Artificial intelligence applications in oncology to augment data and support decisions

Artificial Intelligence Applications in Oncology to Augment Data and Support Decisions

Zhixiang Wang

Artificial Intelligence Applications in Oncology to Augment Data and Support Decisions

Dissertation
to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus,
Prof. dr. Pamela Habibović,
in accordance with the decision of the Board of Deans,
to be defended in public

on Tuesday 11th July 2023, at 10.00 AM

by

Zhixiang Wang

Supervisor:
Prof.dr.ir. A.L.A.J. Dekker

Co-supervisor:
Dr. A. Traverso
Dr. L.Y.L. Wee

Assessment Committee:
Prof.dr. Frank Verhaegen (Chair)
Dr. Cecile Wolfs
Prof.dr. Marius Staring (Leiden University)
Dr. Zhenwei Shi (Guangdong Provincial People's Hospital)

Table of contents

Chapter 1: Introduction and Outline of Thesis

Rationale

Artificial Intelligence (AI) is being used in oncology to analyze and compare huge amounts of routinely collected data to find patterns and correlations with an outcome of interest (e.g., survival, treatment response). AI can potentially help clinicians to diagnose, treat, and track the progress of disease of cancer patients more quickly and accurately.

Definitions

AI is a new technological science that investigates and develops theoretical methods, technologies, and application systems for simulating, extending, and expanding human intelligence. The relationship between machine learning (ML), deep learning (DL), and AI is shown in Figure 1.1.

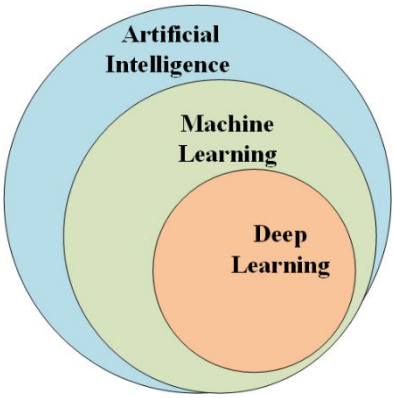


Figure 1.1 The relationship between machine learning (ML), deep learning (DL), and Artificial intelligence.

Machine Learning

Machine Learning (ML) is the most important branch of AI. It spans several disciplines including computer science, engineering, and statistics. ML can excavate large amounts of high dimensional historical data and use them for prediction or classification. More specifically, ML attempts to build a function between the input (data) and the output (outcome) [1]. ML can be divided into supervised and unsupervised learning

Supervised learning includes classification and regression models. The classification model and regression model can predict specific labels (discrete

value) and values within a certain range (continuous value) from input features[2]. It is widespread in ranking, recommendation systems, recognition tasks, etc.

Unsupervised learning does not require one to have "labels" in the dataset. It applies statistical methods to find common features within the dataset. The core applications of unsupervised learning are density estimation and cluster analysis[3]. In the medical field, especially in clinical research, manually labeling retrospective cohorts of patients requires massive human and material resources, which usually makes it impossible to obtain large-scale real data. Unsupervised learning algorithms can be used to automatically label data by performing clustering, which can be helpful in cases where labeled data is not available. Additionally, unsupervised learning algorithms can be used to identify and group similar types of data, making it easier to identify patterns or outliers.

Because there is no need of labelled data, unsupervised learning algorithms can learn from larger scale data, making it easier for the algorithm to process and better understand the task for more accurate performance. By combining unsupervised learning with supervised learning methods, machine learning models that make more accurate predictions and better decisions can be created.

Deep Learning

Deep learning (DL), a branch of ML, refers to algorithms that simulate human neurons by building large neural networks, also known as deep neural networks, that are trained using computers with high-performance computing and vast amounts of data[4].

DL has achieved impressive results in the analysis of medical images, beyond visual inspection, as demonstrated in 2012 by Krizhevsky, who presented a deep convolutional neural network called AlexNet, which won the first place in a classification task containing 1.2 million high-resolution images in 1,000 categories, It demonstrated the ability of deep learning to automatically process and classify images[5].

As more and more algorithms were proposed, it was found that deep learning for medical image analysis achieved remarkable results in tasks such as organ segmentation[6], disease diagnosis[7], tumor detection[8], etc. DL is trained so

that the convolutional layers automatically extract the features needed for the task from the image and make predictions.

Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs), a subset of DL based on game theory, became popular in the medical imaging domain, mainly for synthetic data generation[9]. Briefly, GANs consist of two competing actors: a generator and a discriminator. They are used to generate synthetic images/samples and "judge" the quality of the generated images, respectively. The equilibrium is reached when the synthetic (i.e., "fake") samples cannot be distinguished from the real distribution [10].

Since GANs were proposed in 2014 by Ian Goodfellow[11], it has been widely used by an increasing number of applications in the standard of care medical imaging. For example, GAN models can be used to simulate certain patient conditions to better evaluate different treatment options. In addition, GAN models can optimize Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) scan data to obtain a clearer 3D view that can be used to pinpoint treatment targets and precisely place treatment sources. Thus, in the medical field, the application of GAN can improve image analysis, accelerate diagnostic results, and can improve treatment simulation and target localization.

AI applications on decisions

Traditional computer vision involves defining handcrafted features that are extracted from medical imaging data, for example, to predict prognosis or segment tumors. This technology is particularly advantageous in oncology because it can automatically analyze a large amount of data. In recent years, the Convolutional Neural Network (CNN) was widely used as an advanced medical image feature extraction method to find the best features automatically.

AI can improve the accuracy and efficiency of diagnosis by providing preliminary diagnosis information and highlighting important features in the diagnosis process. AI, with its ability to summarize and analyze large amounts of data to solve multifaceted problems, has become an important tool for clinicians to identify similarities between heterogeneous cohorts of patients. Furthermore, with the training of large amount of data, AI shows strong

generalizability and precision, leading to the belief that it can be utilized to aid physicians in making data-driven clinical decisions. Additionally, this technology can offer a disease probability predicted by AI as supplemental information to assist physicians in making the correct diagnoses[12].

AI applications on data augmentation

Oncology data is one of the most important resources for AI research in the medical field[13]. Oncology data consist of a large amount of information, from patient records to medical imaging and genomic and biological markers, which can utilize AI to identify meaningful connections between features and specific diseases and outcomes. In this way, oncology data is a critical resource for the further development of AI in healthcare.

Artificial intelligence can automatically or semi-automatically segment tumor regions of interest (ROI) from medical images such as CT and MRI [14]. This can improve efficiency, accuracy, reproducibility, and consistency, and is more attractive and has clinical potential than the current. fully manual method. For solid tumors with clear boundaries, the results of semi-automatic and automatic delineation are very similar to those of manual delineations [15].

The generation of radiotherapy plans requires the search for optimal treatment parameters, such as the number of beams, clinical constraints, etc. Although several traditional automated or semi-automated planning methods have been proposed, they are time-consuming and highly dependent on the experience and expertise of the radiation oncology staff [16]. Automated planning methods based on deep learning have the potential to significantly reduce the care professionals' work time, increase efficiency, and reduce inter-observer variability.

Medical data generation is a method of data simulation using AI technologies such as neural networks, machine learning, and deep learning to generate high-quality and precise virtual data which can provide virtual and simulated data for medical research to address the current shortage of experimental data.
Medical data generation technology can improve the efficiency of medical laboratories and save time and cost.

It enables researchers to quickly obtain large-scale realistic samples in a shorter time. Such samples can help researchers to train AI and junior clinicians. AI-generated samples can simulate realistic images containing lesions to help researchers explore lesion structure and function so that junior clinicians and researchers can better understand and more accurately identify lesions.
In addition, this technology is also helpful to improve the safety and accuracy of clinical trials. Using AI to simulate clinical data, simplifies the complex experimental process and helps to achieve higher levels of performance in terms of security and reliability by predict the progression of disease, take preventive measures in advance, and decide whether intervention is needed.

Objectives

This thesis proposes two hypotheses: first, an artificial intelligence (AI) model can generate high-quality synthetic medical data from multiple sources, thereby improving the performance of classification models when combined with real data. Second, AI models can attain high accuracy in diagnosing patients from multiple types of medical data, and they can significantly enhance the efficacy of clinicians during the diagnostic process. To reach the above objectives, this thesis has first focused on analyzing the limitations of the state-of-the art solutions

Limitations

The main reason why it is difficult to obtain high-quality curated large datasets may be due to the following reasons. First, the hardware for collection and storage requires dedicated costs, which not all hospitals and research facilities can afford. Secondly, labeling and cleaning routinely acquired unstructured big data requires working time from medical professionals. Finally, the confidentiality of medical data is also a very important responsibility [17] because of the following reasons. Firstly, medical data involves confidential information such as patients' medical records, examination results, test results and other treatments, and if this information is compromised, it could cause serious damage to patients' privacy. At the same time, a breach of medical data has the potential to trigger legal disputes.
The difficulty of obtaining quality data will have a negative impact on the progress of AI and medical technology. First, a low quality of data may reduce the accuracy of data analysis, which may mislead AI decision-making; Second,

insufficient data may cause the model to overfit, which results in the model not being able to generalize to real-world applications. Finally, the lack of accurate data will limit AI's development in the medical field.

Data imbalance refers to the unequal number of instances of different categories, such as in a binary classification problem, when there is many positive samples compared to the negative class. This is a common problem in ML, which affects the results of AI applications [18]. This will bring issues during model training. First, overfitting: a large difference in the number of samples from different categories in the dataset prevents the model from being adequately trained, which leads to over-fitting the training data of the most frequent class with best performances, while reducing them significantly on the other categories[19]. Second, the generalization ability is weakened. Due to the large deviation of the training set, when the model is applied to external data, the performance of the model may decline significantly. Third, accuracy distortion: Data imbalance will reduce the classification accuracy of the model because the model will tend to predict the categories with more samples, which may lead to misclassification.

The impact of data imbalance on AI is significant, which affects not only the results of AI but also their clinical value. Therefore, effective measures must be taken to overcome this problem.

Because DL requires a large amount of data for training, it often needs to be centralized and unified for training: a concept known as centralized learning. However, this method will consume a lot of hardware storage and transmission time. Moreover, especially for medical data,

legal and privacy issues are also important obstacles to centralized learning [20]. To solve the problems of data island (interpretation) and privacy protection, in 2016, H. Brendan McMahan [21] proposed a federated learning algorithm and achieved comparable results to the centralized approach. Differently from centralized methods, federated learning does not move data, but trains the model by moving the model to all the data centers ('nodes'). That is, instead of transmitting data to the central server, the method of transmitting the model realizes that the data is available and invisible to the server, to achieve the effect of privacy protection.

