Explainability of CNN Classification Models Using CycleGAN and Their Application to Medical Imaging

Taiga Nakajima[1], Yoshua Kazukuni Nomura[2], Narufumi Suganuma[2], and Shinichi Yoshida[1]

[1]Kochi University of Technology, Kochi 782-8502, Japan
[2]Kochi University, Kochi 783-8505, Japan
275107g@gs.kochi-tech.ac.jp

Abstract. In recent years, there has been a rapid increase in interest regarding the widespread application of Convolutional Neural Networks (CNNs) in Computer-Aided Diagnosis (CAD) and image-based diagnosis. However, CNN-based diagnostic approaches still face numerous challenges in terms of interpretability. Previous studies have proposed the use of CycleGAN to analyze the classification processes of CNNs, suggesting its potential to enhance interpretability. CycleGAN is characterized by its ability to transform specific parts of an image without altering the background, allowing it to capture more detailed information such as differences in shapes and patterns within regions, compared to traditional methods like Grad-CAM. This study aims to apply CycleGAN to the disease pneumoconiosis to visualize which parts of the image the classification model focuses on when making its determinations. The results reveal that the brightness across the entire lung field changes before and after transformation, suggesting that the classification model may be focusing on the degree of brightness within the lung region when identifying pneumoconiosis.

Keywords: Computer Aided Diagnosis(CAD), Convolutional Neural Network(CNN), Generative Adversarial Network, CycleGAN, explainable artificial intelligence(XAI), Pneumoconiosis

1 Introduction

Pneumoconiosis is a lung disease caused by the long-term inhalation and accumulation of dust and fine particles, leading to various pathological changes in the lungs[1]. Pneumoconiosis is identified using chest X-ray images by detecting the presence of ground-glass opacities. However, because these opacities often overlap with structures such as blood vessels, the interpretation process is time-consuming and burdensome for physicians. As a result, research on Computer-Aided Diagnosis (CAD) systems has been gaining attention. CAD includes medical image analysis and are used for a second opinion to support physicians in their diagnosis with its objective decision. Convolutional Neural

Networks (CNN) is widely employed[2]. As examples of high-precision diagnosis using CNN, Manickavasagam et al. have proposed a model for detecting lung nodules with an accuracy of 98.88%[3], while Alkurdi et al. proposed a model for detecting breast cancer with an accuracy of 98%[4]. However, CNN-based diagnosis have a difficulty of interpretability of those results. Therefore, it is important to ensure that diagnostic results are explainable. To achieve this, the explainable AI (XAI) is required. XAI is a technique that explains how an AI system makes specific decisions or predictions in a way that humans can understand. In the medical field, it is essential to clearly present the basis for diagnostic results and the features or regions the AI used for its decisions. This enables physicians to trust the AI-generated results and incorporate them into their diagnoses and treatment plans. One example of XAI is Class Activation Map (CAM) [5]. CAM are used for CNN results to identify regions contributing to the classification results. CAM provides interpretability to CNN, however, CAM provides only the information of location of regions which affects to the result and does not provide other information such as shape, intensity, or pattern. In the medical analysis, the differences of such information are important for the interpretability. In the study by Tsutsui and Yoshida [6], they employed CycleGAN [7], a type of Generative Adversarial Network (GAN) [9], as an analytical method to acquire both the regions contributing to classification and the differences in shapes and patterns within those regions. In previous studies, high-accuracy classification results (95.43%) were achieved by inputting images of the extracted lung field regions into a CNN. Therefore, in this study, we aim to apply CycleGAN to images of pneumoconiosis, a lung disease, to identify the regions that contribute to the classification of pneumoconiosis, as well as their shapes and patterns.

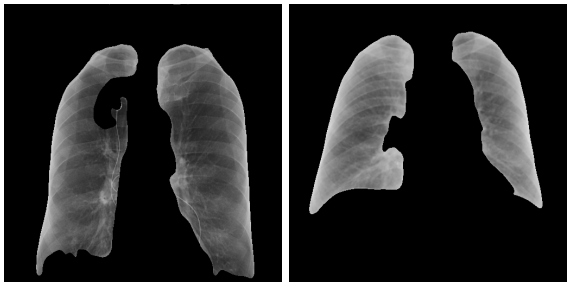
2 Method

2.1 Explainable CNN using CycleGAN

We use the explainable artificial intelligence technique proposed by Tsutsui and Yoshida [6], which uses CycleGAN. This method learns the transformation of images between two classes, which mean no-finding class and pneumoconiosis class. Then the an image is converted to the other class. The difference of images are computed between post- and pre-transformation.

