Automatic lung abnormality detection using deep convolutional neural networks

Justin Hwang

Abstract

In medical image analysis, machine learning models can help healthcare professionals with the diagnosing and monitoring of lung abnormalities, which are prevalent and diverse in nature. This paper presents a comprehensive approach to lung abnormality detection using deep learning techniques. The proposed methodology leverages the power of convolutional neural networks (CNNs) to accurately and efficiently identify the presence of various lung abnormalities from chest radiographs. The research focuses on a dataset comprising a diverse range of lung abnormalities. Deep learning architectures have been used to process 1,740 chest radiographs. The model yielded an accuracy of 93.10% after various preprocessing steps to ensure quality, even with a relatively small sample size. The results demonstrate that deep learning can effectively identify individuals with lung disease, streamlining the process of diagnosis even in the absence of direct medical supervision. This procedure holds the potential to enhance personalized prevention and treatment approaches, thereby contributing to life-saving measures.

1 Introduction

Medical imaging is critical in the diagnosis of disease. Through the precise visualization of internal anatomical structures, techniques like computed tomography (CT) scans and chest X-rays have revolutionized disease diagnosis and treatment planning [1][2]. Among the numerous medical applications, the detection of lung abnormalities emerges as a task of paramount significance.

Lung abnormalities encompass a wide spectrum of conditions. Timely and accurate identification of these abnormalities is vital for patient care, as it directly influences treatment strategies and clinical outcomes [3]. However, the intricate anatomy of the lungs, coupled with the subtlety of pathological variations, presents challenges for medical professionals in interpreting images and making precise diagnoses [4].

In response to these challenges, machine learning technologies have emerged as a transformative diagnosis technique in medical imaging. These technologies offer the promise of enhancing the diagnostic accuracy and efficiency of lung abnormality detection. By training on large datasets of medical images, machine learning models can learn intricate patterns and nuances that may elude human observation [5]. Furthermore, the model's automatic functionality

enables convenient and timely identification of lung diseases. Although not a substitute for professional medical diagnosis, the model is well-suited for individual screenings [6].

Here, we apply convolutional neural networks for lung abnormality detection, providing insights into the current landscape, methodologies, and empirical results that underscore their potential in transforming medical practice. Notably, while existing papers predominantly focus on devising segmentation and detection algorithms, we incorporate outlier data identification through lung segmentation masks. This innovative integration underscores the importance of data quality within medical image analysis.

2 Method

2.1. Lung segmentation

Lung segmentation was done by obtaining a dataset of lungs and their corresponding ground truth masks, applying pre-processing transformations, and training a U-Net model to segment the lungs.

The dataset used for lung segmentation has been publicly documented in the U.S. National Library of Medicine [7] and originates from two sources. The radiographic images have been amassed in partnership with the Department of Health and Human Services at Montgomery County, Maryland, as well as Shenzhen No. 3 People's Hospital located in Guangdong Medical College, Shenzhen, China. Both sets contain manually segmented lung masks by radiologists for evaluation. In total, 5,290 pairs of chest X-ray images and their corresponding lung masks have been collected for the process of lung segmentation.

Contrast normalization was performed by expanding the intensity value range of the image to encompass a predefined span of values. The application of this filter contributed to an elevated distinction between lung and bone constituents. Subsequently, the gaussian blurring transformation was employed, entailing the application of a mathematical function to the image. This imparted a gentle blurring effect, effectively mitigating noise and finer details within the image. Additionally, median filtering was applied to further eliminate noise. This entailed the exclusion of pixels with luminosities that substantially deviate from neighboring pixels, subsequently replacing the central pixel with the median brightness value derived from the examined pixels. Ultimately, the images were resized to dimensions of 256 by 256 pixels, ensuring uniformity in model training while balancing computational efficiency with loss of image quality.

The dataset encompassed a diverse collection of 5,290 images. To prevent overfitting and to assess the model's performance, the dataset was divided into an 80% training subset and a 20% testing subset.

For evaluating the model's performance and steering the training process, the Dice coefficient loss function was chosen. It is based on the Dice coefficient, also known as F1 score, a statistical formula quantifying the similarity of two samples. It is defined as twice the area of the intersection of A and B, divided by the sum of the areas of A and B. In the lung segmentation

task at hand, minimizing this loss function promotes the accurate alignment of predicted and actual segmentations.