AI models rely heavily on data to create algorithms that can help machines learn how to recognize patterns in data. However, there are still some defects in medical data. First, as privacy protection is a consequence of national regulations, it may be difficult to collect complete human medical data, especially genetic data. Second, the human body is an integrated system and

the data from the single modality used, such as CT or a specific disease biomarker, can be incomplete information, which can lead to errors in the analysis of the patient. In this way, when the data is incomplete, the AI model may not be able to accurately identify the pattern or establish a biased model, resulting in unreliable results [22].

Robustness refers to the ability of a system or program to maintain accurate predictions in data that has not been seen before. Due to the lack of reliability and the complexity of medical data, AI is prone to unexpected results when dealing with rare diseases [23]. Therefore, robustness is needed for AI models in healthcare to operate accurately and reliably when faced with different data sets, including rare diseases or unusual cases. It is generally achieved by building models that can deal with both data that has been seen and data that has not been seen before. This is achieved through developing models that can generalize better and generalization techniques such as regularization, cross-validation, and data augmentation. It is also important to keep track of data drift and use transfer learning to ensure the right set of weights and architectures to maintain an acceptable level of accuracy over time.

In the process of clinical diagnosis in the real world, clinicians need strong evidence to support the clinical recommendations and explain them in the consultation with the patient. Even though AI can approach or even exceed the accuracy of clinicians in some medical imaging tasks such as the detection of pulmonary nodules[26], tumor segmentation[27], etc, the AI prediction process is a "black box" for people because the decision-making principle underlying AI algorithms is often unclear. The AI diagnosis results given in this way are not necessarily convincing for clinicians. Therefore, it is still unclear whether AI can help clinicians in diagnosis and treatment, and further experiments are needed to prove it.

The contribution of the thesis related to solving the above issues

The outline of thesis chapters displayed in Figure 1.2

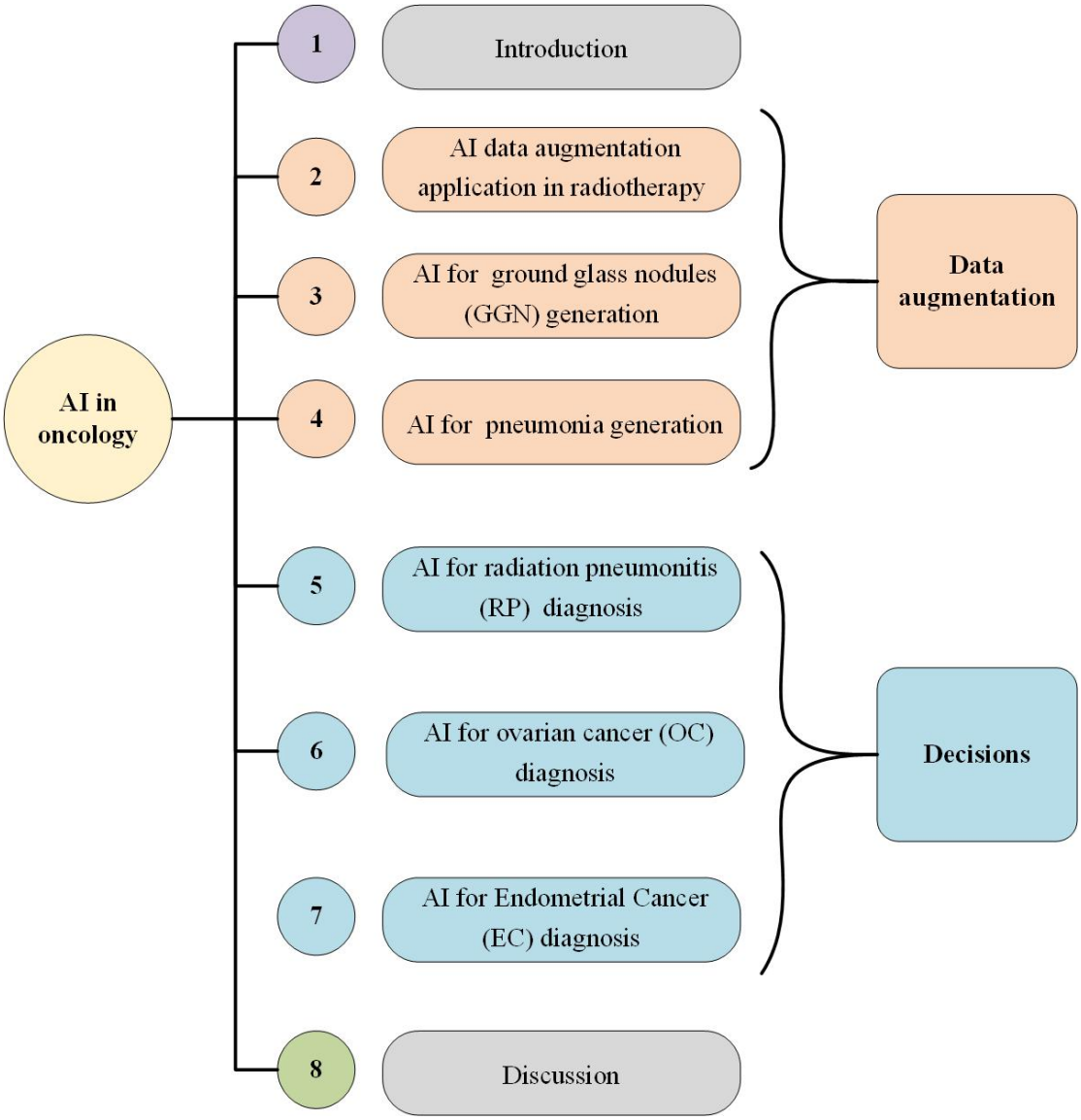


Figure 1.2 the outline of the thesis chapters.

For data augmentation

AI can generate synthetic data by simulating the real-world data distribution. It can help alleviate data privacy problems, fill the gaps in traditional data sets, test AI systems, and make AI more accessible. However, it has been difficult to verify whether the generated data is true enough by appropriate methods.

In Chapter 3, I designed an experiment to compare the gap between the generated data and the real sample through the Turing test and numerical methods and proved that the generated data is very close to the real data and difficult to distinguish.

Data generated by AI can help solve the problem of uneven data distribution.

One of the biggest problems AI faces is insufficient data or unbalanced data, which will cause AI to tend to multi-category, ignore a few category samples, and significantly reduce the accuracy and availability of the model when making decisions.

In Chapter 4, I designed an AI algorithm for data generation and demonstrated that synthetic data can overcome the shortcomings of data imbalance and can thus improve the accuracy of classification models.

For decisions

Designing an AI-based predictive system for medical diagnosis is a complex process, as it typically requires gathering different types of data to face different diseases. However, AI has enough potential to extract the necessary information from a variety of highly disparate data sources to perform a specific task, whether macroscopic medical images or microscopic biological markers.

In Chapters 5 and 6, I proved that despite different types of data from macroscopic to microscopic , AI still can predict specific disease properties.

AI can be used to automate certain diagnostic tasks and to help clinicians make more informed decisions. For example, AI can be used to identify patterns in data, such as imaging data or omics data, to make predictions about the diagnosis, disease progression, and therapeutic potential of certain diseases.

In Chapter 7, I compared the clinician's performance such as efficiency and accuracy and discovered that the clinicians with AI support have a significant improvement in diagnosis accuracy, which means AI can support for diagnosis.

References

1. El Naqa I, Murphy MJ (2015) What is machine learning? In: machine learning in radiation oncology. Springer, pp 3-11
2. Cunningham P, Cord M, Delany SJ (2008) Supervised learning. In: Machine learning techniques for multimedia. Springer, pp 21-49
3. Barlow HBJNc (1989) Unsupervised learning. 1 (3):295-311
4. LeCun Y, Bengio Y, Hinton GJn (2015) Deep learning. 521 (7553):436-444
5. Krizhevsky A, Sutskever I, Hinton GEJCotA (2017) Imagenet classification with deep convolutional neural networks. 60 (6):84-90
6. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang XJPM (2021) A review of deep learning based methods for medical image multi-organ segmentation. 85:107-122
7. Bakator M, Radosav DJMT, Interaction (2018) Deep learning and medical diagnosis: A review of literature. 2 (3):47
8. Saba T, Mohamed AS, El-Affendi M, Amin J, Sharif MJCSR (2020) Brain tumor detection using fusion of hand crafted and deep learning features. 59:221-230
9. Zhu J-Y, Park T, Isola P, Efros AA Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, 2017. pp 2223-2232
10. Jiang Y, Chen H, Loew M, Ko HJIJoB, Informatics H (2020) COVID-19 CT image synthesis with a conditional generative adversarial network. 25 (2):441-452
11. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio YJCotA (2020) Generative adversarial networks. 63 (11):139-144
12. Haleem A, Javaid M, Khan IHJCMR, Practice (2019) Current status and applications of artificial intelligence (AI) in medical field: An overview. 9 (6):231-237
13. Kann BH, Hosny A, Aerts H (2021) Artificial intelligence for clinical oncology. Cancer cell 39 (7):916-927. doi:10.1016/j.ccell.2021.04.002
14. Hesamian MH, Jia W, He X, Kennedy PJJodi (2019) Deep learning techniques for medical image segmentation: achievements and challenges. 32 (4):582-596
15. Wang J, Lu J, Qin G, Shen L, Sun Y, Ying H, Zhang Z, Hu WJMp (2018) A deep learning-based autosegmentation of rectal tumors in MR images. 45 (6):2560-2564

16. Fan J, Wang J, Chen Z, Hu C, Zhang Z, Hu WJMp (2019) Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. 46 (1):370-381
17. Ngiam KY, Khor IW (2019) Big data and machine learning algorithms for health-care delivery. Lancet Oncol 20 (5):e262-e273. doi:10.1016/s1470-2045(19)30149-4
18. Huynh T, Nibali A, He Z (2022) Semi-supervised learning for medical image classification using imbalanced training data. Comput Methods Programs Biomed 216:106628. doi:10.1016/j.cmpb.2022.106628
19. Wang C, Wu J, Xu L, Zou Q (2020) NonClasGP-Pred: robust and efficient prediction of non-classically secreted proteins by integrating subset-specific optimal models of imbalanced data. Microb Genom 6 (12). doi:10.1099/mgen.0.000483
20. Yang Q, Liu Y, Cheng Y, Kang Y, Chen T, Yu HJSLoAI, Learning M (2019) Federated learning. 13 (3):1-207
21. McMahan HB, Moore E, Ramage D, y Arcas BAJapa (2016) Federated learning of deep networks using model averaging. 2
22. Yu KH, Beam AL, Kohane IS (2018) Artificial intelligence in healthcare. Nat Biomed Eng 2 (10):719-731. doi:10.1038/s41551-018-0305-z
23. Yang ZY, Ye ZF, Xiao YJ, Hsieh CY, Zhang SY (2022) SPLDExtraTrees: robust machine learning approach for predicting kinase inhibitor resistance. Brief Bioinform 23 (3). doi:10.1093/bib/bbac050

Chapter 2: Applications of generative adversarial networks (GANs) in radiotherapy: narrative review

Adapted from ***Zhixiang Wang****, Glauco Lorenzut, Zhen Zhang , Andre Dekker ,Alberto Traverso *Applications of generative adversarial networks (GANs) in radiotherapy: narrative review. 2022. Precision Cancer Medicine.* *https://dx.doi.org/10.21037/pcm-22-28*

Abstract

Objective: This review aims to provide an up-to-date snapshot of Generative Adversarial Network (GAN) applications in radiotherapy.

