Research on Lung Diagnosis Methods based on Data Augmentation

Ruoshi Zhu

School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China

zhuruoshi@shu.edu.cn

Abstract. The objective of this research endeavor is to propose an innovative methodology for diagnosing lung cancer through a sophisticated approach to data augmentation. In essence, the proposed method harnesses the potential of the Swin-Unet model, a unique architectural design for feature extraction and classification. Further, it employs StyleGAN3-a state-of-the-art technique from the realm of Generative Adversarial Networks - to enhance and expand the dataset. In tandem with these techniques, the Copy-Paste method is deployed to amplify the diversity and volume of the dataset, effectively bolstering the network model's generalization capabilities. A comparative analysis, observing the impact of dataset enhancement and different data augmentation techniques on the Swin-Unet's classification task, is conducted to validate the study's hypothesis. The study aims to elucidate the effectiveness of using Generative Adversarial Networks for dataset expansion and their role in improving the diagnostic precision of the model used for lung cancer diagnosis. The research findings aspire to contribute valuable insights that could potentially enhance the accuracy, standardization, and efficiency of lung cancer diagnosis. This is particularly beneficial in scenarios where the available sample sizes are limited, posing challenges to effective diagnosis and treatment planning. As such, the value of the proposed method is paramount, given its potential to revolutionize current practices in lung cancer diagnostics.

Keywords: Lung cancer, CT scans, Machine learning, Artificial intelligence

1 Introduction

Lung cancer, a grave illness posing a severe threat to human life, continues to be a global health problem with rising rates of incidence and death. In 2020 alone, it is estimated that roughly 205 individuals succumbed to lung cancer on an hourly basis worldwide [1]. A regime of early screening and detection is pivotal to obtaining favorable outcomes in the treatment of lung cancer. Presently, the primary mode of screening for lung cancer is medical imaging techniques, with a special emphasis on Computerized Tomography scans. Regrettably, drawbacks of this approach include the high operating costs, extensive radiation exposure, and low specificity in cancer

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_55

detection, despite its prowess in identifying cancerous tumors at their nascent stage. Hence, it becomes imperative to devise innovative techniques and technologies for early diagnosis and treatment of lung cancer. In this regard, Machine Learning and Deep Learning are emerging as a hopeful research direction for better lung cancer detection regimes. However, the creation of varied and representative datasets is fundamental to enhancing computer-aided diagnosis, which incorporates Artificial Intelligence. Unfortunately, scarcity of medical data continues to hinder progress in this field [2].

In recent years, significant achievements have been made in the medical imaging field with the use of Generative Adversarial Networks (GANs) for expanding datasets. Nevertheless, the process of collecting and annotating medical segmentation datasets presents costly and challenging hurdles. The quantity of medical image datasets is constrained by medical resources, and only trained medical image experts can annotate the data with precision. In such scenarios, semi-supervised medical image segmentation methods that use discriminators and self-learning mechanisms can enhance the performance and generalizability of medical semantic segmentation models, even when there are few annotated pixels. However, these methods still primarily focus on generating annotations for medical images and have not effectively augmented medical image data, a key issue requiring attention [3]. Moreover, quality medical datasets are vital for propelling healthcare innovations and elevating patient care, but their availability is scarce owing to the sensitive nature of medical data and various barriers that hinder data sharing and analysis. These obstacles include privacy concerns, a lack of standardization, competition among healthcare providers, and legal impediments [4]. Sanitizing data is an essential step to safely and legally publish datasets that contain private information. Differential privacy is generally a prevalent framework used for data sanitizing which adjusts based on input information for a specific task. Consequently, this technique for sanitizing data severely restricts the type and form of datasets that can be published and exhibits poor adaptability for unforeseen new tasks. However, in comparison to conventional data privacy protection techniques, Generative Adversarial Networks have demonstrated superior performance in safeguarding privacy. These networks can generate synthetic data that resemble the input distributions of the original dataset while also ensuring data privacy. Specifically, Generative Adversarial Networks add calibrated random noise to the generated gradient information during the backward propagation stage of model training to ensure privacy protection. Although these methods have yielded reasonable results, most have not yet achieved the quality level of the original dataset [5].

