Explainable Pneumonia Learning: A Comprehensive Study

HIZIR CAN BAYRAM[1], EMRE CETIN[2]

[1]Istanbul Technical University, Sariyer, Istanbul 34000 TR (e-mail: hizircanbayram@gmail.com)
[2]Istanbul Technical University, Sariyer, Istanbul 34000 TR (e-mail: emrec.tin@gmail.com)

This work was carried out under Trustable AI course, given by Zehra Cataltepe at Istanbul Technical University, 2021-2022 Fall Semester.

ABSTRACT Pneumonia causes certain deaths every year all over the world. Even though early diagnosis might be likely by various imaging techniques such as CT, CXR, it is a vast amount of time for radiologists to analyze medical images by conventional techniques since it requires certain amount of experts and deep knowledge. However, number of experts in this area makes early diagnosis inaccessible for everyone. To eliminate this, CAD systems can help average radiologists to catch expert-level performance in terms of having more meaningful and interpretable results from these imaging techniques. In this work, we explored pneumonia detection in chest X-ray images and how CAD systems can help us to have an interpretable outcome. So, our work's main focus is not confined with diagnosis. We rather explored if we could interpret X-ray images in a meaningful manner, how the interpretability of these models can be boosted up, when considering robustness of the models especially. To do so, we performed various experiments such as training different deep net architectures, measuring how they differ from unlearnt deep net architectures, how lung segmentation affect interpretability of these models etc. We showed that creating interpretable models is possible and we can outperform these models' performances using different techniques. The app, code snippets, and notebooks are available at https://github.com/hizircanbayram/TrustableAIProject.

INDEX TERMS explainable machine learning, grad cam, layer cam, pneumonia detection, rsna

I. INTRODUCTION

CHEST X-ray or CXR is one of the most widely used diagnosing tools for lung and heart related diseases due to its easy access, low risk of side effects and inexpensiveness [1]. According to [2], pneumonia diseases cause almost 50.000 people in the United States to die each year, a major evidence to the importance of diagnosis. Although early diagnosis is possible by doctors using CXR, it is a great burden for them to evaluate these CXRs from scratch because it is time consuming and labor intensive.

In recent years, the development of computer-aided diagnosis (CAD) in medicine [4] has gained special attention in order to ease the burden of doctors. Availability of open-source databases such as [5], [6] and a wide variety of deep learning architectures such as [7]–[9] were crucial to develop such CAD systems and diagnosis algorithms in particular using CXR. Erdi et al. [10] showed a studiously categorized list of algorithms from medical image segmentation to diagnosis and from domain adaptation to medical image retrieval using CXR images.

A serious number of patients from a wide variety of ages lose their lives to pneumonia, a mortality disease caused by either bacterial or viral infection, each year [11], [12]. Early diagnosis of pneumonia can significantly reduce the number of deaths. However, availability of expert radiologists, especially in low-income countries, makes early diagnosis inaccessible. In order to avoid this, CAD systems can help naive doctors to navigate and have some insights from an X-ray image.

Although various pneumonia detection systems are offered in the literature such as [13], [14], a lot of systems in the literature are not suitable for clinical setting. Roberts et al. [15] point this out by examining more than 2000 papers for detection of COVID-19 and they explain various criterias to consider a paper clinically-suitable. When we did an investigation similar to the work by them, we found out that this problem also occurs in the pneumonia literature. Having an interpretable artificial intelligence system that automatically detects pathological conditions in X-ray images can assist in quantifying the severity of the condition and necessary clinical-support.

Interpretability of machine learning and deep learning

models' output is a newly yet major field [16]. In order to evaluate a model's clinical suitability, one needs to consider how and where the model predicts such an output. This necessity leads to a new literature for interpretable machine learning for medical image analysis. Van der Velden Bas HM et al. [17] give a broad taxonomy for XAI in medical image analysis and list state of the art algorithms, their advantages and disadvantages.

In this work, we'll investigate pneumonia detection in chest X-ray images. Our method will not be limited to diagnosis. We'll rather investigate how we can interpret the model outcome, how the interpretability can be leveraged in terms of robustness and enriched variety of insights. To do so, we'll use different datasets [18]–[20] that have extra information about the pathological regions.