In compiling the model, a batch size of 16 was adopted for over 20 epochs, optimizing the utilization of computational resources while facilitating efficient weight updates. Furthermore, the Adam optimizer was employed. Adam's adaptive learning rate and momentum-based strategies expedited convergence, enabling efficient weight updates across the network's layers. A learning rate of 1e-4 was experimentally chosen in order to balance convergence speed with overshooting.

2.2. Identifying outlier data

Identifying outlier image data is necessary to ensure the quality of the overarching step, detecting lung abnormality. To weed out such samples, a two-step algorithm was used taking into account the lung segmentation masks.

First, the mask was cleaned up using morphological functions. The opening operation from the cv2 library helps to separate connected components, fill gaps, and smooth the mask. It removes small noise and fine structures while preserving the overall shape of the larger objects. Then, the clean_up_small_objects function from Scikit-image was applied to eliminate artifacts that may not be relevant to the analysis.

After refining the mask, a criterion was designed to detect atypical X-ray samples. First, if the region with the greatest area comes within 60% of that of exactly one other region, then the image passed. In addition, the largest component had to measure no greater than 20,000 pixels (out of a possible 65,536 pixels in a 256 by 256 image) for clearance.

2.3. Region of interest (ROI) extraction

The predicted lung mask was used to make a “bounding box”, formed by taking four points: the minimum and maximum horizontal and vertical coordinates of the highlighted areas in the mask. These points form a box that tightly encloses desired regions of the segmented mask. To capture areas immediately around the lungs, padding was added uniformly for each side of the bounding box by a value of 25 pixels. This region was then resized to the desired 256 by 256 pixels, without necessarily preserving original aspect ratios. These transformations were applied to form the corresponding cropped region from the original image, resized to match the dimensions of the ROI.

2.4. Lung abnormality detection

The dataset used for lung abnormality detection was sourced from a Kaggle competition, provided by the Vingroup, a data science and artificial intelligence research group based in Vietnam [8]. All images were annotated by experienced radiologists for the presence of 14 critical radiographic abnormalities as listed below:

- Aortic enlargement
- Atelectasis

- Calcification
- Cardiomegaly
- Consolidation
- ILD
- Infiltration
- Lung Opacity
- Nodule/Mass
- Other lesion
- Pleural effusion
- Pleural thickening
- Pneumothorax
- Pulmonary fibrosis

In total, 2,100 images have been used prior to filtering out poor data.

Similar to the previous section on pre-processing lung images for segmentation (Section 2.1.2), analogous techniques have been employed to standardize and optimize the model training process. Contrast normalization, gaussian filter, and median filter have all been used to enhance image contrast and reduce noise. Furthermore, the images were resized to dimensions of 512 by 512 pixels in order to compromise between memory storage and maintaining high image quality. This particular resizing will be important for extracting the ROI as its dimensions will not exceed that of the original image. In addition, the procedure mentioned in Section 2.2 has been applied to filter out unwanted image data.

The dataset consisted of 1,740 images, with 80% allocated for the training subset and a 20% for the testing subset.

The loss function chosen for this stage was binary cross-entropy with logits loss, a popular choice for binary classification tasks. It first applies the sigmoid activation function, which converts raw outputs to a probability distribution between 0 and 1. Then, the values are fed to a standard binary cross-entropy loss function, which calculates the difference between predicted probabilities and actual labels. This measures how well a model's predictions match the true outcomes, facilitating the accurate classification of lung abnormality.

In compiling the model, a batch size of 16 was adopted for over 20 epochs, optimizing the utilization of computational resources while facilitating efficient weight updates. The Adam optimizer was used with a learning rate of 1e-3.

3 Results

3.1. U-Net achieves highly accurate lung segmentation

The X-ray images have been amassed in partnership with the Department of Health and Human Services at Montgomery County, Maryland, as well as Shenzhen No. 3 People’s Hospital located in Guangdong Medical College, Shenzhen, China. The deliberate selection of these two sources contributes to the diversification of patient data across varied geographic regions.