In an effort to address these issues, this paper proposes a novel method for lung cancer diagnosis through data augmentation. This approach involves the use of the Swin-Unet model for feature extraction and classification, the StyleGAN3 for dataset enhancement and expansion, and the Copy-Paste method to improve the diversity and quantity of the dataset and to augment the generalization capability of the network model. By comparing the impact of dataset enhancement and different data augmentation methods on the Swin-Unet classification task, the effectiveness of expanding datasets through Generative Adversarial Networks and improving the

diagnostic accuracy of the model for lung cancer diagnosis can be studied. This research offers a valuable reference for enhancing the accuracy, standardization, and speed of lung cancer diagnosis in scenarios with few samples [6].

2 Methodology

The proposed lung diagnostic method in this study consists of three parts: the GAN image synthesis network, the data augmentation part, and the Swin-Unet part.

2.1 Generative Network Structure

This study employs a generator identical to styleGAN3 as shown in the Fig. 1. Initially, a mapping network is used to transform the initial latent code, which follows a normal distribution, into an intermediate latent code, denoted as $w \sim \mathcal{W}$. Subsequently, the synthesis network G begins generating images $Z_N = G(Z_0; w)$. To facilitate precise continuous translation and rotation of input values, Fourier features are utilized. Specifically, frequencies are uniformly sampled within a circular frequency domain with $F_c = 2$, and these frequencies remain fixed during the training process. By employing Fourier features, the generator becomes more suitable for simulating unaligned and arbitrarily oriented datasets, as any geometric transformations of intermediate features Z_i are directly propagated to the final Z_N.

Here, N refers to a sequence of layers consisting of convolutional, non-linear, and upsampling layers. In contrast to the previously successful styleGAN2, the N-layer sequence does not include a noise component. This modification ensures that the precise sub-pixel positions of each feature are inherited entirely from the underlying coarse features. Although this change does not significantly affect the model's performance, it reduces computational complexity to some extent [7].

The intermediate latent code w controls the convolutional kernel of the synthesis network G. During training, the exponential moving average (EMA) of the mean squared value $\sigma^2 = \mathbb{E}[x^2]$ is computed for all pixels and feature maps. The feature maps are then normalized by dividing them by $\sqrt{\sigma^2}$ before convolutional operations are applied. After incorporating the bias term b_i, each layer at every resolution undergoes a $2\times$ upsampling, reducing the number of feature maps by half. Subsequently, the data passes through the Leaky ReLU activation function and undergoes a $2\times$ downsampling.

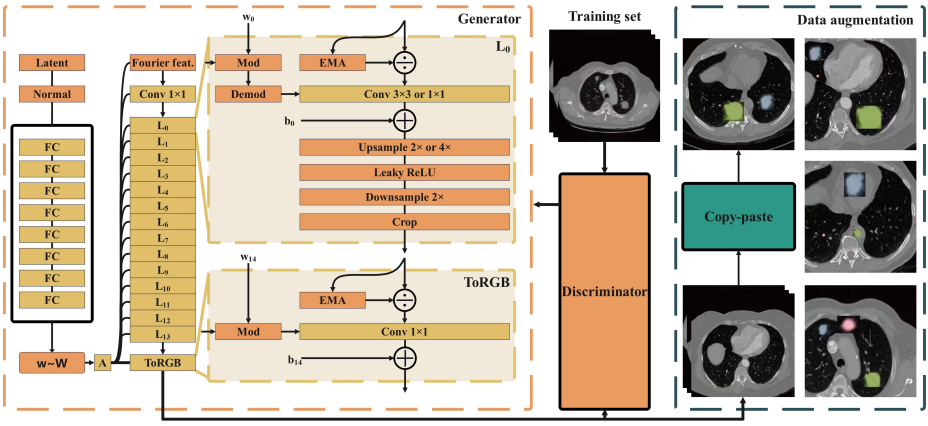

Fig. 1. System Block Diagram (Photo/Picture credit: Original)