2.2 Dataset

For medical applications, we use chest X-ray images of pneumoconiosis patients as the dataset. In order to improve the quality of diagnose pneumoconiosis on the lung field region, we use image segmentation to extract the lung field regions using U-net(Fig. 1). The original images were obtained from the chest X-ray image datasets provided by NIOSH (National Institute for Occupational Safety and Health), Kochi University Medical School (KM), and NIHCC (National



(a) healthy **(b)** pneumoconiosis

Fig. 1: Pneumoconiosis images

Institutes of Health Clinical Center) [14]. The dataset included a total of 234 images, comprising 117 images each of pneumoconiosis patients and healthy individuals. To ensure consistency during the training process, the images were standardized to a size of 512x512 pixels. Of these, 212 images were used for training and 22 for testing.

2.3 Model structure

CycleGAN consists of a discriminative model and a generative model, which are learned adversarial. In this section we describe the generative model and the discriminative model.

Generative model The generative model consists of an encoder (feature extraction component), 9 residual blocks, a Self-Attention layer, and a decoder (output component). Instance Normalization is used for normalization, and ReLU is employed as the activation function. The input images are grayscale images with dimensions of 512×512. In the encoder, the input images undergo downsampling using a stride-1 convolutional layer followed by two stride-2 convolutional layers. Subsequently, various image features are extracted using 9 residual blocks. The structure of the generator is shown in Table. 1.

Discriminative model The discriminative model extracts features from images and outputs a 30×30 feature map. Batch Normalization is employed for normalization, and the ReLU activation function is used. The structure of the discriminator is detailed in Table 2.

2.4 Experiment

After training CycleGAN using the dataset, the images are transformed between pneumoconiosis and healthy classes. Then the difference between the original

Table 1: Internal structure of the encoder

Encoder			
kernel size	connection method	stride	Number of filters
7×7	Convolution; Instance Normalization; ReLU	1	64
3×3	Convolution; Instance Normalization; ReLU	2	128
3×3	Convolution; Instance Normalization; ReLU	2	256

Transformer
Residual blocks
Residual blocks1
Residual blocks2
Residual blocks3
...
Residual blocks9

Decoder			
kernel size	connection method	stride	number of filters
3×3	Transposed Convolution; Instance Normalization; ReLU	2	124
3×3	Transposed Convolution; Instance Normalization; ReLU	2	64
7×7	Transposed Convolution; Instance Normalization; ReLU	1	3 or 1

Table 2: Internal structure of the discriminator model.

kernel size	connection method	stride	number of filters
3×3	Convolution; Batch Normalization; ReLU	1	64
3×3	Convolution; Batch Normalization; ReLU	1	64
2×2	Max Pooling	2	64
3×3	Convolution; Batch Normalization; ReLU	1	128
3×3	Convolution; Batch Normalization; ReLU	1	128
2×2	Max Pooling	2	128
3×3	Convolution; Batch Normalization; ReLU	1	256
3×3	Convolution; Batch Normalization; ReLU	1	256
3×3	Convolution; Batch Normalization; ReLU	1	256
2×2	Max Pooling	2	256
3×3	Convolution; Batch Normalization; ReLU	1	512
3×3	Convolution; Batch Normalization; ReLU	1	512
3×3	Convolution; Batch Normalization; ReLU	1	512
2×2	Max Pooling	2	512
3×3	Convolution; Batch Normalization; ReLU	1	512
3×3	Convolution; Batch Normalization; ReLU	1	512
3×3	Convolution; Batch Normalization; ReLU	1	512
3×3	Convolution; Batch Normalization; ReLU	1	1

and transformed images is calculated. This difference is converted into an image to evaluate the change pattern before and after the transformation. The difference image is designed to appear brighter when pixel values increase and

Table 3: Hyperparameters of dataset(CycleGAN)

epoch	Batch Size	Optimizer	Learning Late
700	1	Adam	2×10^{-4}

darker when pixel values decrease. Gray regions show no change between pre- and post-transformation. In addition, the mean and standard deviation of the pixel values of the original and converted lung images are calculated to evaluate the influence of the pixel values during conversion.

