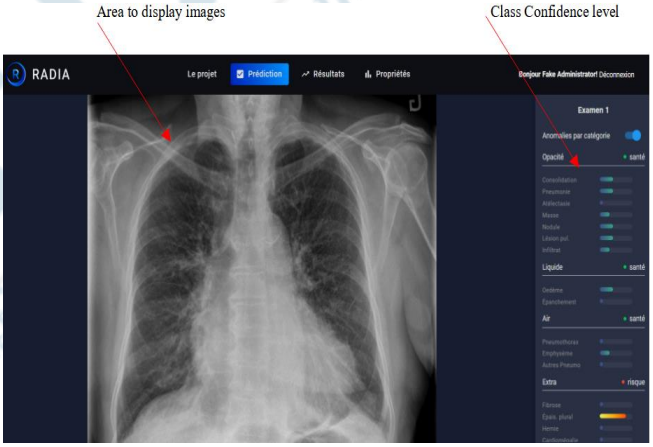

Figure 12: RADIA WebTool with the probability values per class

XII. CONCLUSION

Healthcare software projects can be difficult to manage because of several reasons. Firstly, the projects involve handling confidential patient information, which presents challenges in terms of privacy and security. Secondly, healthcare organizations that do not utilize software tools to improve their operational efficiency may struggle to adapt to new technologies. In addition, while data storage is commonly seen as the final destination for medical examinations, it is actually a crucial component in training artificial intelligence projects.

Furthermore, it is necessary to have a more balanced dataset to improve training. Adding more public datasets and

taking advantage of private datasets from CIUSSS will help to achieve this. Additionally, accuracy can be improved by combining human expertise and models, as shown in the ChexZero [20] paper, and implementing more effective parameters and hyperparameters like Focal Loss [9] to optimize algorithms. For web tools, dynamic thresholds and feedback analysis will be investigated.

Also, our team's efforts will focus on implementing various localization features such as class activation mapping (heatmap), bounding boxes, and instance segmentation. By highlighting specific areas, these features draw the viewer's attention, and the NLP algorithm can provide automatic labeling and reporting generators that radiologists may also benefit from for daily work.

REFERENCES

[1] Akosa, Josephine. "Predictive accuracy: A misleading performance measure for highly imbalanced data." Proceedings of the SAS global forum. Vol. 12, April 2017.

[2] Allaouzi, I. and Ahmed, M.B. A novel approach for multi-label chest X-ray classification of common thorax diseases. IEEE Access, 7, pp.64279-64288, May 2019.

[3] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.

[4] Espíndola, Rogério P., and Nelson FF Ebecken. "On extending f-measure and g-mean metrics to multi-class problems." WIT Transactions on Information and Communication Technologies 35, pp. 25-34, May 2005.

[5] Forsberg, D., Rosipko, B. and Sunshine, J.L., 2017. Radiologists' variation of time to read across different procedure types. Journal of digital imaging, 30(1), pp.86-94, February 2017.

[6] Glossary of Common Terms and API Elements, scikit-learn 0.23.2,Retrieved from https://scikitlearn.org/0.23/glossary.html

[7] Irvin, Jeremy, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." In Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, pp. 590-597, 2019.

[8] Johnson, J.M. and Khoshgoftaar, T.M. Survey on deep learning with class imbalance. Journal of Big Data, 6(1), pp.1-54, December 2019.

[9] Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal loss for dense object detection." In Proceedings of the IEEE international conference on computer vision, pp. 2980-2988, 2017.

[10] Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A convnet for the 2020s." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976-11986,2022.

[11] Omarova, G. S., and V. V. Starovoitov. "Application of the Clahe Method Contrast Enhancement of X-Ray Images," 2022.

[12] P. Kisilev; E. Walach; E. Barkan; B. Ophir; S. Alpert; S. Y. Hashoul, "From Medical Image to Automatic Medical Report Generation", IBM Journal of Research and Development (Volume: 59, Issue: 2/3, pp. 2-1 to 2-7), March-May 2015.

[13] Pham, H.H., Le, T.T., Tran, D.Q., Ngo, D.T. and Nguyen, H.Q., 2021. Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing, 437, pp.186-194, May 2021.

[14] PyTorch V2.0, CROSSENTROPYLOSS. Retrieved from https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html#crossentropyloss

[15] Rajpurkar, Pranav, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." arXiv preprint arXiv:1711.05225, November 2017.

[16] Saxena, S., Sharma, S. and Sharma, N. Research Article Parallel Image Processing Techniques, Benefits and Limitations. Research Journal of Applied Sciences, Engineering and Technology, 12(2), pp.223-238, 2016.

[17] Shi S, Malhi I, Tran K, Ng AY, Rajpurkar P. CheXseen: Unseen disease detection for deep learning interpretation of chest X-rays. arXiv preprint arXiv:2103.04590. March 2021.

[18] Suganyadevi, S., Seethalakshmi, V. and Balasamy, K. A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval, 11(1), pp.19-38, March 2022.

[19] Syeda-Mahmood, Tanveer, et al. "Chest x-ray report generation through fine-grained label learning." Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23. Springer International Publishing, 2020.

[20] Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y. and Rajpurkar, P.Expert-level detection of anomalies from unannotated chest X-ray images via self-supervised learning. Nature Biomedical Engineering, pp.1-8, September 2022.

[21] Wu, Joy T., et al. "Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents." *JAMA network open* 3.10: e2022779-e2022779, October 2020.