II. RELATED WORK

This section discusses some influential papers in the literature related to one of the following topics: model inspection and explainability in lung segmentation, pneumonia or COVID-19 identification in CXR/CT images. Moreover, we also discuss potential limitations, biases, and problems of COVID-19 or pneumonia identification given the current state of available databases. It is important to observe that as the identification of COVID-19 in CXR/CT images is a hot topic nowadays due to the growing pandemic, it is unfeasible to represent the actual state-of-the-art for this task since new works are emerging every day. Nevertheless, we may observe that most of those works aim to investigate configurations for Deep Neural Networks, which is already different from our proposal.

In order to show how fast is growing the research content around the topic of Machine Learning applications on COVID-19 and Pneumonia Diseases , we can briefly present some surveys and reviews published in the literature.

A. DEEP LEARNING METHODS FOR PNEUMONIA DIAGNOSIS

Most of the previous works in this field use a pediatric dataset published by Kermany et al. [28], consisting of a total of 5856 radiographs divided in three classes (2780 bacterial, 1493 viral and 1583 normal) of patients between 1 and 5 years of age. Other works in the field [29] use other datasets and a classification hierarchy. First, radiographs are classified as "pneumonia" or "normal". Then, radiographs labeled as pneumonia are classified as "viral" or "bacterial". Kermany et al. obtained an AUC of 0.85 for the first classification (pneumonia versus normal) and 0.81 for the second (viral versus bacterial). We can also find the same classification system in [28], which will be explained later.

B. EXPLAINABLE AI METHODS IN MEDICINE

As was previously mentioned, CNNs are considered black-box algorithms, which increases the difficulty of applying them in areas such as Medicine. The reason is that even these algorithms can be very precise classifying radiographs or making predictions, the end-users (medical staff in our case) will need to interpret and understand how and why the algorithm has reached the conclusion (the outcome provided to the end-user). For this reason, it is important to develop eXplainable AI (XAI) systems that allow end-users to understand how the system works (i.e. how and why classifies the radiography in our case). We can distinguish between two main techniques, the first being object detection systems. These systems are based on CNN models that locate different objects in images. An example of this type of model is CoupleNet, which classifies and locates signs of pneumonia on a chest radiography and produces a visualization of the original image with bounding boxes in lung areas that show signs of disease [30]. These algorithms are used in different works to detect and locate pneumonia signs [31] combining two CNNs for the detection and location of pneumonia signs on lungs. The second kind of methods uses additional techniques to visualize how the model classifies, such as heatmap generation. This method is used in a wide variety of problems including pneumonia diagnosis. For example in [32], where different lung areas can be seen with different color intensities according their relevance to the prediction made by the model. Other similar work is presented in Zech et al. [33]. A substantial difference between the two previous methods is that in the first one it is necessary to build a training dataset where the areas with signs of the disease have been marked by the experts. Whereas, in the second type of methods, it is only necessary to label each radiography of the training set with the classification given by the experts (consolidation/non consolidation in our case). After training, medical staff need to review the heatmaps generated by the model to validate the clinical sense.

III. MATERIALS AND METHOD

In this work, we aimed to investigate how different training setups affect the forming of saliency maps. To do so various experiments were conducted. Firstly, we chose two fundamental deep neural network architectures of medical domain: Dense-Net and Res-Net, namely. We approached to the problem as a binary classification problem and trained these neural nets with RSNA Pneumonia dataset [18] accordingly. Then saliency maps acquired from Grad-CAM and Layer-Cam techniques were generated. Next, in order to determine if these saliency maps were learnt based on our binary problem or not, we designed an experiment that differentiates learnt neural nets from randomly trained ones. Our hypothesis was that the more the difference between them was high, the likely the saliency maps were generated according to our problem. Finally, so as to measure how non-lung areas of chest X-ray images change our saliency maps, we trained a deep lung segmentation model and passed X-ray images to it. After the lung mask was detected, we colored non-lung areas black so that our deep models could not learn anything from those regions. Our main metric for evaluating how good the generated saliency maps were was the Dice coefficient between saliency maps and bounding box of lung

opacity disease in X-ray images. To compare these two, we needed to turn our saliency maps into binary masks. One critical questions arised here: how to select a tresholding value for it when optimizing time to choose it and accuracy of saliency maps? An appropriate strategy was designed for it. Below, we gave further details about the experiments and designs mentioned above.