The U-Net, a widely adopted CNN architecture, plays a pivotal role in the lung segmentation process due to its aptitude for precise feature extraction and spatial contextualization. Originally proposed for biomedical image segmentation, the U-Net architecture has demonstrated its efficacy across various segmentation tasks, making it a suitable choice for delineating lung structures within medical images [9].

In the assessment of lung segmentation efficacy, three key evaluation metrics have been employed: binary accuracy, dice coefficient (also known as F1 score), and the area under the receiver operating characteristic curve (AUROC).

Binary accuracy gauges the proportion of accurately predicted pixel values within the binary segmentation mask compared to the ground truth mask. This metric is determined by tallying the count of correctly identified pixels, both foreground (lung) and background (non-lung), and dividing it by the total count of pixels in the mask. Binary accuracy provides a comprehensive assessment of the model's performance with respect to pixel-level classification. In the experiment, the calculated binary accuracy achieved a value of 97.73%.

The dice coefficient, often referred to as the F1 score, quantifies the agreement between the predicted segmentation and the ground truth mask. This metric is calculated by assessing the overlap between the two masks. The dice coefficient provides insights into the model's ability to balance precision and recall, making it a valuable measure for segmentation tasks. In the experimentation, the computed dice coefficient yielded a value of 95.91%.

The AUROC assesses the model's ability to discriminate between foreground and background pixels across various probability thresholds. This metric captures the trade-off between true positive and false positive rates and lies between 0 and 1, where a perfect AUROC score of 1 indicates flawless discrimination, while a score of 0.5 signifies random guessing. The computed AUROC achieved a value of 99.82% **(Figure 1)**.

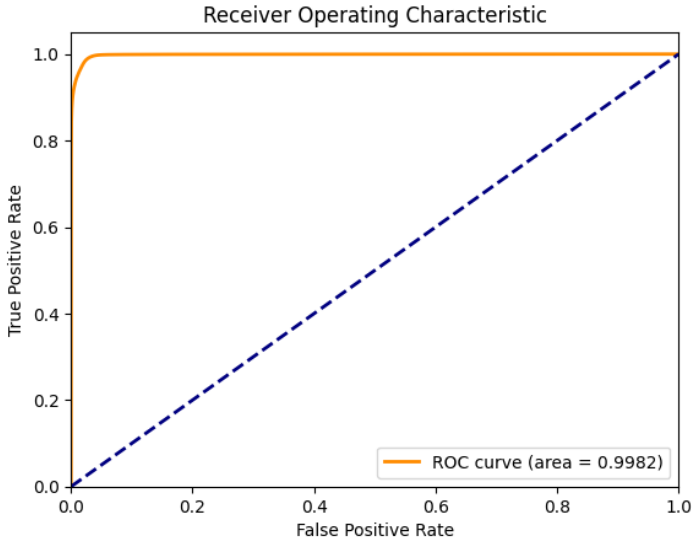

Figure 1: Lung segmentation AUROC curve

A significant portion of the predicted masks closely resemble their corresponding ground truth counterparts, underscoring the model's substantial accuracy **(Figure 2)**. Out of the 1,038 validation images, analysis of the worst six predictions by dice coefficient revealed that those X-rays have been taken too far or have been taken at unusual angles **(Figure 3)**. Furthermore, inverted images, such as the one in the last column, became a more prevalent issue in the process of lung abnormality detection.