Background: Radiation Therapy (RT) is the dominant method for clinical cancer treatment, which aims to ensure that Planning Target Volume (PTV) receives a sufficient dose while Organs-At-Risk (OARs) are exposed to little or no radiation. However, obtaining a clinically acceptable radiotherapy plan often requires a long time, tedious work, and a high level of physician experience. The general steps to perform RT include planning (CT/MRI/PET) image acquisition, contouring the treatment area (Gross Tumor Volume, OARs, etc.), and developing a treatment plan and treatment implementation. But there are still some challenges that need to be overcome. Fortunately, with the development of the computer science, Generative Adversarial Network (GAN) which is composed of a Generator and Discriminator with opposing optimized goals has been widely used by an increasing number of applications in various fields, especially in CT, MRI, and other images and plays a great role in RT.

Methods: We searched for studies published from January 2018 to March 2022, with English language restrictions on PubMed and IEEE Xplore databases.

Conclusion: GAN model has already been widely used in RT. Thanks to their ability to automatically learn the anatomical features from different modalities images, improve quality images, generate synthetic images and make less time consumption automatic dose and plan calculation. Even though the GAN model cannot replace the radiotherapy doctors' work, it still has great potential to enhance the radiologists' workflow. There are lots of opportunities to improve the diagnostic ability and decrease potential risks during radiotherapy and time cost for plan calculation.

Keywords: Generative Adversarial Network, radiotherapy, applications

Introduction

Radiation Therapy (RT) is the most used method for cancer treatment, which aims to deliver the prescribed dose to the Planning Target Volumes (PTVs), while simultaneously reducing at minimum the dose to Organs-At-Risk (OARs) (1). Obtaining a clinically acceptable RT treatment plan often requires a long time, tedious work, and a high level of physician/technician experience(2).
The general steps to perform RT planning include (CT/MRI/PET) image acquisition, contouring of treatment area (Gross Tumor Volume, OARs, etc.), treatment plan optimization, and treatment delivery.
However, there are still some challenges: first, sometimes a diagnosis is performed on MR scans because of better soft-tissue contrast, but a CT is always required to make a plan. Second, contouring and treatment planning is time-consuming and dependent on expertise which is prone to inter-observer variability.. Finally, in-room imaging is of a lower quality than diagnostic computed tomography (CT).
In the last 10 years, the RT research community has focused on optimizing and automizing the above-presented steps by using Artificial Intelligence (AI). With the development of computer science, deep learning (DL) algorithms, a branch of AI, are widely applied by researchers to solve the above-mentioned issues. Generative Adversarial Networks (GANs), a subset of DL, became popular in the medical imaging domain, mainly for synthetic data generation(2). Since GAN was proposed in 2014 by Ian Goodfellow (3), it has been widely used by an increasing number of applications in standard of care medical imaging, especially in CT, MRI, and plays a great role in RT (4). A deep understanding of GANs requires specific knowledge of computer science, often not available at RT clinics.
Therefore, in this review, we will introduce the development of the GAN models, their structures, their improvements, and their applications in RT which can help the researchers have a preliminary knowledge about these deep learning models.
For readers interested in a specific application of GANs, we have grouped the GAN applications into three clusters: CT translation and synthesis (see later GANs for synthetic imaging), dose and plan calculation (see later GANs for dose and plan calculation), and image quality improvement(see later GANs for quality improvement). Finally, we have discussed the limitations and future directions to give some hints for the following researchers who want to develop GAN applications in RT.

We present the following article using a Narrative Review reporting checklist

Methods

The search was performed from PubMed and IEEE Xplore datasets according to multiple keyword combinations and the related MESH terms including: “Radiotherapy”, “generative adversarial network(GAN)”, and “application”. January 1, 2018, was set as the cut-off date because we only considered the research within 5 years. The inclusion criteria were: original research articles (proceedings included), English language, and the development of a GAN model using a RT dataset. 100 publications were extracted according to the search string above. Two researchers with expertise in deep learning, quickly scanned the abstracts to exclude irrelevant articles. Subsequently, we scanned the reference list from the selected articles to include the related ones which were not found by the initial search. Finally, 23 articles refer to the applications of GANs in radiotherapy, most of which were appropriately referenced in this review (Table 1).

Table 1 The search strategy summary

Items	**Specification**
Date of Search (specified to date, month and year)	2022-04-14
Databases and other sources searched	PubMed & IEEE Xplore
Search terms used (including MeSH and free text search terms and filters)	“Radiotherapy”, “generative adversarial network(GAN)”, and “application”
Timeframe	January 1, 2018
Inclusion and exclusion criteria (study type, language restrictions etc.)	original research articles (proceedings included), English language, and the development of a GAN model using a RT dataset.
Selection process (who conducted the selection, whether it was conducted independently, how consensus was obtained, etc.)	Two researchers with expertise in deep learning, quickly scanned the abstracts to exclude irrelevant articles.
Any additional considerations, if applicable	the reference list from the selected articles to include the related ones which were not found by the initial search.

DEVELOPMENT OF GENERATIVE ADVERSARIAL MODELS

Generative Adversarial Networks (GAN) are inspired by game theory. The basic structure of the GANs model is shown in Figure 1(a). It is composed of a Generator G and Discriminator D and it aims to optimize these components alternately according to the Minimax game logic until they can't beat each other (Nash equilibrium).

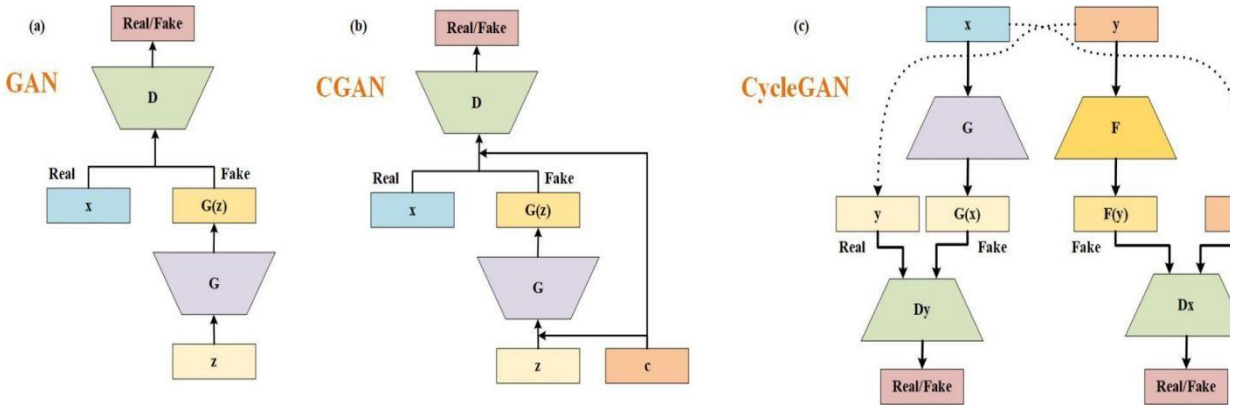


Figure 1.The structure of GANs model. (a) Generative Adversarial Network (GAN), (b) Conditional GAN (CGAN) ,(c)Cycle GAN. G and D are respectively the generator and discriminator. x and y present the real data come from different dataset. z present random noise. G and F are the Generator. D , Dx, and Dy are discriminators. c is the class information of x.

The generator G takes as input a vector z obeying the standard normal distribution N(0,1) and creates the target data distribution G(z). The goal of generator G is to synthesize new data in such a way that the discriminator D cannot distinguish it from a real one. The discriminator D can be seen as a classification network that distinguishes whether this new input data is real or not.

The Nash equilibrium is reached when the generator G synthesizes data G(z) hard to distinguish from the real ones and the discriminator D can classify real and fake data with high precision.

In the process of training, indeed, the update of the generator G tries to make the synthetic data classified as the real ones, so that the synthetic data is closer

to the decision boundary and the real image. While, the discriminator D plays the role of a binary classifier, and each update of the discriminator enhances its ability to distinguish between real data and synthetic data, which means dividing the correct decision boundary between the two kinds of data. With the continuous training of alternate iterations, the synthetic data will close to the real image, which will eventually make them indistinguishable to the discriminator D, so that the generator G can fit the real data with a high degree of realism.

However, the basic GAN generates data from noise so there are still some shortcomings: first, the class of the generated data cannot be controlled and, secondly, the transferring between two different clinical imaging modalities cannot be done by a basic GAN. To overcome these problems, some improvements to GAN based model were made.

CGAN (Conditional GAN)

In the training processing of GAN, the random noise Z is used as a priori information in the comparison training process, which greatly improves the calculation efficiency when the amount of data is too large. However, too much random noise will lead to the uncontrollability of the training process and experimental results, which greatly reduces the accuracy of the network. To solve this problem, supervised learning or semi-supervised learning is added based on GAN to effectively restrict the generation process and increase the stability of the network during training. Conditional GAN is such an improved model (1). The CGAN structure shown in Figure 1(b)

Cycle GAN

In the real world, it is difficult to obtain a large amount of paired image data that arises from the same individual at different modalities or machines. Therefore, Zhu et al proposed the Cycle-consistent Generative Adversarial Networks (Cycle GAN) in 2017 to solve the problem of converting the images between different modalities with unpaired data (2).

The Cycle GAN consists of two identical GAN models with a generator and discriminator respectively. The generator is trained for getting a mapping between data source distribution x and y.