2.2 Data augmentation

In the data augmentation part, a simple Copy-Paste method is employed, which has been shown to be effective in previous studies, particularly in semi-supervised models. This method randomly selects two images and applies random perturbations and flips to both images. Then, a subset of the segmentation from one image is pasted onto the other image. The scaling factor of the images ranges from 0.8 to 1.25 to ensure the coherence of the output image distribution [8].

During the process of image pasting, a binary mask α for the pasted object can be generated based on the annotations of the real image. To ensure the smoothness of the mask, a Gaussian filter is applied. Subsequently, the new image is computed as $I_1 \times \alpha + I_2 \times (1 - \alpha)$, where I_1 represents the pasted object and I_2 represents the main image.

2.3 Segmentation Network Structure

The design of the segmentation network in the study is based on Swin-Unet, which comprises of encoder, bottleneck, decoder, and hop connections. The Swin Transformer block serves as the fundamental building block of the network. By segmenting the input image into non-overlapping patches of size 4×4 , the input sequence is transformed into an embedding sequence. By applying this approach, the feature dimension of each patch is transformed to 48 [9].

The decoder section of the network is composed of Swin Transformer blocks and patch expansion layers. To address the information loss resulting from downsampling, the contextual features are combined with multi-scale features from the encoder using skip connections [10]. The patch expansion layers reshape the adjacent dimension feature maps into a larger feature map and perform a $2 \times$ upsampling of the resolution. Following this, another patch expansion layer is used to perform a $4 \times$ upsampling of the image, restoring the resolution of the feature maps to match the input resolution

[11]. Finally, a linear projection layer is utilized to generate the segmentation predictions based on these outcomes.

3 Experiment and analysis

3.1 Dataset

This study utilizes the LUNA16 dataset, which is publicly available and derived from the LIDC/IDRI database under the Creative Commons Attribution 3.0 Unported license. The LUNA16 dataset itself is also licensed under the Creative Commons Attribution 4.0 International license. Scans with a slice thickness greater than 2.5 millimeters were excluded, resulting in a total of 888 CT scans included in the dataset. The LIDC/IDRI database also contains annotations collected through a two-stage annotation process performed by four experienced radiologists. Each radiologist marked non-nodule lesions, nodules with a size less than 3 millimeters, and nodules with a size greater than or equal to 3 millimeters. The reference standard selected for this study includes all nodules with a size greater than or equal to 3 millimeters that were confirmed by at least three radiologists [12].

The focus of this study is to investigate data augmentation methods in the context of limited data. Thus, despite the availability of 1,186 nodules with valid annotations in the LUNA16 dataset, this study selected a subset of 300 images with more pronounced features to simulate the scenario of data scarcity. Out of these, 270 images were allocated for training purposes, while the remaining 30 images were used for testing. By adopting this approach, the aim is to examine the effectiveness of data augmentation techniques for nodule detection in situations where data is insufficient.

3.2 Metrics

Hausdorff Distance (HD).

HD proposed by the German mathematician Felix Hausdorff in 1914, is a distance metric used to measure the similarity between two non-empty finite sets. It is defined as the maximum of the minimum distances from each point in one set to the other set and vice versa. Mathematically, the Hausdorff Distance between two sets A and B is denoted as H(A, B) and calculated as follows:

$$H(A,B) = \max\big(h(A,B), h(B,A)\big) \tag{1}$$

where h(A, B) represents the minimum distance from each point in set A to the nearest point in set B, and h(B, A) represents the minimum distance from each point in set B to the nearest point in set A.

The computation of Hausdorff Distance involves finding the distances between all points in the two sets and determining the maximum distance. This enables Hausdorff Distance to capture the largest difference between corresponding points in the two sets, making it a valuable metric for quantifying shape dissimilarity.