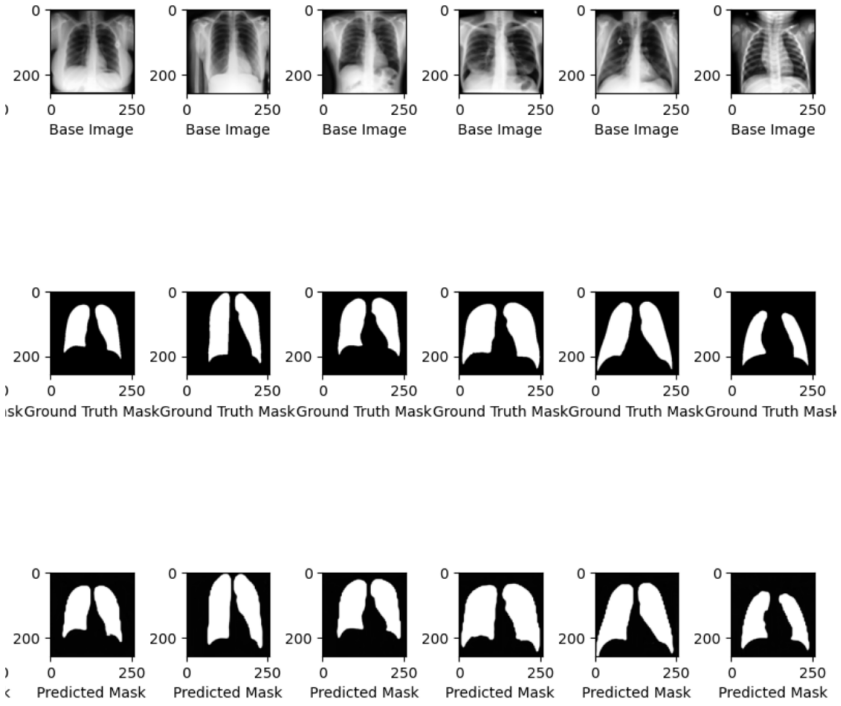

Figure 2: Sample predicted lung masks

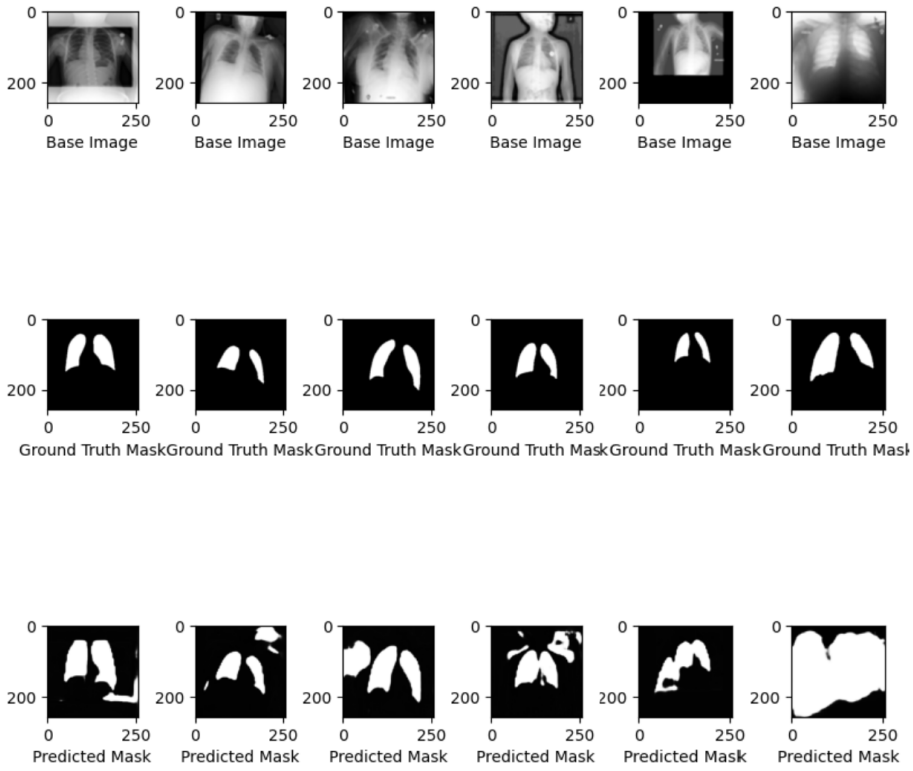

Figure 3: Six worst predicted lung masks by dice coefficient

3.2. Effective outlier data identification

The two-part criterion designed to detect atypical X-ray samples have effectively weeded out undesired data samples. First, if the region with the greatest area comes within 60% of that of exactly one other region, then the image passed. This case accounted for the vast majority of X-ray images, as accurately predicted lung segmentations reliably fulfilled the requirement.

In cases where the segmentation yielded connected lung structures, despite applying morphological operations, the largest component had to measure no greater than 20,000 pixels (out of a possible 65,536 pixels in a 256 by 256 image) for clearance. This condition effectively filtered out chest X-ray scans that did not match the anticipated color distribution of bright bone matter and dark air and lungs.