The discriminators are the same as the traditional GAN model to determine whether the data is real or synthetic.The structure of the Cycle GAN structure is shown in Figure 1(c).

x, z, and G(z) represent respectively the real data, the input data, and the synthetic data generated by generator G.

For (c), F and G present two generators that generate fake data from y and x. The Dx and Dy present the discriminators for distinguishing between real and fake data created by the generators.

Evaluation metrics

The evaluation metrics needed to evaluate the quality of the synthetic data in GAN-based radiotherapy applications are divided into three different groups: image similarity, dose performance, and plan evaluation (shown in Table 1). They can be selected according to the specific RT tasks.

Table 2 Evaluation metrics for GAN applications.

Category	Metrics	Full Name	Definition	Function
Image	ME	Mean error	The average difference between the estimated values and the actual value.	$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n} f_i - y_i$
Image	MSE	Mean square error	The average squared difference between the estimated values and the actual value.	$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (f_i - y_i)^2$
Image	MAE	Mean absolute error	The average absolute difference between the estimated values and the actual value.	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} \lvert f_i - y_i \rvert$
Image	MRE	mean relative error	The ratio of the mean absolute error to the mean value of the quantity being measured.	$\mathrm{MRE} = \frac{1}{n}\sum_{i=1}^{n} \frac{\lvert f_i - y_i \rvert}{y_i}$

Image	SNU	Spatial Nonuniformity	The maximum and minimum percentage differences from the mean irradiance.	$\mathrm{SNU} = \frac{\overline{\mathrm{HU}_{\max}} - \overline{\mathrm{HU}_{\min}}}{1000} \times 100\%$
Image	PSNR	Peak signal to noise ratio	The ratio between the maximum value of an image and the value of corrupting noise affects the fidelity of its quality.	$\mathrm{PSNR} = 10 \log_{10} (\frac{\mathrm{MAX_i}}{\mathrm{MSE}})$
Image	SSIM	Structural similarity metric	The method to evaluate the quality of images.	$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x{}^2 + \mu_y{}^2 + c_1)(\sigma_x{}^2 + \sigma_y{}^2 + c_2)}$
Image	NCC	normalized cross correlation	The normalized inverse Fourier transform of the convolution of the Fourier transform of two images.	$\mathrm{NCC} = \frac{\mathrm{cov}(I_0, I)}{\sigma_{I_0}\sigma_I}$

Image	LPIPS	Learned Perceptual Image Patch Similarity	The method to measure the perceptual difference between two images	$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|\|w_l \odot (y^l_{hw} - y^l_{0hw})\|\|_2^2$
Dose	DVH difference	Dose-volume histogram difference	Difference between the dose-volume histogram (DVH) is a histogram relating radiation dose to tissue volume in radiation therapy planning.	$\text{DVH difference} = \sum_{i=1}^{n} \text{DVH}_{xi} - \text{DVH}_{yi}$
Dose	Hausdorff distance	Hausdorff distance	The method measure the distance between two point sets.	$H(A,B) = \max(h(A,B), h(B,A))$ $h(A,B) = \max_{a \in A} \{ \min_{b \in B} \|a - b\| \}$ $h(B,A) = \max_{b \in B} \{ \min_{a \in A} \|b - a\| \}$

Plan	CI	Conformity index	The method to quantitatively assess the quality of radiotherapy treatment plans, and represents the relationship between isodose distributions and target volume.	$\mathrm{CI} = \frac{\mathrm{V_{RI}}}{\mathrm{TV}}$
Plan	HI	Homogeneity Index	The method to calculate the uniformity of dose distribution in the target volume.	$\mathrm{HI} = \mathrm{D_{2\%}} - \mathrm{D_{98\%}}$
Plan	APE	Averaged Prediction Error	The average of the actual and the estimated doses difference	$\mathrm{APE} = \frac{1}{n}\sum_{i=1}^{n} \lvert Ground\ Truth_i - Prediction_j \rvert$

Abbreviations:n: the amount of the pixel in the images. is the pixel value of the synthetic image. is the pixel value of the target image. is the maximum possible pixel value of the image. x is the target images, y is the synthetic images. are the mean value of x and y. is the variance of x. is the variance of y. is the covariance of x and y. c1 and c2 are two variables to stabilize the division with weak denominator c1=(k1L)2,c2=(k2L)2,L is the dynamic range of the pixel-values, k1 =0.01 and k2=0.03 by default. presents covariance, and are the standard deviation of target and generated images, respectively. h(A,B) and h(B,A) calculate the maximum distance

between point groups. is the reference dose volume, and TV is the target volume. and are the percentage dose to 2% and 98% target volume. and present the Dose-volume histogram difference distribution of two sets. and are the averaged maximum and the minimum intensity of region of interests (ROIs) of patients' data.

Evaluation metrics for calculation similarity between synthetic data and target

Normally, it is difficult for human eyes to evaluate the similarity between synthetic images and the target ones. These metrics can quantify the similarity between them are shown in Table 2 image-related metrics.

The Mean Squared Error (MSE) and Mean Average Error (MAE) are the metrics that refer to the expected value of the difference between the synthetic and target data. The higher score means the bigger difference between them. The Peak Signal to Noise Ratio (PSNR) represents the ratio of the energy of the peak signal to the average energy of the noise. The higher the score, the smaller the distinction between target and synthetic data.

The above metrics only calculate the gap between one-to-one correspondence pixels without considering the other positions. This treats the image as isolated pixels, while ignoring visual features, especially the local structural information. The structural information has great influence on the subjective evaluation of medical images.

Conversely, to address above shortage, the Structural Similarity Index Measure (SSIM) consider a region of pixels when calculating the difference between two images. When the two images are identical, the value of SSIM is equal to 1.

To align more the quantitiave evaluation of image similarity to the visual inspection, the Learned Perceptual Image Patch Similarity (LPIPS) (5) was proposed also known as Perceptual Loss. It is used to measure the difference between two images in subject feeling contains rich image information such as texture color and texture primitives.The lower the value of LPIPS, the more similar the two images are.

Evaluation metrics for synthetic Dose performance

This type of evaluation is different from the evaluation method which compares the similarity between the real target and the synthetic data. The similarity metrics are not the best method considering the evaluation method between synthetic and target doses. Based on the aforementioned motivation, to evaluate the performance of the synthetic dose, the commonly used methods are shown in Table 2 dose-related metrics.

Dose-Volume Histogram (DVH) difference compares the difference of the dose-volume histogram in RT planning between the generated and the real one. The lower the score, the more realistic the generated doses are. Moreover, the Hausdorff distance also can be used to calculate the difference between the synthetic DVH and the real one.

Evaluation metrics for plan evaluation

For evaluation of the feasibility of the generated plan, a comparison between the real and generated plan is required (shown in Table 2 plan-related metrics.). Furthermore, the conformity index (CI) and Homogeneity Index (HI) scores that can evaluate the conformity and uniformity of dose distribution also should be considered. Averaged Prediction Error (APE) calculates the averaged ratio of the prescription dose and the difference between the ground truth and the prediction(4).

GENERATIVE ADVERSARIAL MODELS FOR CT TRANSLATION AND SYNTHESIS

Different modalities of medical images can provide multimodal information, that can be used for a better diagnosis and RT planning.

However, in a realistic situation, limitations due to unnecessary costs and radiation protection of the patient, make it hard to collect all the desired imaging modalities from a single patient. Fortunately, while there is a different focus and range distribution between modalities, there is still some hidden information in one type of image that may prevent the need to take another one. This is why the cross-mode image synthesis method is feasible(6).

For treatment planning, CT is always required, while delineations are often performed on MR, for example for pelvic or head and neck tumors. However, the transformation from MR to CT will lead to an undesirable 2-5 mm systematic error(7).

To address the systematic error, MR-only treatment planning was proposed, in which only MRI is required as the sole input modality. It can protect the patient from CT radiation doses and benefit a pediatric patient who has less dose upper(8).

The GAN-based method has the feasibility of mapping the information and generating the image from different modalities. GAN can generate synthetic CT (sCT) from MRI, thanks to allow performing the calculation of the dose accuracy with a single MRI-only workflow.

For synthetic CT methods, Yingzi Liu et al. (2019) tried to integrate dense block into 3D Cycle GAN to effectively generate the CT from T2-weighted MRI. Dense block connects all blocks that make up the model directly into each other, leading to each block gets additional inputs from all previous blocks and passes on its own output to all subsequent blocks. This ensures the maximum transmission of information between blocks in the model(see in Figure 2(a)).The proposed method achieved (51.32 16.91HU and less than 1% DVH difference compare to the one generated by real CT. This demonstrates the feasibility of GAN-based applications for the development of the MRI-only workflow for prostate proton radiotherapy(9).

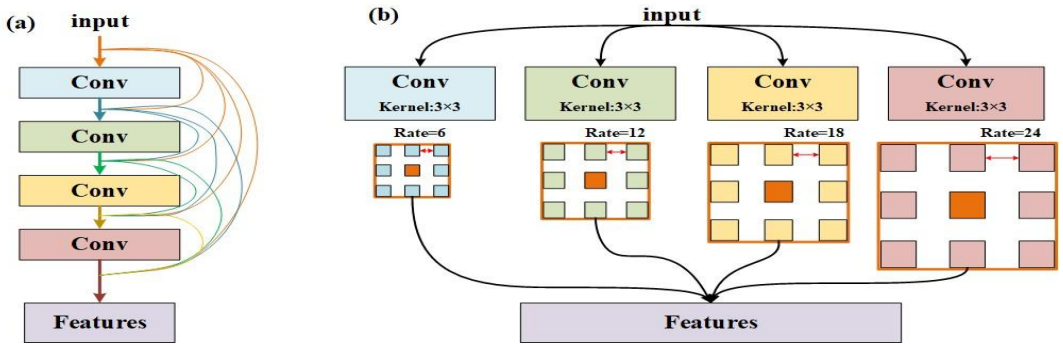


Figure 2. The structure of blocks. (a) dense block, (b) Atrous Spatial Pyramid Pooling (ASPP) block. Conv presents the convolutional layer.

To evaluate whether the sCT is accurate enough for MRI-only treatment planning, Samaneh Kazemifar et al. (2020) used Mutual Information (MI) which is used to evaluate non-linear relations between two variables as the loss function in GAN to overcome the misalignment between two modalities in training model processing. In fact, one of the largest issues when translating MR to CT is that the two images belong to different spaces: the frequency one and the tissue density respectively. The above methods achieve a mean absolute difference below 0.5%(0.3Gy) for the prescription dose for CTV and below 2%(1.2Gy) for OARs. The excellent result illustrates that GAN is a potentially promising method to generate sCT for MRI-only treatment planning of patients with brain tumors in intensity-modulated proton therapy(10, 11).