Note that in CycleGAN training, the hyperparameters are set as shown in table 3. Set the number of epochs to 700 and the batch size to 1. Also, use the Adam optimizer and set the learning rate for both the generative and discriminative models to 2×10^{-4}.

3 Result

The results of the experiments conducted using the prepared datasets are detailed below. Using CycleGAN, we performed transformations from pneumoconiosis to healthy and from healthy to pneumoconiosis, and we describe the changes in appearance and pixel value variations observed in the resulting images.

3.1 transformation results

In the transformation from pneumoconiosis to healthy, a general decrease in the brightness of the lung fields was observed(Fig. 3). This is attributed to the emergence of characteristic brightness patterns typical of healthy lungs in the transformed images. Conversely, in the transformation from healthy to pneumoconiosis, an increase in the brightness of the lung fields was noted(Fig. 5). This is likely due to the addition of features associated with pneumoconiosis, which result in higher overall brightness in the lung fields.

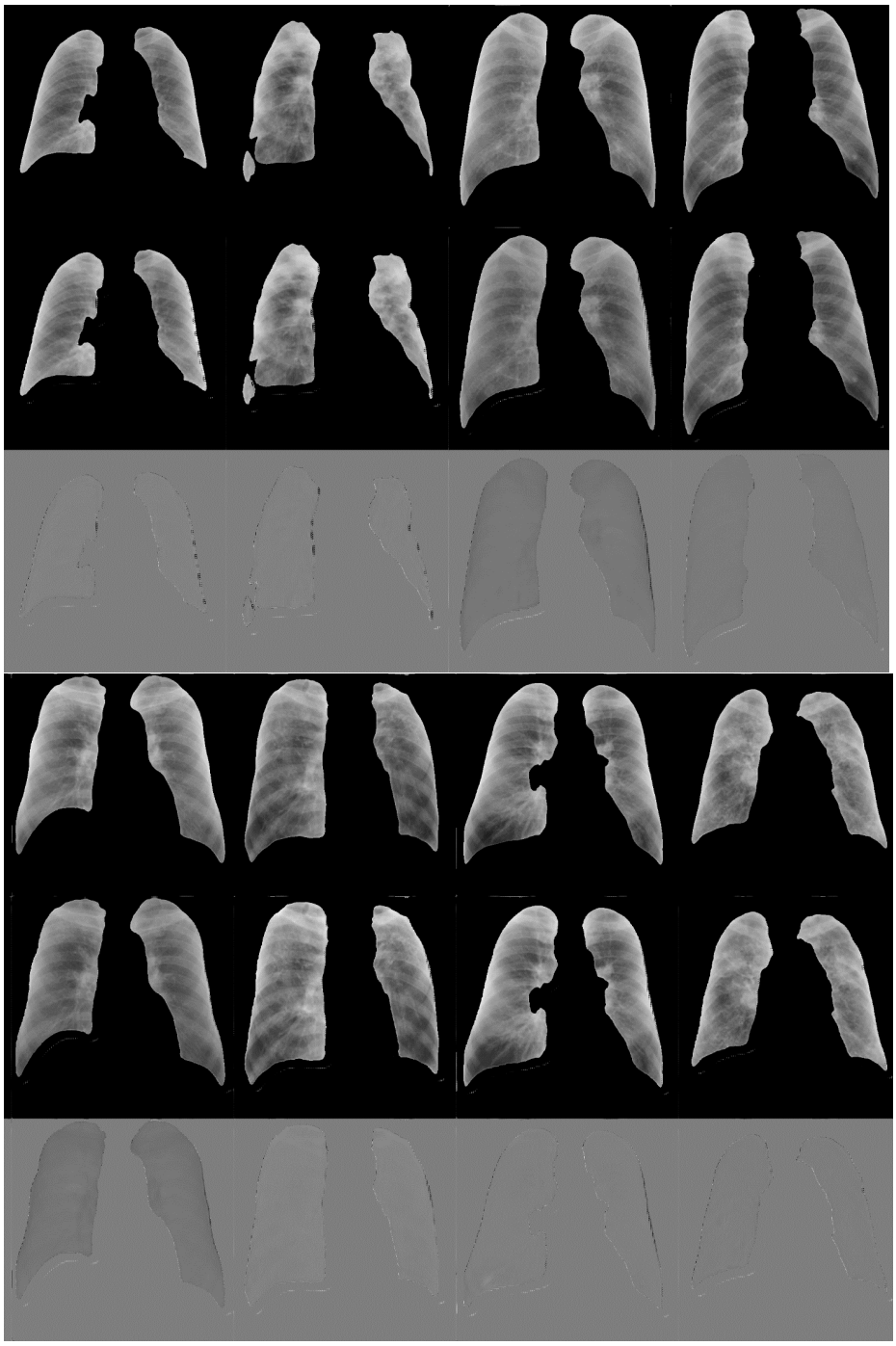


Fig. 2: Transformation from pneumoconiosis to healthy (1/2)

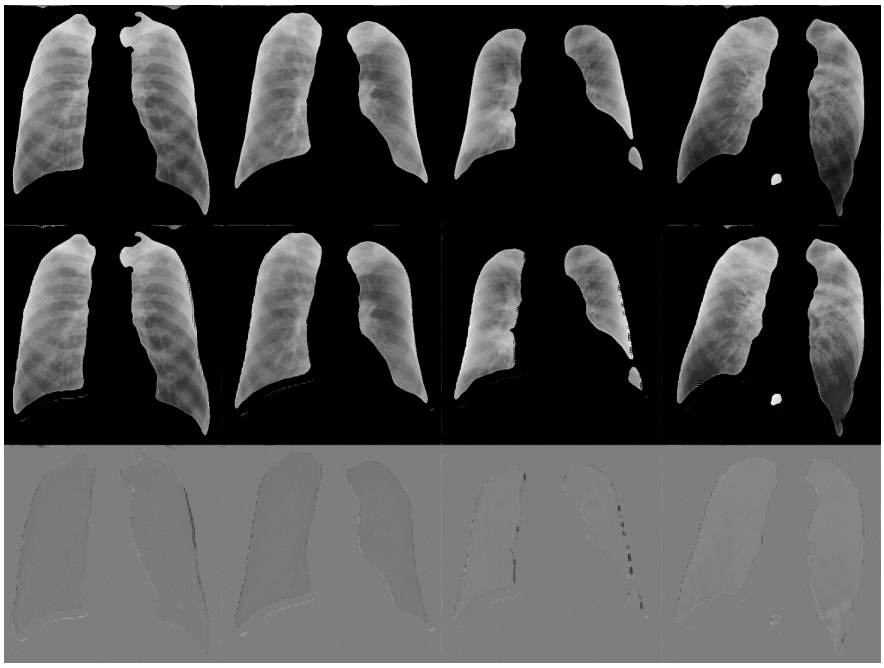


Fig. 3: Transformation from pneumoconiosis to healthy (2/2)

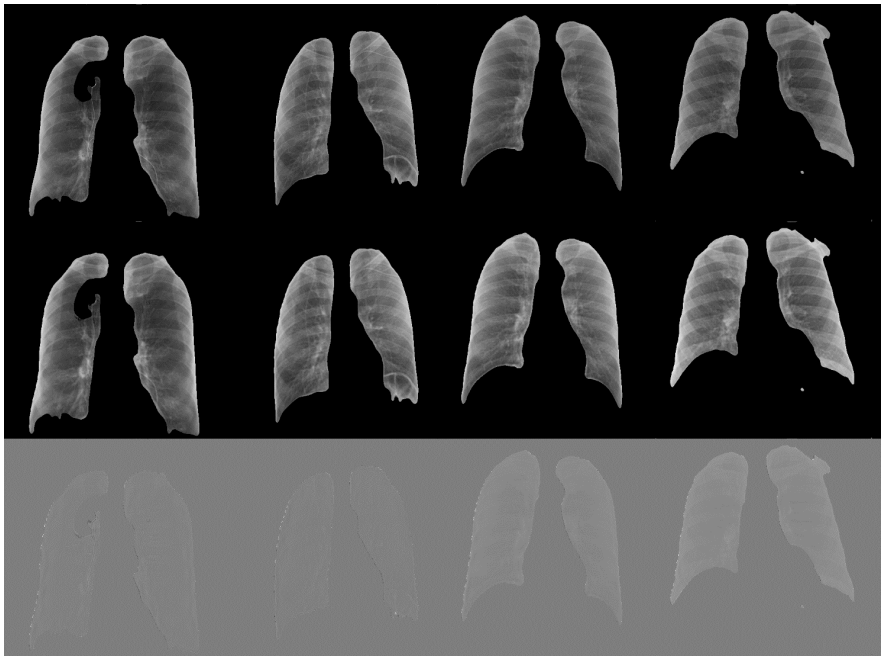


Fig. 4: Transformation from healthy to pneumoconiosis (1/2)