In the Vingroup dataset, 360 out of 2,100 images have been removed from the dataset, the vast majority of which did not match the expected color scheme.

3.3. ROI mitigates confounding effects

Confounding effects refer to external factors or variables that can influence the observed outcomes, potentially leading to misleading interpretations or conclusions [10]. These effects can inadvertently impact the accuracy or reliability of the predictions in the subsequent step, lung abnormality detection. Addressing confounding effects is crucial to ensure that the performance metrics of the model will accurately reflect their true capabilities without irrelevant bias. In the

task of lung abnormality detection, confounding effects could arise from various sources, such as variations in image quality or radiographic watermarks, possibly introducing variability in the data that the model processes. As a result, its predictions could be influenced by these extraneous factors rather than solely reflecting its ability to accurately identify abnormal lungs.

To mitigate these effects, the ROI has been created to only focus on relevant areas, namely the lung and its immediately surrounding areas. This region was extracted by performing a series of image manipulation techniques with the aid of predicted lung guidelines.

3.4. ResNet-18 achieves high lung abnormality detection performance

ResNet-18, short for Residual Network-18, is a CNN architecture that has exhibited exceptional performance in various computer vision tasks [11]. ResNet-18 offers a solution to the problem of vanishing gradients in deep neural networks, enabling the training of significantly deeper networks without degradation of performance.

The evaluation of the lung abnormality model is crucial in understanding its effectiveness in identifying lung abnormalities and distinguishing between normal and abnormal lung images. The key performance metrics include the AUROC, binary accuracy, specificity, and sensitivity.

Binary accuracy measures the proportion of correctly predicted samples among all samples. The model achieved a binary accuracy of 93.10%, highlighting its high accuracy in classifying lung images as normal or abnormal. This metric provides a clear indication of the model's ability to make accurate predictions on the dataset.

The model achieved an AUROC score of 95.21% **(Figure 4)**, indicating its strong ability to distinguish between normal and abnormal lung images. A higher AUROC score signifies a better overall performance of the model in terms of correctly classifying samples from both classes.

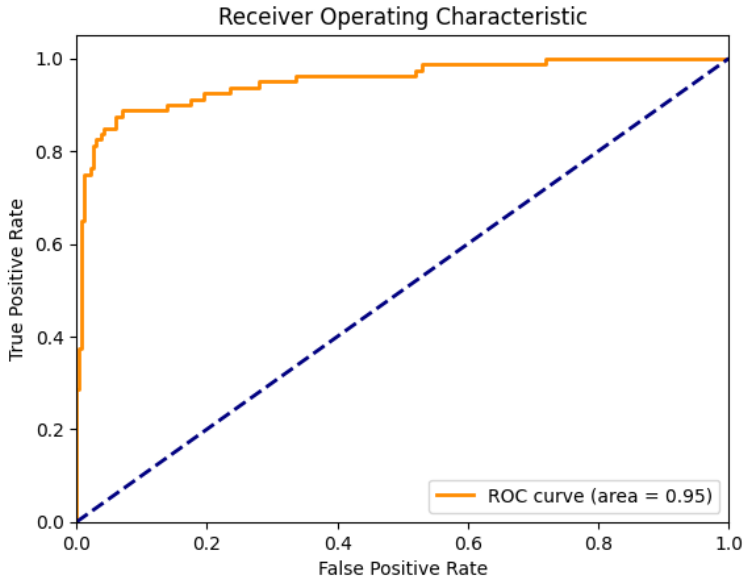

Figure 4: Lung abnormality detection AUROC curve

Specificity, also known as the true negative rate, represents the proportion of correctly predicted negative (normal) samples among all actual negative samples. A higher specificity value indicates a lower rate of false positives, thus highlighting the model's capability to minimize misclassifications of normal cases as abnormal. The model demonstrated a specificity of 97.39%.

Sensitivity, also referred to as the true positive rate or recall, measures the proportion of correctly predicted positive (abnormal) samples among all actual positive samples. A higher sensitivity value underscores the model's capability to detect true positive cases effectively. The model exhibited a sensitivity of 78.75%.