A. RSNA PNEUMONIA DATASET

This dataset is formed by Radiological Society of North America(RSNA) by benefiting ChestXray-14 dataset [5] in order to diagnose thoracic diseases. Where this dataset differs from ChestXray-14 is that the images are shared in DICOM format and the dataset is arranged so that it can be used to diagnose pneumonia in particular. DICOM format consists of pixel values of X-ray images as well as the metadata that belongs to the images. The dataset contains 26.684 images where each one is labeled either "Not Normal/No Lung Opacity", "Normal", "Lung Opacity". "Normal" label represents neither pneumonia nor any other conditions detected in the image. "Not Normal/No Lung Opacity" represents opacity in the X-ray image is caused by a condition other than pneumonia, hence it is related to any other conditions. Finally "Lung Opacity" represents opacity in the image caused by pneumonia. Opaque areas are segmented by radiologists using bounding box technique. More than one bounding box can be captured in the images. Images labeled as "Lung Opacity" are assumed as pneumonia. In other words, the dataset consists of 9555 pneumonia positive images as well as 20672 pneumonia negative images. Distribution of the labels are shown in the table below.

Label	Number of Images
Normal	8851
Lung Opacity	9555
No Lung Opacity / Not Normal	11821

TABLE 1. RSNA Pneumonia Dataset distribution

B. TRAINING DEEP NEURAL NETS

1) Data Preprocessing

Because the dataset is imbalance, we randomly sampled 8800 images from each class and trained our models. We read images as dicom format and then performed preprocessing steps; resizing to (224,224), cropping its center, flipping horizontally and rotating it up to 10 degrees randomly, respectively. We finally normalized images according to the ImageNet mean and std. This normalization technique is common in the literature [22], [23]. Summary of our process can be seen below.

2) Training Phase

We followed the best practises of medical image analysis and chose DenseNet [7] and Resnet [8] architectures for our experiments. In order to see the impact of saliency maps, we simplified the problem as we chose only two classes: normal and lung opacity. Binary cross-entropy loss function is used as a loss function. Networks take 3-dimensional 224x224 chest X-ray images and outputs a probability of how likely the image is lung opacity. Our optimization algorithm is Adadelta, with default hyperparameters of Pytorch framework. We trained models for 50 epochs using batch size of 16.

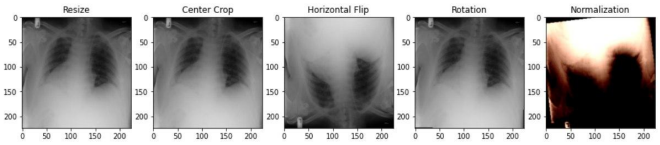

FIGURE 1. Data preprocessing pipeline

C. TRAINING RANDOMLY ASSIGNED DEEP NEURAL NETS

We designed an experiment to differentiate properly learnt deep models and the others. To do so, we trained another DenseNet and ResNet architectures with random labels. Instead of feeding correct labels to models, we generated random values of 0 or 1 and fed these values to the networks. Our aim here is generating saliency maps from non-learnt weights of the networks. Comparing saliency maps of learnt and unlearnt models would prove that the saliency maps were more likely interpretable or not.

D. DESIGNING LUNG SEGMENTATION BASED EXPERIMENT

1) Lung Segmentation Model

In order to segment lung areas of chest X-ray images, an updated version of Rula Amer et al.'s work derived from Frid Adar et al.'s work was preferred. This work uses U-Net [9] as a baseline with minor difference of its encoder part. U-Net network consists of encoder and decoder subnetworks and it is a fully convolutional network. Frid-Akar et al. initialized U-Net's encoder part by switching it with the feature extractor of VGG-16 architecture.