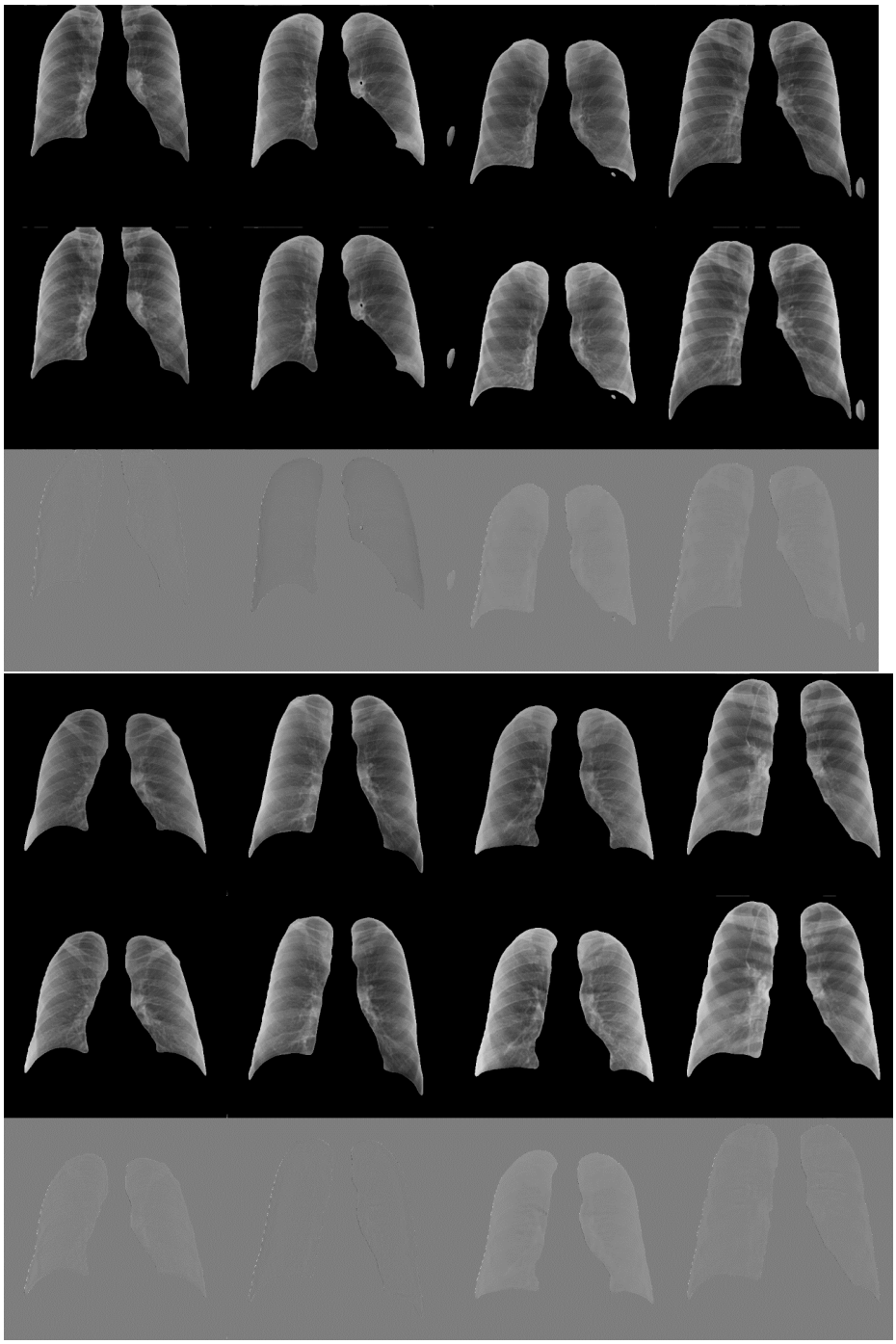


Fig. 5: Transformation from healthy to pneumoconiosis (2/2)