The presented heatmaps offer a visual representation of the model's attention to different regions, depicted through varying intensities of color. Deeper shades indicate higher significance. Notably, the model assigns considerable importance to the contour of the right lung, particularly in proximity to the heart **(Figure 5, 6)**.

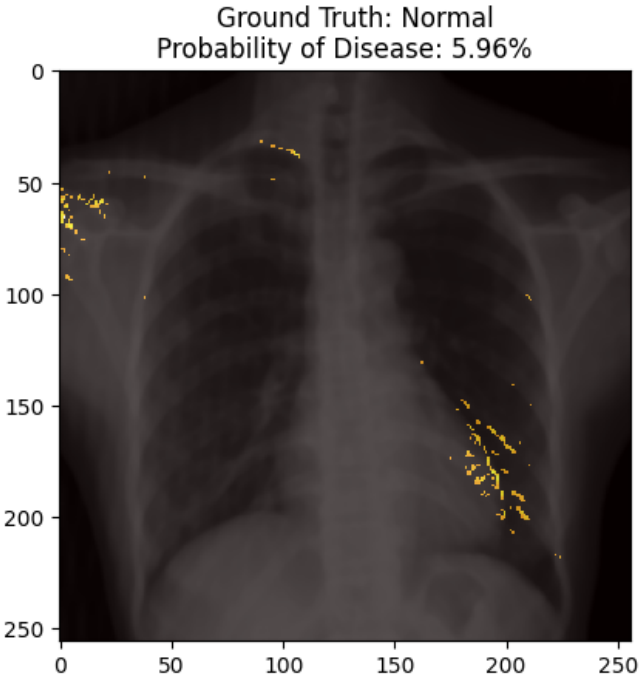

Figure 5: Lung abnormality detection heatmap for normal lungs

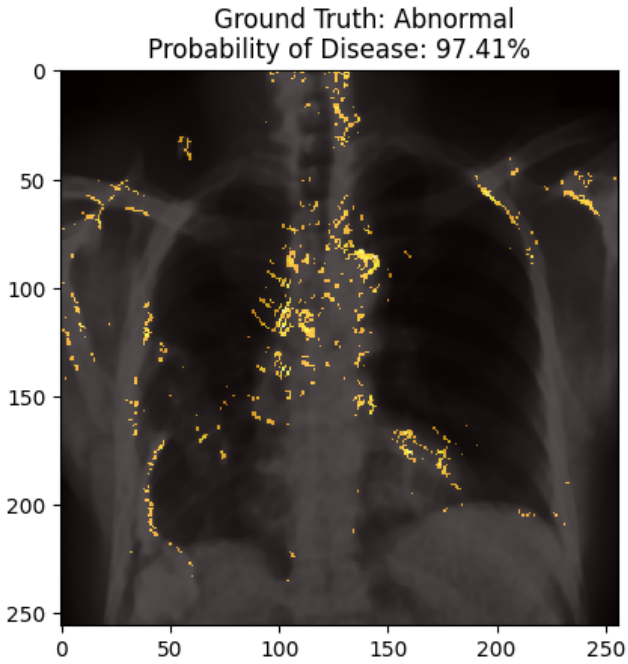

Figure 6: Lung abnormality detection heatmap for abnormal lungs

4 Discussion

The presented work revolves around the utilization of deep learning architectures for the identification of lung abnormalities from chest radiographs. The achieved accuracy of 93.10% underscores the model's proficiency in recognizing lung disease instances within the dataset. The success of the model can be attributed to its capability to learn complex patterns and features from the radiographic images, which are often imperceptible to the human eye.

The high AUROC score of 95.21% further accentuates the model's robustness in distinguishing between normal and abnormal cases. This performance metric is indicative of the model's effective classification ability.

The calculated sensitivity of 78.75% highlights an important aspect of the model's performance in the context of lung abnormality detection. Sensitivity, also known as the true positive rate or recall, measures the proportion of actual positive cases correctly identified by the model. In the context of medical diagnostics, sensitivity is crucial as it indicates the model's ability to correctly identify individuals with lung abnormalities, ensuring that genuine cases are not missed.

The observed sensitivity implies that the model successfully identified approximately 78.75% of the actual lung abnormality cases present in the dataset. However, the remaining 21.25% of true positive cases were not detected by the model, which raises concerns as missed detections could potentially delay timely medical intervention for patients with lung abnormalities [12].