Zizhao Zhang et al. (2018) proposed a Cycle- and Shape-Consistency GAN to synthetic realistic looking 3D images using unpaired data and improve the volume segmentation by generated data. With the extensive experiment on a 4496 CT and MRI dataset, it proves that both tasks are beneficial to each other, and coupling them has better performance than exclusively(12).

In another study, Sven Olberg et al. (2019) designed an Atrous Spatial Pyramid Pooling (ASPP) structure in GAN model. ASPP is a structure that captures objects and image features on multiple scales, thanks to the introduction of multiple filters that have complementary effective fields of view (as shown in Figure 2(b)).

The proposed method achieved a RMSE(17.7±4.3), a SSIM(0.9995±0.0003), a PSNR (71.7±2.3), and great dose performance based on sCT with more than 98% passing rates in the 1042 images test set. The excellent result illustrates the designed structures can improve the performance of traditional GAN(13).

Meanwhile, different from traditional sCT which is generated from a single MR sequence, Yuhei Koike et al. (2019) tried to generate and assess the feasibility of sCT from multi-sequence MRI using GAN for brain radiotherapy treatment planning. With the small, clinically negligible difference (less than 0.1% DVH difference and 0.6 1.9 mm overall equivalent path length difference), CGAN is feasible to generate sCT from multi-sequence include T1-weighted,T2-weighted and fluid-attenuated inversion recovery (Flare) MRI(1).

Vincent Bourbonne et al. (2021) was the first study which demonstrate the GAN-generated CT from diagnostic brain MRIs have comparable performance to initial CT for the planning of brain stereotactic RT. In their study , the 2D-UNet was selected as the backbone of generator. Through experiments on a dataset of 184 patients, there were no significant statistical differences regarding ICRU 91s endpoints, which means the synthetic CT and initial CT has high similarity for both the organs at risk and the target volumes(14).

On the other hand, the time cost is also worth considering the feature of radiotherapy applications.

Matteo Maspero et al. et al. (2018) tried to assess whether the GAN method can rapidly generate sCT to be used for accurate MR-based dose calculations in the entire pelvis. As the result, the CGAN required 5.6s and 21s for a single patient on GPU and CPU, respectively. It achieves less than 2.5% calculated DVH differences on sCT and CT. Results suggest that the sCT generation was

sufficiently fast and accurate to be integrated into an MR-guided radiotherapy workflow(15).

The 8 key publications focusing on GAN application for CT translation and synthesis are presented in Table 3.

Table 3 Key publications focusing on GAN application for CT translation and synthesis

Authors	Year	country	dimension	model information	target	Region	patient	Evaluation method	conclusion
Yingzi Liu	2019	USA	3D	cycleGAN	T2-weighted MRI to CT	pelvic	17	MAE, DVH difference	We applied a novel learning-based approach to integrating dense-block into cycleGAN to synthesize pelvic sCT images from routine MR images (Lei et al 2019) for potential MRI-only prostate proton therapy. The proposed method demonstrated a comparable level of precision in reliably generating sCT images for dose calculation, which supports further development of MRI-only treatment planning. Unlike photon therapy, the accuracy of proton dose calculation is highly dependent on stopping power rather than HU values. Therefore, the future directions of MRI-only proton treatment planning include prediction of the stopping power map based on the MR images and generating

									elemental concentration maps that can be used for Monte Carlo simulations.
Samaneh Kazemifar	2019	USA	2D	GAN	MRI to CT	Brain	77	MAE	In conclusion, MRI-only treatment planning will reduce radia tion dose, patient time, and imaging costs associated with CT imag ing, streamlining clinical efficiency and allowing high-precision radiation treatment planning. Despite these advantages, several challenges prevent clinical implementation of MRI-only radiation therapy. Through the method we have proposed here, synthetic CT images can be generated from only one pulse sequence of MRI images of a range of brain tumors. This method is a step toward using artificial intelligence to establish MRI-only radiation therapy in the clinic.
Samaneh Kazemifar	2020	USA	2D	GAN with Mutual Informati on (MI) as the loss function	MRI to CT	Brain	77	DVH difference	This work explanted the feasibility of using sCT images generated with a deep learning method based on generative adversarial networks (GANs) for intensity-modulated proton therapy. We tested the method in brain tumors—some of them located close

									to complex bone, air, and soft-tissue interfaces—and obtained excellent dosimetric accuracy even in those challenging cases. The proposed method can generate sCT images in around 1\,s without any manual pre- or post-processing operations. This opens the door for online MRIguided adaptation strategies for IMPT, which would eliminate the dose burden issue of current adaptive CT-based workflows, while providing the superior soft-tissue contrast characteristic of MRI images.
Zizhao Zhang	2018	USA	3D	Cycle- and Shape-Consisten cy GAN	MRI to CT and Segmentati on task	heart	4496 images	segmentati on score	we present a method that can simultane ously learn to translate and segment medical 3D images, which are two significant tasks in medical imaging. Train ing generators for cross-domain volume-to-volume transla tion is more difficult than that on 2D images. We address three key problems that are important in synthesizing realis tic 3D medical images: 1) learn from unpaired data, 2) keep

									anatomy (i.e. shape) consistency, and 3) use synthetic data to improve volume segmentation effectively. We demon strate that our unified method that couples the two tasks is more effective than solving them exclusively. Extensive ex periments on a 3D cardiovascular dataset validate the effec tiveness and superiority of our method.
Sven Olberg	2019	USA	2D	atrous spatial pyramid pooling (ASPP)GAN	MRI to CT	breast	2400	RMSE, SSIM, PSNR	In this study, we have evaluated the robustness of the con ventional pix2pix GAN framework that is ubiquitous in the image-to-image translation task as well as the novel deep spa tial pyramid framework we propose here. The proposed framework demonstrates improved performance in metrics of training time and image quality, even in cases when training data are limited. The success of the framework in sCT genera tion is a promising step toward an MR-only RT workflow that eliminates the need for CT simulation and setup scans while enabling online adaptive therapy

									applications that are becom ing ever more prevalent in MR-IGRT.
Yuhei Koike	2019	Japan	3D	CGAN	multi-sequence MRI to CT	Brain	580	DVH difference，clinically negligible difference	images from multi-sequence brain MR images using an adversarial network. The performance of the model was evaluated by comparing the image quality and the treatment planning with those of the original CT images. The use of multiple MR sequences for sCT generation using cGAN provided better image quality and dose distribution results compared with those from only a single T1w sequence. The CT number of the generated sCT images showed good agreement with the original images, but not in the bone regions. Impacts on the dose calculations were within 1%. These findings demonstrate the feasibility and utility of sCT-based treat ment planning and support the use of deep learning for MR-only radiotherapy
Matteo Maspero	2018	the Netherl ands	2D	cGAN	MRI to CT	abdomi nal, pelvic	91	DVH,mean Dice similarity	To conclude, this study shows, for the first time, that sCT images generated with a deep learning approach employing a cGAN

								coefficient	and multi-contrast MR images acquired with a single acquisition facilitated accurate dose calculations in prostate cancer patients. It was further shown that without retraining the network, the cGAN could generate sCT images in the pelvic region for accurate dose calculations for rectal and cervical cancer patients. A particularly attractive feature of our method is its speed as it allows sCT generation within 6 seconds on a GPU and within 21 seconds on a CPU. This could be of particular benefit for MRgRT applications.
Yingzi Liu	2019	USA	3D	cycle GAN	MRI to CT	abdominal	21	MAE,DVH difference	We applied a novel learning-based approach to integrate dense-block into cycle GAN to synthesize abdominal sCT images from routine MR images for potential MRI-only liver proton therapy. The proposed method demonstrated a comparable level of precision in reliably generating sCT images for dose calculation, which supports further development of

									MRI-only treatment planning. Unlike photon therapy, the accuracy of proton dose calculation is highly dependent on stopping power rather than HU values. Therefore, the future directions of MR-only proton treatment planning include prediction of the stopping power map based on the MR images or generating elemental concentration maps that can be used for Monte Carlo simulations.

GENERATIVE ADVERSARIAL MODELS FOR DOSE AND PLAN CALCULATION

The RT planning is highly dependent on the clinical experience and skills of the radiotherapy physicist or dosimetrist, as well as their knowledge of radiotherapy physics and understanding of the Treatment Planning System (TPS). With advances in DL, especially GANs, automatically generating 3D RT dose distributions from medical images like CTs and MRIs became possible. In the past few years, several methods can generate dose distribution or Intensity-Modulated Radiation Therapy (IMRT) plans from different kinds of inputs.

To overcome an limited dataset situation for a deep learning task, Wentao Liao et al. (2021) proposed an Auxiliary Classifier GAN (ACGAN) to synthesize dose distribution according to tumor types and beam types. The proposed with excellent PSNR (75.6032) and MS-SSIM (0.95120) results, that demonstrate the synthetic dose distribution is close to the real one which can be used for increasing the training set for dose prediction tasks(16).

In order to different organs to jointly constrain the dose distribution of each organ in model training to achieve better PTV dose coverage and OARs sparing. Chongyang Caoet al. (2021) designed an Adaptive Multi-organ Loss (AML) - based Generative Adversarial Network (AML-GAN). The AML loss can measure the gap between synthetic dose and real one on whole dose , OAR and PTV distribution. The experiment demonstrates the proposed method achieves state-of-art APE (0.021±0.014) in terms of OARs and PTV(16).

Dose calculation is a time-consuming task, which sharply decreases the RT workflow efficiency. Therefore, some GAN-based dose simulation methods to decrease the time cost and generate accurate dose distribution were proposed.

In another work, Xinyi Li et al. (2021) designed a CGAN-based model which can real-time generate fluence map from CT. This model containing a novel PyraNet that implements 28 classic ResNet blocks in pyramid-line concatenations as the generator. The proposed method was evaluated on 15 plans, the AI model only cost 3 s to predict a fluence map without statistical significance from the real one. This approach holds great potential for clinical applications in real-time planning(17).

While Xiaoke Zhang et al. (2021) proposed a discovery cross-domain GAN (DiscoGAN) to generate comparable accuracy to Monte Carlo simulation

without much time cost. The proposed method was evaluated on abdominal cases, thoracic cases, and head cases by MRE and achieved no systematic deviation. It demonstrates the proposed method has great potential for accurate dose calculation compared to the Monte Carlo simulation method(18).