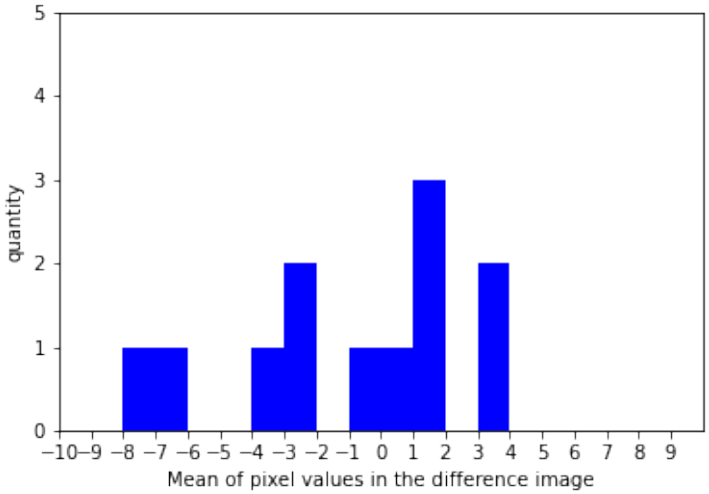


Fig. 6: Distribution of pixel value changes in images transformed from pneumoconiosis to healthy

3.2 Distribution of pixel value changes in images transformed

Describe the variations in pixel values. Observing trends in changes by averaging the pixel values of difference images before and after transformation for each of the 22 test images, the following observations were made: When transforming from pneumoconiosis images to healthy images, there was a tendency for many pixel values to decrease (Fig. 6). Additionally, some cases showed no significant difference in pixel values before and after transformation. Conversely, when transforming from healthy images to pneumoconiosis images, there was an overall tendency for pixel values to increase (Fig. 7). Similar to the transformation from pneumoconiosis to healthy images, some cases exhibited minimal differences in pixel values before and after transformation.

In summary, when using CycleGAN to transform pneumoconiosis and healthy lung images, there was a trend observed towards changing the overall brightness of the lung field regions.

4 Discussion

From the experiment results, it was found that transforming pneumoconiosis images and healthy images led to changes in brightness. This is believed to be due to an overall increase in brightness across the lung field affected by pneumoconiosis. However, no specific parts of the lung, such as fibrous objects characteristic of pneumoconiosis symptoms, were generated or removed; only the brightness

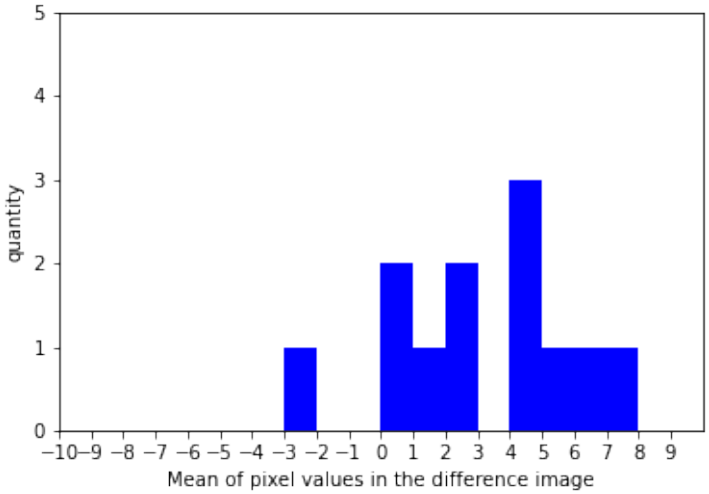


Fig. 7: Distribution of pixel value changes in images transformed from healthy to pneumoconiosis

changed. This suggests that CycleGAN may be discerning pneumoconiosis based solely on changes in brightness.

Furthermore, some images showed little change or changes opposite to expected patterns after transformation. This indicates instances where CycleGAN did not perform the transformation effectively. The reasons for these unsuccessful transformations remain unclear and will be investigated as future work.

From these results, CycleGAN's potential to identify pneumoconiosis based on changes in brightness is suggested. However, it was observed that the addition or removal of dust, a symptom of pneumoconiosis, was not achieved. Therefore, this approach may not be suitable for diseases where features contributing to symptoms, such as changes in lung roughness, are localized within the overall region of the lung.