During training, dice-coefficient was used a loss function. The model takes an image of size 448x448x3 and outputs an image of size 448x448. The reason why the input image is 448x448x3 is to have an Image-Net pre-trained encoder for U-Net. The model was optimized using Adam's algorithm with default parameters. The batch size, epoch number and initial learning rate were 16, 30 and $10^{-3}, respectively.$

2) Lung Segmentation Model Dataset

Various datasets are combined for lung segmentation task. The reason behind it is that the number of lung-segmented images is not that common. A detail explanation is given below.

About Lung Segmentation Dataset

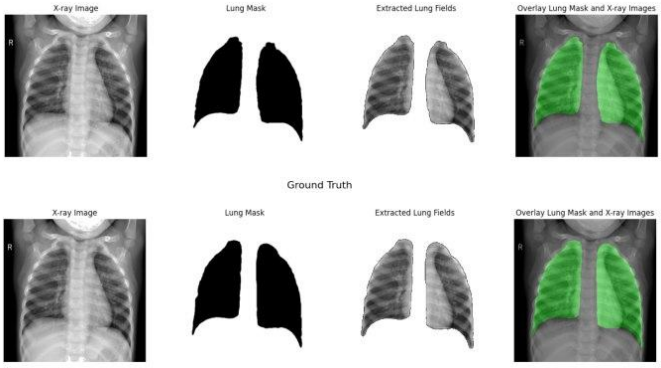

FIGURE 2. Lung segmentation output

1) JSRT [36] and SCR [37] Databases: JSRT database consists of 247 posteroanterior X-ray images coming from 13 Japan and 1 USA medical centers. Images are of size 2048x2048 and 12-bits grayscale with big-endian.
2) Montgomery County X-ray and Shenzhen Hospital X-ray Datasets [38]: Montgomery database is collected by a colloboration with Department of Health and Human Services of Maryland. The database consists of 138 frontal chest X-ray images taken through a tuberculosis program of Montgomery region.
3) V7Labs COVID-19 X-ray Dataset [39]: This database is formed by V7Labs of DarwinAI. The database consists of 6500 AP/PA chest X-ray images and pixel-wise polygonal lung segmentation masks.
4) XLSor Lung Segmentation Dataset [40]: XLSor database is generated using NIH, JSRT and Montgomery databases. The database consists of 2500 chest X-ray images and their corresponding masks.

Dataset	Number of Images
JSRT and SCR Databases	247
Montgomery County X-ray and Shenzhen Hospital X-ray Datasets	704
V7Labs COVID-19 X-ray Dataset	6266
XLSor Lung Segmentation Dataset	2500
Total	9717

TABLE 2. Lung Segmentation Dataset Distribution

3) Training Phase

We used lung segmentation model to have a lung mask of the feed-forwarded chest X-ray images. Then we combined this mask with the corresponding X-ray image. Thus, non-lung areas are colored black. Then, we followed the same procedure as in 3.2.1. We trained the model with the same procedure as in 3.2.2.

IV. RESULTS

This section presents an overview of our experimental findings and a preliminary analysis of each contribution individually.

Classification was performed using different models and different labeling techniques as stated in the Materials and Method section. We benchmarked these models for comparison. Lung Segmented model was trained with our dataset that consists of various pneumonia and COVID-19 datasets. Interpretability of our models was measured using Dice coefficient and we benchmarked different saliency map, deep architecture and lung segmentation supported techniques in order to see the big picture.

Table 3 presents results for our classification scenario for different networks, different labeling techniques and lung segmented support. Models using segmented CXR images presented better results than the models that used segmented images. There was no superior relation between models with or without lung segmented support. Both settings were on par in the normal class. In all cases, the models using segmented images performed similar, considering the selected metric. That result alone might discourage the usage of segmentation in practice in terms of classification.

Models	Sensitivity	Specificity	Accuracy
Random DenseNet	0	1	0,507
Random ResNet	0	1	0,507
Normal DenseNet	0,903	0,939	0,921
Normal ResNet	0,895	0,954	0,924
Lung Segmented DenseNet	0,9	0,94	0,916
Lung Segmented ResNet	0,91	0,93	0,92

TABLE 3. Models test results

Table 4 presents results for lung segmentation model. Results are promising in terms of segmenting lung areas in X-ray images, even for various datasets.