Dice similarity coefficient (DSC).
DSC also known as Dice coefficient or Dice index, is a widely used evaluation metric for image segmentation. It measures the similarity between predicted segmentation results and ground truth segmentation results. The DSC ranges from 0 to 1, with a value closer to 1 indicating a higher degree of overlap between the predicted and ground truth segmentations.

The calculation of DSC is based on the intersection and union of the predicted and ground truth segmentations. Specifically, it is computed as twice the intersection area divided by the sum of the areas of the predicted and ground truth segmentations. The mathematical expression is as follows:

$$DSC = \frac{2 \times S_{Intersection}}{S_{Predicted} + S_{GT}} \tag{2}$$

By computing the DSC, the accuracy and performance of segmentation algorithms in image segmentation tasks can be evaluated. A higher DSC indicates that the algorithm captures the target structures in the image more accurately and is more consistent with the ground truth segmentation.

In this study, certain modifications have been made to the Dice similarity coefficient for the semantic segmentation task of lung tumors, with the aim of encouraging the model to predict the lesion region more accurately. Specifically, a hyperparameter α that is multiplied with the intersection area and the total area is introduced. Through experimental analysis, it was found that a value of 1.2 is more suitable for α. Therefore, the modified mathematical expression is as follows:

$$dice\_loss^{*} = \frac{2\alpha S_{Intersection}}{S_{Predicted} + \alpha S_{GT}} \tag{3}$$

3.3 Ablation experiments

During the model training process, there were significant challenges in fitting and oscillation of the loss curve. This was attributed to the fact that, unlike other semantic segmentation tasks, the majority of lung tumor areas are much smaller than the image area. To address this issue, a sensitivity analysis of the loss function was conducted in this study. Specifically, the LUNA16@300 dataset was utilized for training and prediction, with 800 epochs performed using different combinations of loss functions, including 0.6Loss_ce+0.4Loss_dice, 0.6Lloss_ce+0.4Loss_dice*, Loss_dice, and Loss_dice*. The results obtained from these experiments are presented in Table 1.

Table 1. Result of the sensitivity experiment

Loss	DSC	HD	Precision
Loss_dice	0.062	71.07	0.3
0.4Loss_dice*+0.6Loss_ce	0	43.53	0.13
0.4Loss_dice+0.6Loss_ce	0	15.34	0.07
Loss_dice*	0	0	0

The sensitivity analysis revealed that for semantic segmentation of the extremely small lesion areas in thoracic tumors, the use of Loss_dice proved to be the most effective. When combined with the cross-entropy loss function, the proposed Loss_dice* demonstrated a certain improvement in accuracy compared to the original method. Therefore, in subsequent experiments, the Dice loss function is adopted as the primary loss function.

In this section, experiments were conducted using the Swin-Unet network to investigate the effectiveness of generating images. From the 1000 images generated by the generator, 279 images with distinct features were selected and annotated. These images were then combined with the LUNA16@300 dataset, resulting in an augmented dataset named LUNA16+@579. Semantic segmentation experiments were performed on the LUNA16@300, LUNA16+@379, LUNA16+@479, and LUNA16+@579 datasets, and the DSC (Dice Similarity Coefficient) and HD (Hausdorff Distance) metrics were compared across different datasets.

4 Result

In this study, the Swin-Unet network was trained for 800 epochs using the LUAN16@300 dataset and various augmented datasets of different sizes. The obtained results are presented in Table 2. It can be observed from the results that the proposed data augmentation method significantly improves the segmentation performance of the semantic segmentation model in scenarios with limited data availability. As shown in Fig. 2.