5 Conclusion

In this study, we used CycleGAN to perform mutual transformations between the lungs of pneumoconiosis patients and healthy individuals. By investigating the changes before and after the transformations, we analyzed the criteria and trends in CycleGAN's pneumoconiosis-to-healthy and healthy-to-pneumoconiosis conversions. Additionally, we calculated the differences in pixel values before and after the transformations and analyzed the distribution of the average changes to substantiate the observed trends. As a result, when transforming from pneumoconiosis to a healthy image, a decrease in the brightness across the entire lung

field was observed. Conversely, when transforming from healthy to pneumoconiosis, an increase in brightness across the entire lung field was noted.

These results suggest that CycleGAN may be determining pneumoconiosis based on the degree of brightness. Additionally, since no specific parts of the lung were generated or removed, it is implied that CycleGAN might not be using or placing much importance on features other than brightness changes. Moreover, there were instances where the transformations by CycleGAN were not successfully executed.

Future research will aim to improve the accuracy of CycleGAN transformations and apply other explainability methods that can capture more detailed features. This will help explore whether features other than brightness are involved in the determination of pneumoconiosis. On the other hand, in XAI research, another approaches are also studied. For example, in [16], AI-generated images are used to explain the result of CNNs. Furthermore, other studies use other types of GANs to achieve explainability for diseases of the lungs and liver[17]. It is important to compare our method with these methods to further refine and advance the research.

Acknowledgement

This work was supported by JSPS KAKENHI, Grant Numbers JP22K12786, JP22K19650, JP21H03553, JP22H03699, and JP20H00267.

References

1. Yang, Fan, et al. "Pneumoconiosis computer aided diagnosis system based on X-rays and deep learning." BMC medical imaging 21 (2021): 1-7.
2. Bo Zhou Daozheng Chen Jun Gao, Qian Jiang. "Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: An overview." In *Mathematical Biosciences and Engineering*, Vol. 16, pp. 6536–6561, 2019.
3. R. Manickavasagam, S. Selvan, and Mary Selvan. "Cad system for lung nodule detection using deep learning with cnn." *Medical & Biological Engineering & Computing*, Vol. 60, No. 1, p. 221–228, 2021.
4. Dunya Alkurdi, Muhammad Ilyas, and Akhtar Jamil. "Cancer detection using deep learning techniques." *Evolutionary Intelligence*, pp. 1–9, 07 2021.
5. Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016.
6. Tsutsui Yasuyuki, Yuki Shinomiya, and Shinichi Yoshida, "Analysis of Trained Convolutional Neural Network using Generative Adversarial Network", In *International Workshop on Advanced Computational Intelligence and Intelligent Informatics(IWACIII)*, Oct.31-Nov.3, 2021
7. Zhu, Jun-Yan and Park, Taesung and Isola, Phillip and Efros, Alexei A. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks." In *The Institute of Electrical and Electronics Engineers(IEEE) International Conference on Computer Vision (ICCV)*, Oct 2017.

8. Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments." *University of Massachusetts, Amherst, Technical Report 07-49*, October, 2007.
9. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In *Advances in Neural Information Processing Systems*, Vol. 27, pp. 2672–2680. Curran Associates, Inc., 2014.
10. Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena. "Self-Attention Generative Adversarial Networks." In PMLR, 2019
11. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
12. Alrashed, Ahmed Raed Sabah, and Timur Inan. "Age And Gender Detection By Face Segmentation And Modefied CNN Algorithm." 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1-7, 2023.
13. Liew, Shan Sung, et al. "Gender classification: a convolutional neural network approach." *Turkish Journal of Electrical Engineering and Computer Sciences* 24.3 (2016): pp. 1248-1264.
14. Xiaosong Wang, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097–2106, 2017.
15. Radford, Alec. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434*, 2015.
16. Bird, Jordan J., and Ahmad Lotfi. "Cifake: Image classification and explainable identification of ai-generated synthetic images." *IEEE Access* (2024).
17. Hasenstab, Kyle A., et al. "Feature Interpretation Using Generative Adversarial Networks (FIGAN): A framework for visualizing a CNN's learned features." *IEEE Access 11* (2023): 5144-5160.