Metric Type	Score
Jaccard Index	0.91
Pixel Accuracy	0.97
Validation Loss	0.1177

TABLE 4. Lung segmentation test results

First images from left side in Figure 3 normal chest X-ray, second images Lime method using 10 future, third images Lime method using 15 future.

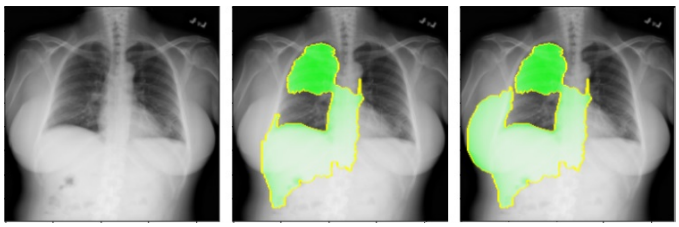

FIGURE 3. LIME Example.

When using the GradCam and LayerCam methods, the pixel values are set to 0 or 1 according to the threshold value. Images were turned gray without setting this threshold

value. The threshold value have a critical role in measuring the Gcam and Layercam methods. The success rate varies according to the threshold value determined. Performing performance analysis using equal and frequent intervals is very costly. Testing a threshold value with each model takes approximately 2 hours. Since there are 6 models, this period takes 13-14 hours. In order to minimize this cost, 5000 images were analyzed. We aimed to determine the threshold value ranges, start and end points of the images according to the pixel distributions. We observed that there is a distribution between 100-200 pixel values according to the histogram distribution. The rate of change was determined as 25 pixels. Therefore, analysis was performed using values of 100, 125, 150, 175 and 200. The histogram distribution is shown on the 4. Since the background color is black in lung images, the number of pixels in the 0-50 range is quite high. Because of this, the graph was corrupted. To prevent this, the 0-50 range has been removed from the chart.

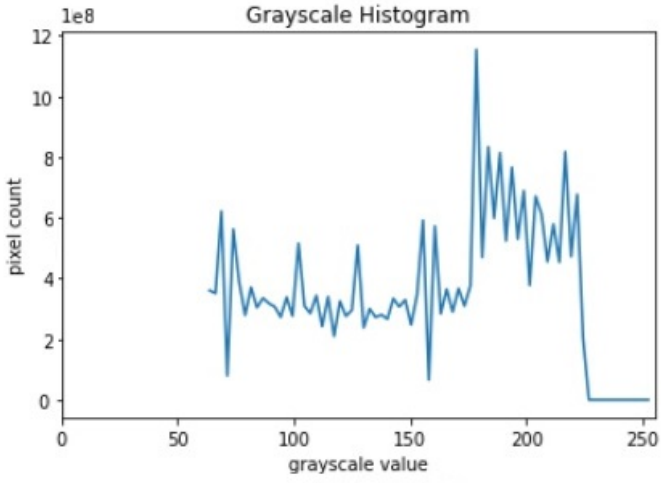

FIGURE 4. Histogram distribution for gray image

After the threshold value is determined, the area below the threshold value is marked as 0. The area above the threshold value is marked as 1. In the picture, 0s are represented by purple pixels. 1 is represented by yellow pixels. After the threshold value is applied, the middle color values disappear. It is clearly visible in the Figure 6. The degree of importance in the color distribution is shown in the Figure 5. The most important regions are shown in red, the least important regions in blue.

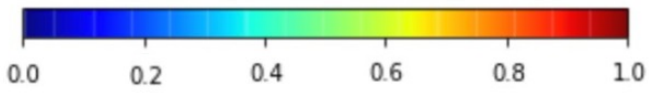

FIGURE 5. Color importance level

Figures 6, 7 and 8 present the LIME and Grad-CAM heatmaps for our classification s cenario. F igure 6 shows the difference between the randomly labeled model and the correctly labeled model. Since learning did not occur from the randomly labeled model,random models the class heat maps almost did not focus on the lung area at all.