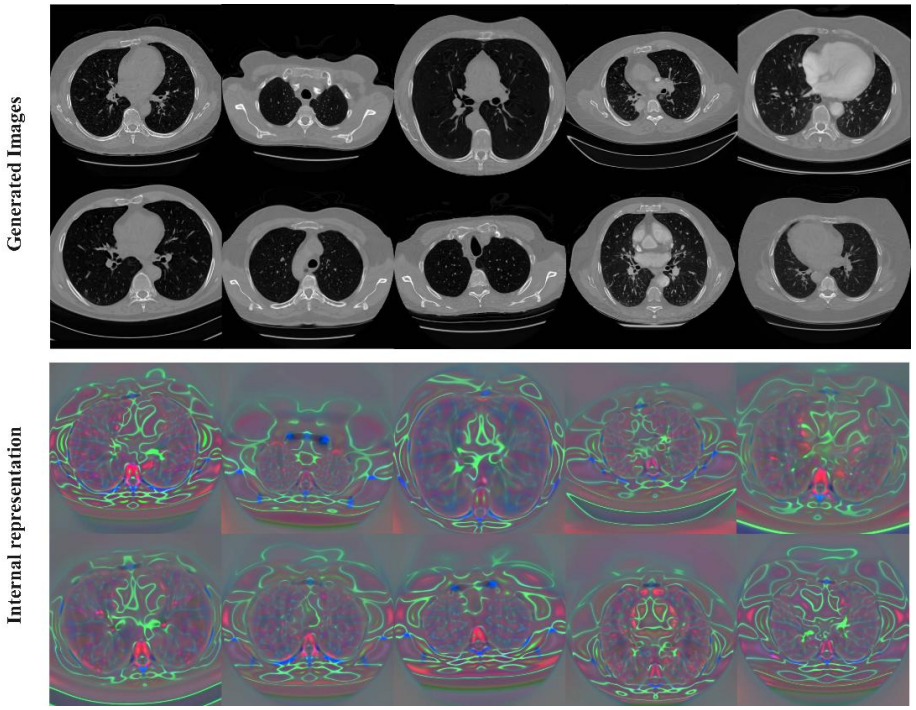

Fig. 2. Synthetic Image Generation and Internal Representations via styleGAN3se (Photo/Picture credit: Original)

In this study, the Swin-Unet network was trained for a total of 800 epochs, and experiments were conducted using the LUAN16@300 dataset as well as multiple augmented datasets of varying sizes as the training set. The experimental results are presented in Table 2. As the proportion of augmented data increased, there was an increasing trend in the Dice similarity coefficient of the predicted results, indicating improved segmentation performance. Furthermore, there was a significant improvement in prediction accuracy. Therefore, it can be concluded that the proposed data augmentation method significantly enhances the segmentation performance of the semantic segmentation model for lung tumors, particularly in scenarios with limited data availability.

Table 2. Result of the experiment

Dataset	DSC	HD	Precision
LUNA16@300	0.06	71.07	0.31
LUNA16+@379(ours)	0.19	73.24	0.55
LUNA16+@479(ours)	0.19	62.74	0.48
LUNA16+@579(ours)	0.38	69.5	0.61

The primary focus of this study is to enhance the generator of GAN networks, with minimal modifications to the discriminator. However, it appears that modifying the

discriminator has the potential to improve the quality of the synthesized images. For instance, in the training results, some images do not exhibit the desired nodular features, which could be attributed to issues with the discriminator.

In addition, the LUNA16 dataset used in this study consists of black and white images. However, the preprocessing technique employed in this paper, namely point sampling, results in all pixels in the training images being black and white, leading to the appearance of jagged features at the image edges. This severe aliasing in the training data is detrimental to conventional GANs. On one hand, the generator needs to translate the output smoothly at a sub-pixel level, but on the other hand, the edges must retain the jagged appearance to preserve the characteristics of the training data. This issue poses a significant and challenging problem in the field of medical image generation.

The medical images generated by the generator do not include annotation information. Therefore, the challenge of obtaining high-quality annotated information in medical datasets still needs to be addressed. In the future, it may be possible to modify the input of the generator by incorporating the annotation information from the training set along with the image information into the GAN generator to achieve the synthesis of annotated information and synthesized images. However, this remains an open and highly challenging problem.