A sensitivity of 78.75% suggests room for improvement in the model's ability to detect lung abnormalities, especially in cases where subtle or atypical signs are present. Strategies to enhance sensitivity involve adjusting the classification threshold or expanding the dataset through the means of obtaining additional data or incorporating data augmentation techniques [13]. Furthermore, employing more complex architectures or fine-tuning the existing model might allow it to capture intricate features that it currently misses.

Moreover, the generated heatmaps provide insight into the areas of interest that the model deems significant during classification. The model's emphasis on the outline of the right lung, particularly near the heart, is indicative of its ability to recognize patterns of abnormalities often associated with specific anatomical regions. Nonetheless, the variation in heatmap intensities highlights potential areas for refinement and further model enhancement. Introducing techniques like Grad-CAM++ might offer improved visualizations by focusing on higher-level features and allowing for better localization of abnormalities [14].

The automatic and timely diagnosis potential of the model is noteworthy. However, it is crucial to emphasize that the model's predictions should serve as complementary information to professional medical diagnoses rather than a substitute. Clinical expertise remains vital in the interpretation of results and decision-making regarding patient care [15].

In conclusion, this study presents a promising application of deep learning in lung abnormality detection. The achieved results underscore the model's potential utility in assisting medical professionals and enhancing diagnostic processes.

5 References

1. Smith-Bindman, R., Lipson, J., Marcus, R., Kim, K. P., Mahesh, M., Gould, R., Berrington de González, A., & Miglioretti, D. L. (2009). Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. *Archives of internal medicine*, *169*(22), 2078–2086. https://doi.org/10.1001/archinternmed.2009.427
2. National Research Council. 2006. *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11340.
3. MacMahon, H., Naidich, D. P., Goo, J. M., Lee, K. S., Leung, A. N. C., Mayo, J. R., Mehta, A. C., Ohno, Y., Powell, C. A., Prokop, M., Rubin, G. D., Schaefer-Prokop, C. M., Travis, W. D., Van Schil, P. E., & Bankier, A. A. (2017). Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology*, *284*(1), 228–243. https://doi.org/10.1148/radiol.2017161659
4. Farias, L. P. G., Fonseca, E. K. U. N., Strabelli, D. G., Loureiro, B. M. C., Neves, Y. C. S., Rodrigues, T. P., Chate, R. C., Nomura, C. H., Sawamura, M. V. Y., & Cerri, G. G. (2020). Imaging findings in COVID-19 pneumonia. *Clinics (Sao Paulo, Brazil)*, *75*, e2027. https://doi.org/10.6061/clinics/2020/e2027

5. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056
6. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Patel, B. N., Yeom, K. W., Shpanskaya, K., Blankenberg, F. G., Seekins, J., Amrhein, T. J., Mong, D. A., Halabi, S. S., Zucker, E. J., Ng, A. Y., … Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine*, *15*(11), e1002686. https://doi.org/10.1371/journal.pmed.1002686
7. Jaeger, S., Candemir, S., Antani, S., Wáng, Y. X., Lu, P. X., & Thoma, G. (2014). Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, *4*(6), 475–477. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20
8. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* (pp. 234-241). Springer International Publishing.
9. "VinBigData Chest X-ray Abnormalities Detection." *Kaggle*, https://www.kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection/data . Accessed 17 August 2023.
10. Phillips, C. V., & Goodman, K. J. (2004). The missed lessons of Sir Austin Bradford Hill. *Epidemiologic perspectives & innovations : EP+I*, *1*(1), 3. https://doi.org/10.1186/1742-5573-1-3
11. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
12. Brady A. P. (2017). Error and discrepancy in radiology: inevitable or avoidable?. *Insights into imaging*, *8*(1), 171–182. https://doi.org/10.1007/s13244-016-0534-1
13. Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J Big Data* 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0
14. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.
15. Rajaraman, S., Kim, I., & Antani, S. K. (2020). Detection and visualization of abnormality in chest radiographs using modality-specific convolutional neural network ensembles. *PeerJ*, *8*, e8693. https://doi.org/10.7717/peerj.8693