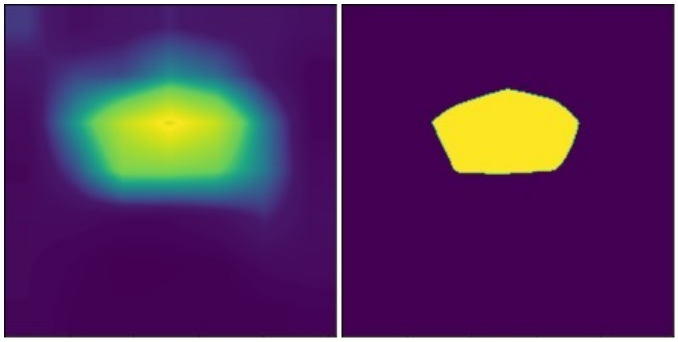

FIGURE 6. Applying a threshold value to the gray image

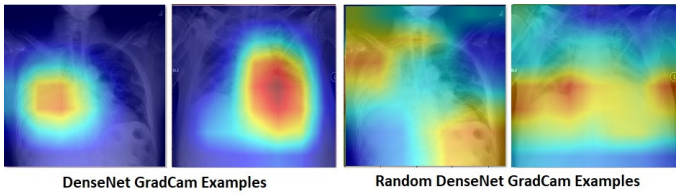

FIGURE 7. Gcam results for the trained model and Randomly labeled model from correctly labeled images.

We can notice that the models created CXR images focused primarily in the lung area in figure 7 and 8. It is seen that Gradcam focuses on a narrower area than Layercam.

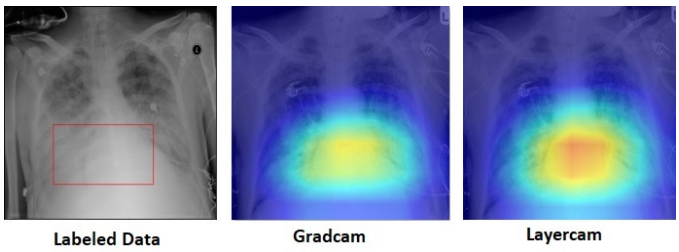

FIGURE 8. The labeled region, Gradcam and layer cam results are shown.

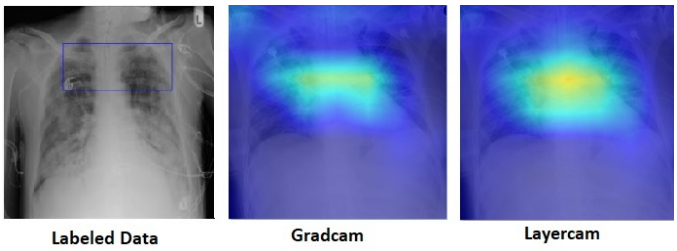

FIGURE 9. The labeled region, Gradcam and layer cam results are shown.

We calculated the Dice score for each model by using the threshold values of 100, 125, 150, 175, 200, which I determined as we mentioned in Figure 4. Dice score was obtained by using separate threshold values for each model. 5000

images were used for analysis. The analysis took approximately 18 hours. Analyzes were performed on a Windows10 operating system device with an Intel Core i7 processor and 16GB ram. The results of the LayerCam method are shown in TABLE 5 and the results of the Gcam method are shown in TABLE 6.

Models / Threshold	100	125	150	175	200
Random DenseNet	0.131	0.125	0.114	0.102	0.065
Random ResNet	0.126	0.122	0.114	0.106	0.081
Normal DenseNet	0.304	0.305	0.303	0.297	0.277
Normal ResNet	0.330	0.327	0.318	0.305	0.260
LungSegm DenseNet	**0.343**	**0.344**	**0.342**	**0.334**	**0.303**
LungSegm ResNet	0.339	0.337	0.330	0.318	0.262

TABLE 5. LayerCam Dice score for different threshold. LungSegm: Lung Segmented Model