5 Conclusion

This investigation has successfully introduced a novel approach to lung cancer diagnosis by means of data augmentation. Utilization of the Swin-Unet model allows for efficient feature extraction and classification, while application of StyleGAN3 aids in enhancing and expanding the dataset. The Copy-Paste technique further contributes to increasing dataset diversity and quantity, thereby significantly improving the model's accuracy in diagnosing lung cancer. The experimental outcomes verify the efficacy of employing generative adversarial networks for dataset expansion, demonstrating a notable enhancement in the performance and generalization capacity of the network model. The contributions of this research are instrumental in advancing lung cancer diagnosis, particularly in scenarios with limited sample availability. The proposed method addresses several drawbacks of existing screening procedures, such as high costs, exposure to radiation, and low specificity, offering a viable alternative for the early detection and treatment of lung cancer. Furthermore, successful implementation of data augmentation techniques using GANs presents a solution to the paucity of medical image datasets and the challenges inherent in collecting and annotating medical segmentation datasets. Enhancements in diagnostic accuracy, standardization, and efficiency achieved through the introduced method bear significant relevance for innovations in healthcare and patient care. Accurate early-stage lung cancer diagnosis can facilitate improved treatment outcomes and, in the long run, decrease the morbidity and mortality rates associated with this formidable disease.

Future avenues for research may include further refinement and optimization of the proposed method, alongside its validation on larger and more diverse datasets. Exploration of potential integration of other cutting-edge technologies, such as deep learning and artificial intelligence, might further augment the overall performance and applicability of the lung cancer diagnosis method. In conclusion, this research furnishes valuable insights and a robust groundwork for the evolution of superior lung cancer diagnosis techniques, contributing significantly to the global initiatives aimed at tackling this formidable health challenge.

References

1. Rajaraman S, Antani S. Weakly labeled data augmentation for deep learning: a study on COVID-19 detection in chest X-rays[J]. Diagnostics, 2020, 10(6): 358
2. Barshooi A H, Amirkhani A. A novel data augmentation based on Gabor filter and convolutional deep learning for improving the classification of COVID-19 chest X-Ray images[J]. Biomedical Signal Processing and Control, 2022, 72: 103326.
3. Chlap P, Min H, Vandenberg N, et al. A review of medical image data augmentation techniques for deep learning applications[J]. Journal of Medical Imaging and Radiation Oncology, 2021, 65(5): 545-563.
4. Tang Y B, Tang Y X, Xiao J, et al. Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation[C]//International Conference on Medical Imaging with Deep Learning. PMLR, 2019: 457-467.
5. Müller-Franzes G, Niehues J M, Khader F, et al. Diffusion probabilistic models beat gans on medical images[J]. arXiv preprint arXiv:2212.07501, 2022.
6. Skandarani Y, Jodoin P M, Lalande A. Gans for medical image synthesis: An empirical study[J]. Journal of Imaging, 2023, 9(3): 69.
7. Man K, Chahl J. A Review of Synthetic Image Data and Its Use in Computer Vision[J]. Journal of Imaging, 2022, 8(11): 310.
8. Atad M, Dmytrenko V, Li Y, et al. Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan[J]. arXiv preprint arXiv:2207.07553, 2022.
9. Gu Y, Chi J, Liu J, et al. A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning[J]. Computers in biology and medicine, 2021, 137: 104806.
10. Toda R, Teramoto A, Kondo M, et al. Lung cancer CT image generation from a free-form sketch using style-based pix2pix for data augmentation[J]. Scientific reports, 2022, 12(1): 12867.
11. Zotov E, Tiwari A, Kadirkamanathan V. Conditional StyleGAN modelling and analysis for a machining digital twin[J]. Integrated Computer-Aided Engineering, 2021, 28(4): 399-415.
12. Bhatt N, Prados D R, Hodzic N, et al. Unsupervised detection of lung nodules in chest radiography using generative adversarial networks[C]//2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021: 3842-3845.