For RSP prediction application, Joseph Harms et al. (2020) used a Cycle GAN, relying on a compound loss function designed for structural and contrast preservation, to predict relative stopping power (RSP) maps from CBCT. With the result of a MAE (0.06±0.01) and a ME (0.01±0.01) between RSP generation from CT and CBCT, this method provides sufficiently accurate prediction which makes CBCT-guided adaptive treatment planning for IMPT become feasible(19).

The 5 key publications focusing on GAN application for dose and plan calculation and synthesis are presented in Table 4.

Table 4 Key publications focusing on GAN application for dose and plan calculation and synthesis

Authors	Year	country	dimension	model information	target	Region	patient	Evaluation method	conclusion
Wentao Liao	2021	China	2D	Auxiliary Classifier GAN(ACGAN)	Synthesis of Radiotherapy dose	head and neck	110	MS-SSIM and PSNR,	We proposed the Dose-ACGAN for Data Enhancement Work of Radiotherapy Deep Learning. The dose distribu tion of specified tumor category or beam number category can be customized successfully, and the desired dose dis tribution map can be customized by controlling two vari ables together. One purpose of Dose-ACGAN is to generate multi-classification dose distribution enhancement data, train and generate dose distribution data of specified tumor type or beam number type. Used for training dose data required by radiotherapy plan using AI model. The next step in future work is to introduce CT data, contour information and beam angle information to customize the dose distribution corre sponding to the

									predicted CT contour. Provide other better ideas for automatic planning, By using the dose distribution data of normal and effective radiotherapy plans, a large num ber of high-quality tagged data can be generated, such as tumor types, beam types, etc. The reliability and accuracy of automatic dose prediction model for radiotherapy will be improved effectively. A further plan is to enhance the data in this paper for a comparative study of different AI of predicting dose tasks.
Xinyi Li	2021	USA	4D	CGAN contains a novel PyraNet that implements 28 classic ResNet blocks in pyramid-line	CT to IMRT plannin g	orophar yngeal	231	CI,HI	In this work, an AI agent was successfully developed as a DL approach for oropharyngeal IMRT planning. Without time-consuming inverse planning, this AI agent could automatically generate an oropharyngeal IMRT plan for the primary target with acceptable plan quality. With its high implementation efficiency, the developed AI agent holds great potentials for clinical application after future development validation studies.

				concatenations as generator					
Chongyang Cao	2021	China	3D	an Adaptive Multi-organ Loss (AML) based GAN (AML-GAN)	automatic dose prediction of cervical cancer,	cervical cancer	75	CI，OAR,APE	we propose an Adaptive Multi-organ Loss based Generative Adversarial Network, namely AML-GAN, to predict the dose distribution map from CT images automatically. Innovatively, besides the global dose prediction loss, we have also considered the dose prediction losses of PTV and individual OAR separately, making sure that the predicted dose distributions of local areas are as accurate as possible. Extensive experiments demonstrate that our proposed AML-GAN outperforms all state-of-the-art approaches.
Xiaoke Zhang	2021	China	3D	A discovery cross-domain GAN (DiscoGAN)	Synthesis of Radiotherapy dose	head,abdomen,thorax	36	MRE,MAE	We developed a novel machine learning model based on DiscoGAN for dose calculation in proton therapy, which offers comparable accuracy (below 5%) to MC simulation but of reduced computational workload. The

									relationship between MRE and other factors such as dose, beam energy and location within the beam cross-section was examined. The proposed DiscoGAN has proven effective in identifying the relationship among dose, SP and HU in three dimensions. If successful, our approach is expected to find its potential use in more advanced applications such as inverse planning and adaptive proton therapy.
Joseph Harms	2020	Atlanta	2D	cycle GAN	RSP prediction	head-and-neck	23	MAE,ME,PSNR,SSIM	This work presents the use of a deep-learning algorithm for generation of RSP maps directly from cone-beam CTs. The proposed method closely matches the quantitative values of the planning CT and the geometric qualities of the daily CBCT. When used for dose calculation, the method shows strong agreement to a DIR based method that is in clinical use for dose evaluation while patients are under treatment. The proposed method was validated on head-and-neck patient CT

									images, a particularly difficult image set to work with due to the presence of several soft tissue structures, changing body shapes, and the frequent presence of metal artifacts. However, the proposed method still produced and median MAE of around 0.06 when compared to the planning CT and a median structural similarity of 0.88. Gamma analysis between the proposed method and the DPCT method using 3% dose difference and 3 mm distance-to-agreement had an average passing rate around 96% showing that the method can be used for dose evaluation.

GENERATIVE ADVERSARIAL MODELS FOR QUALITY IMPROVEMENT

Traditional medical image enhancement methods are mainly used to improve medical images with low contrast, narrow dynamic range, uneven intensity distribution, and blurred edges. This is given by studying effective image enhancement algorithms to improve the image quality of existing medical images, improve their resolution or emphasize the important texture information and suppress noise. After those passages, the images become more standard and suited for computer-aided diagnosis (CAD) systems.

Medical images can be acquired with different methods according to the need that doctors have to treat a particular case. Moreover, medical image data is much more complicated than natural image data, and it is difficult to get detailed information directly on the original data. These characteristics mean that medical images have a relatively greater demand for image enhancement algorithms that are beyond the capabilities of traditional algorithms.

In these cases, the GAN-based models can improve the quality of medical images (e.g., denoising).

Several GAN-based methods are trained using paired data. Serdar Charyyev et al. (2022) designed a residual attention GAN to synthetic dual-energy CT (DECT) from single energy(SECT). The MAE, PSNR, and NCC were applied to evaluate the performance of the synthetic high and low energy CT was 36.9 HU, 29.3 dB, 0.96 and 35.8 HU, 29.2 dB, and 0.96, respectively. The proposed method has potential feasibility for proton radiotherapy by generating DECT from SECT (20).

In another study, Kui Zhao et al. (2020) designed a supervised GAN with the Cycle-consistency loss, Wasserstein distance loss, and an additional supervised learning loss, named S-Cycle GAN, to synthetize full-dose PET (FDPET) from low-dose PET(LDPET). The model was evaluated in 10 testing datasets and 45 simulated datasets by NRMSE, SSIM, PSNR, LPIPS, SUVmaxand SUVmean, and the results show this method achieves accurate, efficient, and robust performance(21).

A study by Dongyeon Lee et al. (2021) used Cycle GAN to synthetic kilovoltage CT (kVCT) from Megavoltage CT (MVCT). With the excellent average MAE, RMSE, PSNR, and SSIM values were 18.91 HU, 69.35 HU, 32.73 dB, and 95.48, respectively. This Cycle GAN can improve the MVCT to KVCT while

maintaining the anatomical structure in radiation therapy treatment planning(22).

For CBCT improvement, some methods can synthesize target images from unpaired data.

Jinsoo Uh et al. (2021) applied Cycle GAN to correct abdominal and pelvic CBCT between children and young adults in the presence of diverse patient sizes, anatomic extent, and scan parameters. The performance of the model has significantly outperformed performance in the 14 patients' test set (47 ± 7 HU versus 51 ± 8 HU; paired Wilcoxon signed-rank test, $P < 0.01$). This proposed method can decrease the impact of anatomic variations in CBCT images for proton dose calculation(23).

Sangjoon Park et al. (2021) designed a spectral blending technique to combine trained sagittal and coronal directions Cycle GAN to synthetic CT from CBCT. The proposed method achieves better performance than the existing Cycle GAN on PSNR (30.6027 versus 29.4991), NMSE (1.3442 versus 1.5874), and SSIM (0.8977 versus 0.8674)(24).

Yingzi Liuet al. (2020) designed a self-attention Cycle GAN to synthetic CT from CBCT. There is no significant different performance between the CT-based contours and treatment plans from sCT on MAE and DVH differences. The result indicates that the sCT from CBCT can be used for accurate dose calculation(25).

Except self-attention, Liugang Gao et al. (2021) proposed an attention-guided Cycle-GAN which contains two equipped with attention module generators to generate attention mask. It can make generator pays attention to the important part of images to eliminate numerous artifacts. By training and testing on a dataset of 170 patients, the proposed method has similar quality with real CT in MAE (43.5 ± 6.69), SSIM(93.7 ± 3.88), PSNR(29.5 ± 2.36), mean and standard deviation (SD) HU values ($P < 0.05$). Besides that, sCT provided the highest gamma passing rates (91.4 ± 3.26) in dose calculation compared with GAN methods. These demonstrate that the proposed method can trained by unpaired data to generate high-quality CT from CBCT(26).

The 6 key publications focusing on GAN application for quality improvement are presented in Table 5.

Table 5 key publications focusing on GAN application for quality improvement.

Authors	Year	country	dimension	model information	target	Region	patient	Evaluation method	conclusion
Serdar Charyyev	2022	USA		a residual attention GAN	single energy CT (SECT) to synthetic Dual energy CT (DECT)	head-and-neck	70	PSNR,NCC,ME	We applied a novel deep learning- based approach, namely residual attention GAN, to synthesize sLECT and sHECT images from SECT images for potential applications in the clinic where a DECT scanner is not available. The proposed method demonstrated a comparable level of precision in reliably generating synthetic images when compared to ground truth, and noise robustness in derived SPR maps.
Kui	2020	China	3D	a	LDPET(l	brain	109	NRMSE，	In conclusion, we have

Zhao				supervised cycleGAN	ow-dose PET) to FDPET(full-dose PET)			SSIM, PSNR, LPIPS, SUVmax and SUVmean	introduced a novel deep learning based generative adversarial model with the cycle consistent to estimate the high-quality image from the LDPET image. The proposed S-CycleGAN approach has produced comparable image quality as corresponding FDPET images by suppressing image noise and preserving structure details in a supervised learning fashion. Systemic evaluations further confirm that the S-CycleGAN approach can better preserve mean and maximum SUV values than other two deep learning methods, and suggests the amount ofdose reduction should be carefully decided according to the acquisition protocols and clinical usages.