Models / Threshold	100	125	150	175	200
Random DenseNet	0.	0.	0.	0.	0.
Random ResNet	0.	0.	0.	0.	0.
Normal DenseNet	0.292	0.293	0.292	0.286	0.268
Normal ResNet	0.327	0.324	0.315	0.302	0.258
LungSegm DenseNet	**0.340**	**0.339**	**0.340**	**0.331**	**0.301**
LungSegm ResNet	0.336	0.330	0.328	0.315	0.260

TABLE 6. GradCam Dice score for different threshold. LungSegm: Lung Segmented Model

V. DISCUSSION

We didn't see a major difference between using DenseNet and ResNet for a wide variety of metrics such as sensitivity, specificity, accuracy and dice coefficient, we recommended using DenseNet since it is stated in the literature and our scores validated this fact once more though. However, we determined a significant improvement when lung segmentation model was used. This is because models trained without lung segmentation model somestimes perturbed gradients' attention outside of the lung area. When we colored non-lung pixels black, model could not learn any misinformation from these regions since there was not any action that can be diffentiated, turning its attention into lung areas. Thus, it led the model to force itself learn features inside lung area. Another important observation is choosing threshold value affects the result of our saliency maps. We chose a threshold value to turn our RGB saliency map into a binary image so that we can benchmark our maps and ground truth bounding boxes in terms of dice coefficient. Choosing this threshold affected how our maps and bounding boxes overlap and union. Since pixel values are between 0 and 255, we needed to choose some good representative points. To do so, we examined where pixel intensivities are higher in our saliency maps. Focusing those areas, we detected some threshold values and generated our binary saliency maps. We detected a threshold value of nearly 150 is promising in terms of dice coefficient.

VI. FUTURE WORK

This work, to the best of our knowledge, presents the first benchmarking between saliency maps and ground truth disease localization. Therefore, there are further improvements possible. Using bounding boxed disease localization information we can derive a unique LIME method for our needs. Shading specific parts of an image in LIME method can be originated considering these bounding boxes. Another work can be generating saliency maps in a differentiable manner. Current technique generates these maps without an end-to-end manner. If we can optimize our maps when they are being generated, we can have more robust saliency maps. Another future work can be using attention mechanism during the training. In this way, models pay attention specific areas of images. If we could combine attention mechanism with our bounding box disease information, that would be another novel improvement.

VII. CONCLUSION

In this work we presented the first benchmark between saliency maps and bounding box disease localization. We showed that different architectures can result different saliency maps. Lung segmentation can help our models to generate more interpretable and robust maps. We also proved that the more the models learn a given problem, the more interpretable and meaningful results they generate. This work fills a significant gap in terms of presenting a new benchmark to the literature, doing a comprehensive experiments and being a first step of a wide variety of further improvements.

REFERENCES

[1] Dimopoulos, Konstantinos, et al. "Cardiothoracic ratio from postero-anterior chest radiographs: a simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease." International journal of cardiology 166.2 (2013): 453-457.

[2] CDC, 2017. URL https://www.cdc.gov/features/ pneumonia/index.html.

[3] Centurión, O. A., et al. "Evaluating Cardiomegaly by Radiological Cardiothoracic Ratio as Compared to Conventional Echocardiography." J Cardiol Curr Res 9.2 (2017): 00319.

[4] Yanase, Juri, and Evangelos Triantaphyllou. "A systematic survey of computer-aided diagnosis in medicine: Past and present developments." Expert Systems with Applications 138 (2019): 112821.

[5] Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[6] Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

[7] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition.

[8] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[9] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[10] Çallı, Erdi, et al. "Deep Learning for Chest X-ray Analysis: A Survey." Medical Image Analysis (2021): 102125.

[11] Zar, Heather J., and Thomas W. Ferkol. "The global burden of respiratory disease—impact on child health." (2014): 430-434.

[12] WHO Pneumonia. World Health Organization. (2019), https://www.who.int/news-room/fact-sheets/ detail/pneumonia.

[13] Gabruseva, Tatiana, Dmytro Poplavskiy, and Alexandr Kalinin. "Deep learning for automatic pneumonia detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.

[14] Rahman, Tawsifur, et al. "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray." Applied Sciences 10.9 (2020): 3233.

[15] Roberts, Michael, et al. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans." Nature Machine Intelligence 3.3 (2021): 199-217.

[16] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Proceedings of the 31st international conference on neural information processing systems. 2017.

[17] van der Velden, Bas HM, et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis." arXiv preprint arXiv:2107.10912 (2021).

[18] Code for 1st place solution in Kaggle RSNA Pneumonia Detection Challenge. GitHub website. github. com/i-pan/kaggle-rsna18. Accessed September 17, 2021.

[19] Nguyen, Ha Q., et al. "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations." arXiv preprint arXiv:2012.15029 (2020).

[20] Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[21] Radiological Society of North America. RSNA Pneumonia Detection Challenge. https://www. kaggle.com/c/rsna-pneumonia-detection-challenge. 2018, November (accessed January 1, 2022).

[22] Chouhan, Vikash, et al. "A novel transfer learning based approach for pneumonia detection in chest X-ray images." Applied Sciences 10.2 (2020): 559.

[23] Rahman, Tawsifur, et al. "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray." Applied Sciences 10.9 (2020): 3233.

[24] Yerushalmy, Jacob. "Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques." Public Health Reports (1896-1970) (1947): 1432-1449.

[25] "Sensitivity and Specificity." Wikipedia, Wikimedia Foundation, 22 December 2022, en.wikipedia.org/wiki/Sensitivity$and specificity cite \mathbf{note-1}$

[26] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

[27] Jiang, Peng-Tao, et al. "LayerCAM: Exploring Hierarchical Class Activation Maps for Localization."

[28] 8] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131.

[29] N. Mahomed, B. van Ginneken, R.H. Philipsen, J. Melendez, D.P. Moore, H. Moodley, T. Sewchuran, D. Mathew, S.A. Madhi, Computer-aided diagnosis for world health organization-defined chest radiograph primary-endpoint pneumonia in children, Pediatr. Radiol. (2020) 1–10.

[30] T.D. Team, Pneumonia detection in chest radiographs, 2018, arXiv:1811.08939.

[31] I. Sirazitdinov, M. Kholiavchenko, T. Mustafaev, Y. Yixuan, R. Kuleev, B. Ibragimov, Deep neural network ensemble for pneumonia localization from a large-scale chest x-ray database, Comput. Electr. Eng. 78 (2019) 388–399.

[32] N. Mahomed, B. van Ginneken, R.H. Philipsen, J. Melendez, D.P. Moore, H. Moodley, T. Sewchuran, D. Mathew, S.A. Madhi, Computer-aided diagnosis for world health organization-defined chest radiograph primary-endpoint pneumonia in children, Pediatr. Radiol. (2020) 1–10.

[33] J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano, E.K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study, PLoS Med. 15 (11) (2018).

[34] Rula Amer andothers. COVID-19 in CXR: from Detection and Severity Scoring to Patient Disease Monitoring. 2020. arXiv: 2008.02150 [eess.IV].

[35] Lane, H. Chad, and Sidney K. D'Mello. "Uses of physiological monitoring in intelligent learning environments: A review of research, evidence, and technologies." Mind, Brain and Technology (2019): 67-86.

[36] Junji Shiraishi andothers. "Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule". in: American Journal of Roentgenology 174.1 (january 2000), pages 71–74. doı: 10.2214/ajr.174.1.1740071.

[37] B. van Ginneken, M.B. Stegmann and M. Loog. "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database". in: Medical Image Analysis 10.1 (2006), pages 19–40.

[38] Sameer Antani Yì-Xiang J. Wang Pu-Xuan Lu Stefan Jaeger Sema Candemir and George Thoma1. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. 2014. doı: 10.3978/j.issn.2223-4292.2014.11.20.

[39] v7labs. COVID-19 Chest X-ray Dataset. https://github.com/v7labs/covid-19-xray-dataset. 2020, January (last access 12 January, 2022).

[40] rsummers11.Lung_Segmentation_XLSor. https://github.com/rsummers11/CADLab/tree/master/Lung_Segmentation_XLSor. 2020, April (last access 12 January, 2022).