Dongyeon Lee	2021	Korea	2D	CycleGAN	MVCT to KVCT	prostate	11	Hausdorff distances，DVH difference，OAR	In this study, we developed a synthetic approach based on cycleGAN to produce skVCT images from MVCT images for applying MVCT to adaptive helical tomotherapy treatment. The proposed method generates clear CT images by including the anatomical features of MVCT images through a deep learning algorithm without an additional calibration process. The cycleGAN employed in this study was optimized using augmented training data derived from a small number of CT images. The proposed method successfully enhanced the quality of the MVCT images, preserving the anatomical structures of MVCT and restoring the HU to values

									similar to those of kVCT, along with providing reduced noise and improved contrast. The MVCT can be better utilized for aligning both the patient setup for daily treatment and the dose re-calculations for the ART process by considering the distributions of the HU values of skVCT images approach and those of the planning kVCT images.
Jinsoo Uh	2021	USA	2D	cycle GAN	correct CBCT between children and young adults	abdominal, pelvic	64	MAE and ME	Using both abdominal and pelvic images for training a single deep learning model and normalizing age-dependent body sizes helped mitigate the impact of anatomic variations in CBCT images. Delivered proton dose can be accurately estimated from the corrected CBCT for children and young adults with abdominal or pelvic

									tumors.
Sangjoon Park	2020	Korea	2D	CycleGAN	CBCT to CT	Lung	10	PSNR, NMSE,SSIM	In this paper, we proposed a novel unsupervised synthetic approach based on CycleGAN to produce CT images from CBCT images, which requires only unpaired CBCT and CT images for training. The proposed method properly combined CycleGAN and spectral blending technique, generating CT images by CycleGAN and further reducing the artifacts from missing frequency problem by spectral blending. Our method outperforms the existing CylceGAN-based method both qual itatively and quantitatively.
Yingzi Liu	2020	USA	3D	CycleGAN	CBCT to CT	pancreas	30	MAE，DVH，SNU,NCC	The image similarity and dosimetric agreement between the CT and sCT-based plans

									validated the dose calculation accuracy carried by sCT. The CBCT-based sCT approach can poten tially increase treatment precision and thus minimize gastrointestinal toxicity

Discussion

This review clusters the recent GANs application in RT articles into three groups: CT translation and synthesis, dose and plan calculation, and image quality improvement.

Among the included studies, in terms of treatment sites, the majority focused on the head and neck (10/23). This was followed by the abdomen (8/23) and the thorax (5/23).

In RT applications, compared with the traditional handcraft method, GAN brings a potential significant improvement. The GAN model was trained to attain target high-level information distribution rather than simple geometrical and texture deformations. That makes GAN capable of establishing a nonlinear mapping between two different modalities for image translation tasks such as MR and CT or CBCT and CT.

Besides the accuracy prediction, the GAN model has less time consumption compared with traditional methods such as Monte Carlo simulation. The GAN applications can only cost a few seconds per patient which is an unimaginable performance using traditional methods. This will enormously increase the efficiency of RT.

All the publications mentioned in this review prove that GAN applications have great performance in modality translation, dose calculation, and image quality improvement tasks by maintaining anatomical and functional information, which has great benefits in RT workflow.

However, the training of GAN models is a challenging task, as it contains two models (generator and discriminator) with their own opposite targets. It differs from traditional DL model which contain a clear target, such as classification model that can stop training when it can achieve high accuracy in validation sets.

Therefore, in order to train the model, the discriminator should train first to make it has a preliminary classification ability to recognize real data from just noise images. Thanks to this preliminary training it is then possible to train the generator and so the whole GAN model. Then, the final training (adding

generator) can stop when the accuracy of discriminator retains 0.5 or the loss function of discriminator is unable to continue descent, that means the discriminator can no longer distinguish between real and synthetic data given by generator.

The setting of hyper parameters also significantly affects whether the model can be successfully trained. Important hyper parameters to take in consideration are batch size (BS), learning rate (LR), epochs and optimizer.

Batch size means the number of data feed to the model in per step of training, which should be the first determined hyper parameter. Too small and too large BS could cause problems of too long training time or difficult convergence of the model, respectively. Though the ratio of image size and GPU's memory will significantly limit the large of BS, the number around 10 is recommended as BS initial setting(27).

The epochs represent the number of times that model is trained on the whole data set. The larger epochs, the more time consumption is required for training. Therefore, 100 epochs were recommended to make the model get sufficient training without too much time consumption.

The optimizer is the algorithm that modifies the weights of the DL model during the trainingphase and the LR determines how much every iteration influences the weights of the model..

For the optimizer and LR, the Adaptive Moment estimation (Adam) optimizer and learning rate at 1e-4 are recommend and widely used as initial settings.

And, It is worth noting that, addition to the hyper parameters of the GANs model, the physical difference (contrasts, scanners, and patients etc.) which often not included during training also have a significant impact on the performance of the generated images. Especially for MR images, image contrast, or absolute image pixel values highly rely on scanning parameters and scanners, which makes the algorithm more difficult to train, less robustness , harder to migrate the model to data generated by other scanners, making it difficult to be widely used. Fortunately, there are many traditional and deep learning based image harmonization methods (adjusting the distribution of images from different sources so that they are close to each other) can overcome the above problem(28-30).

In order to develop and train a GAN model, there are many open-source deep learning framework can be selected such as Pytorch, Keras, Tensorflow and

Caffe. Among them, Pytorch developed by Facebook is the most widely recommended framework by DL researchers.

Thank to convolutional neural networks (CNN), the GAN model can learn the specific tissue and organ textures from the training sets such as brain, breast, and pelvic which is impossible in the traditional handcraft method. It makes GAN automatic maintain useful anatomical and functional details to achieve excellent performance.

But there still some potential risks for GAN applications in medical imaging tasks. Synthetic CT still remain some dimly visible artifacts in top and bottom areas in some applications(24). And GAN's tasks are highly dependent on data quality and quantity which make them have difficulty with nonstandard patient anatomies(7).

Although many studies (9, 13, 15, 25, 31-33) have conducted distribution comparison between synthetic and real one or validated their models in simulated clinical settings and have shown great potential to apply it in the real world. It still needs more evidence to apply it in the clinic, such as conducting clinical trials or embedding it into treatment planning systems to validate its application in daily clinical practice. Setting these exciting results aside, there are still some technical barriers that need to be overcome. For example, for MRI-based generation of synthetic CT, organ effects such as organ motion (13, 33) and organs containing cavities (15) will have an impact on the accuracy of the results or require manual intervention (11, 13), which needs to be supported by more relevant studies, such as advances in deformable alignment techniques. Therefore, we consider that the application of GANs techniques requires a series of synergistic developments from a technical point of view to be accomplished.

Another important consideration is the choice of the GAN model, that is influenced by the typology of task it has to face.As the most common selected model (shown in Figure 3), the Cycle GAN has great advantages over the other two GAN models for unpaired data sets. It is composed of two independent GANs which makes obtaining two different modalities of information distribution become feasible without requiring a one-to-one correspondence. This will significantly decrease the difficulty for the researchers to build a large enough dataset for training. The basic GAN instead, is the second most selected model. It has the simplest structure which makes it the researchers easier to build their own model to address specific tasks. Finally, the CGAN with

additional information as input can be used in the multi-class transformation tasks.

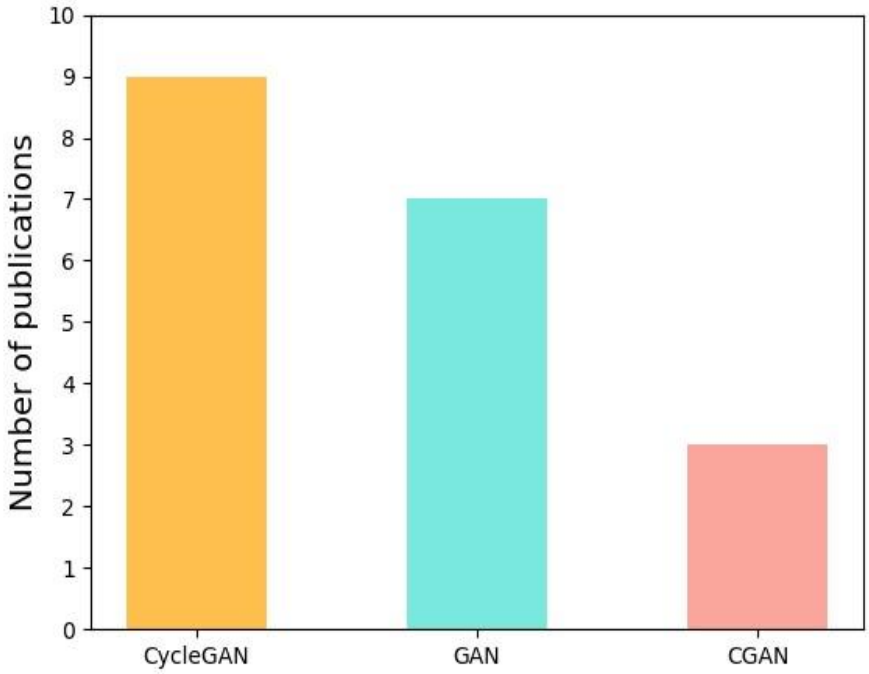


Figure 3. The distribution of the GAN applications in Radiation Therapy (RT).

As mentioned above, when the researchers have paired data but do not require to generate the specific image according to the input class, the basic GAN was recommended as the start. Otherwise, the CGAN is your recommended selection. Finally, when only unpaired data, the Cycle GAN is the only choice for this task(27).

Limitations and Future work

Though the GAN application has a strong power over image generation. There are still some limitations that need to be discussed. First, 3D GANs have a higher requirement in hardware which makes their deployment difficult in most hospitals, so, model compression or small models need to be considered in the future model design and deployment. Second, for plan and dose

calculation, GAN-based applications have less time consumption compared with the traditional methods. However, most of the methods still cannot achieve real-time performance. So, the acceleration methods for GANs are still worth studying. Third, the GAN-based application has not been applied in the real-world clinical trial, which means it is still unclear how much the GAN can help radiotherapy doctors. In this way, the GAN deployment in the real world needs to be done in future works.

Conclusions

In conclusion, the GAN model has already been widely used in RT. Thanks to their ability to automatically learn the anatomical features from different modalities images, improve quality images, generate synthetic images and make less time consumption automatic dose and plan calculation. Even though the GAN model still cannot replace the radiotherapy doctors' work, it still has great potential to enhance the radiologists' workflow. There are lots of opportunities to improve the diagnostic ability and decrease potential risks during radiotherapy and time cost for plan calculation.

References

1. Koike Y, Akino Y, Sumida I, Shiomi H, Mizuno H, Yagi M, et al. Feasibility of synthetic computed tomography generated with an adversarial network for multi-sequence magnetic resonance-based brain radiotherapy. Journal of Radiation Research. 2019;61(1):92-103.
2. Zhu J-Y, Park T, Isola P, Efros AA, editors. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision; 2017.
3. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. 2014;27.
4. Cao C, Xiao J, Zhan B, Peng X, Wu X, Zhou J, et al., editors. Adaptive Multi-Organ Loss Based Generative Adversarial Network For Automatic Dose Prediction In Radiotherapy. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI); 2021 13-16 April 2021.
5. Zhang R, Isola P, Efros AA, Shechtman E, Wang O, editors. The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE conference on computer vision and pattern recognition; 2018.
6. Sorin V, Barash Y, Konen E, Klang E. Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) - A Systematic Review. (1878-4046 (Electronic)).
7. Edmund JM, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. Radiation Oncology. 2017;12(1):28.
8. Dougeni E, Faulkner K, Panayiotakis G. A review of patient dose and optimisation methods in adult and paediatric CT scanning. European journal of radiology. 2012;81(4):e665-83.
9. Liu Y, Lei Y, Wang Y, Shafai-Erfani G, Wang T, Tian S, et al. Evaluation of a deep learning-based pelvic synthetic CT generation technique for MRI-based prostate proton treatment planning. Physics in medicine and biology. 2019;64(20):205022.
10. Kazemifar S, McGuire S, Timmerman R, Wardak Z, Nguyen D, Park Y, et al. MRI-only brain radiotherapy: Assessing the dosimetric accuracy of synthetic CT images generated using a deep learning approach. (1879-0887 (Electronic)).
11. Kazemifar S, Barragán Montero AM, Souris K, Rivas ST, Timmerman R, Park YK, et al. Dosimetric evaluation of synthetic CT generated with GANs for MRI-only proton therapy treatment planning of brain tumors. Journal of applied clinical medical physics. 2020;21(5):76-86.

12. Zhang Z, Yang L, Zheng Y, editors. Translating and Segmenting Multimodal Medical Volumes with Cycle- and Shape-Consistency Generative Adversarial Network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 18-23 June 2018.
13. Olberg S, Zhang H, Kennedy WR, Chun J, Rodriguez V, Zoberi I, et al. Synthetic CT reconstruction using a deep spatial pyramid convolutional framework for MR-only breast radiotherapy. Medical physics. 2019;46(9):4135-47.
14. Bourbonne V, Jaouen V, Hognon C, Boussion N, Lucia F, Pradier O, et al. Dosimetric Validation of a GAN-Based Pseudo-CT Generation for MRI-Only Stereotactic Brain Radiotherapy. Cancers. 2021;13(5).
15. Maspero M, Savenije MHF, Dinkla AM, Seevinck PR, Intven MPW, Jurgenliemk-Schulz IM, et al. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. Physics in medicine and biology. 2018;63(18):185001.
16. Liao W, Pu Y. Dose-Conditioned Synthesis of Radiotherapy Dose With Auxiliary Classifier Generative Adversarial Network. IEEE Access. 2021;9:87972-81.
17. Li X, Wang C, Sheng Y, Zhang J, Wang W, Yin FF, et al. An artificial intelligence-driven agent for real-time head-and-neck IMRT plan generation using conditional generative adversarial network (cGAN). Medical physics. 2021;48(6):2714-23.
18. Zhang X, Hu Z, Zhang G, Zhuang Y, Wang Y, Peng H. Dose calculation in proton therapy using a discovery cross-domain generative adversarial network (DiscoGAN). Medical physics. 2021;48(5):2646-60.
19. Harms J, Lei Y, Wang T, McDonald M, Ghavidel B, Stokes W, et al. Cone-beam CT-derived relative stopping power map generation via deep learning for proton radiotherapy. (2473-4209 (Electronic)).
20. Charyyev S, Wang T, Lei Y, Ghavidel B, Beitler JJ, McDonald M, et al. Learning-based synthetic dual energy CT imaging from single energy CT for stopping power ratio calculation in proton radiation therapy. The British journal of radiology. 2022;95(1129):20210644.
21. Zhao K, Zhou L, Gao S, Wang X, Wang Y, Zhao X, et al. Study of low-dose PET image recovery using supervised learning with CycleGAN. PloS one. 2020;15(9):e0238455.
22. Lee D, Jeong SW, Kim SJ, Cho H, Park W, Han Y. Improvement of megavoltage computed tomography image quality for adaptive helical

tomotherapy using cycleGAN-based image synthesis with small datasets. Medical physics. 2021;48(10):5593-610.
23. Uh J, Wang C, Acharya S, Krasin MJ, Hua CH. Training a deep neural network coping with diversities in abdominal and pelvic images of children and young adults for CBCT-based adaptive proton therapy. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2021;160:250-8.
24. Park S, Ye JC, editors. Unsupervised Cone-Beam Artifact Removal Using CycleGAN and Spectral Blending for Adaptive Radiotherapy. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 2020 3-7 April 2020.
25. Liu Y, Lei Y, Wang T, Fu Y, Tang X, Curran WJ, et al. CBCT-based synthetic CT generation using deep-attention cycleGAN for pancreatic adaptive radiotherapy. Medical physics. 2020;47(6):2472-83.
26. Gao L, Xie K, Wu X, Lu Z, Li C, Sun J, et al. Generating synthetic CT from low-dose cone-beam CT by using generative adversarial networks for adaptive radiotherapy. Radiation oncology (London, England). 2021;16(1):202.
27. He F, Liu T, Tao DJAiNIPS. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. 2019;32.
28. Song S, Zhong F, Qin X, Tu C, editors. Illumination harmonization with gray mean scale. Computer Graphics International Conference; 2020: Springer.
29. Pinto MS, Paolella R, Billiet T, Van Dyck P, Guns PJ, Jeurissen B, et al. Harmonization of Brain Diffusion MRI: Concepts and Methods. Frontiers in neuroscience. 2020;14:396.
30. Jiang Y, Zhang H, Zhang J, Wang Y, Lin Z, Sunkavalli K, et al., editors. Ssh: A self-supervised framework for image harmonization. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021.
31. Babier A, Mahmood R, McNiven AL, Diamant A, Chan TCY. Knowledge-based automated planning with three-dimensional generative adversarial networks. Medical physics. 2020;47(2):297-306.
32. Maspero M, Savenije Mhf Fau - Dinkla AM, Dinkla Am Fau - Seevinck PR, Seevinck Pr Fau - Intven MPW, Intven Mpw Fau - Jurgenliemk-Schulz IM, Jurgenliemk-Schulz Im Fau - Kerkmeijer LGW, et al. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. (1361-6560 (Electronic)).
33. Liu Y, Lei Y, Wang Y, Wang T, Ren L, Lin L, et al. MRI-based treatment planning for proton radiotherapy: dosimetric validation of a deep learning-based liver synthetic CT generation method. Physics in medicine and biology. 2019;64(14):145015.

Chapter 3: Generation of synthetic ground glass nodules using generative adversarial networks (GANs)

*Adapted from **Zhixiang Wang**, Zhen Zhang, Ying Feng, Lizza E. L. Hendriks, Razvan L. Miclea, Hester Gietema, Janna Schoenmaekers, Andre Dekker, Leonard Wee and Alberto Traverso. Generation of Synthetic Ground Glass Nodules Using Generative Adversarial Networks (GANs). Eur Radiol Exp 2022, 6 (1),59.https://doi.org/10.1186/s41747-022-00311-y.*

Abstract

Background: Data shortage is a common challenge in developing computer-aided diagnosis systems. We developed a generative adversarial network (GAN) model to generate synthetic lung lesions mimicking ground glass nodules (GGNs).

Methods: We used 216 computed tomography images with 340 GGNs from the Lung Image Database Consortium and Image Database Resource Initiative database. A GAN model retrieving information from the whole image and the GGN region was built. The generated samples were evaluated with visual Turing test performed by four experienced radiologists or pulmonologists. Radiomic features were compared between real and synthetic nodules. Performances were evaluated by area under the curve (AUC) at receiver operating characteristic analysis. In addition, we trained a classification model (ResNet) to investigate whether the synthetic GGNs can improve the performances algorithm and how performances changed as a function of labelled data used in training.

Results: Of 51 synthetic GGNs, 19 (37%) were classified as real by clinicians. Of 93 radiomic features, 58 (62.4%) showed no significant difference between synthetic and real GGNs ($p \geq 0.052$). The discrimination performances of physicians (AUC 0.68) and radiomics (AUC 0.66) were similar, with no-significantly different ($p = 0.23$), but clinicians achieved a better accuracy (AUC 0.74) than radiomics (AUC 0.62) ($p < 0.001$). The classification model trained on datasets with synthetic data performed better than models without the addition of synthetic data.

Conclusions: GAN has promising potential for generating GGNs. Though similar AUC, clinicians achieved better ability to diagnosis whether the data is synthetic than radiomics.

Keywords (MESH terms): Deep learning, Computed tomography, Lung, Neural networks (computer), Solitary pulmonary nodule, generative adversarial network

Key points

- We propose a technique that can generate synthetic ground glass opacities.
- Some of the generated images were assessed as real by physicians and imaging quantitative method (radiomics).

- The synthetic data can improve the performance of deep learning classification models.

Abbreviations

3D: Three-dimensional; AUC: area under the curve; CAD: Computer-aided diagnosis; CT: Computed tomography; D L: Deep learning; GGO: Ground glass opacity; GGN: Ground glass nodule; LIDC-IDRI: Lung Image Database Consortium and Image Database Resource Initiative; ROI: Region of interest; SRGAN: Super-resolution generative adversarial network; VTT: Visual Turing test.

Background

Artificial intelligence is a rapidly developing field including many applications in computer vision, such as deep learning (D L) and machine learning methods for the segmentation [1] and the classification [2] of anatomical structures and abnormalities in standard of care diagnostic imaging. A strong effort is dedicated to the implementation of these methods as computer-aided diagnosis (CAD) tools to reduce the time burden of clinical tasks and improve radiologists' detection accuracy. For lung cancer screening, the number of CAD systems to automatically identify the presence of pulmonary nodules has exponentially increased in the last 10 years. D L methods have shown an increased detection accuracy for all the types of pulmonary nodules (solid, part solid, ground glass opacities) compared to traditional machine learning methods in low-dose screening computed tomography (CT) scans [3,4].
The success of developing robust and widely applicable deep learning-based CAD systems relies on the availability of a large amount of curated and annotated data. However, annotating data consistently has a cost and is dependent on radiologists' time and availability. Even when large amount of data is collected for training D L networks, the problem of class imbalance may exist. The class imbalance problem refers to some labels (classes) being more frequent than others. Due to this unbalance, the D L network will learn better how to classify the more frequent samples, with degraded performances for the minority class(es)[5]. In the specific case of pulmonary nodule detection, ground glass nodules (GGN), although account for only 2.7% to 4.4% of all nodules, are malignant in 63% of the cases [6].
Next to classical statistical methods such as SMOTE (synthetic minority oversampling technique), researchers have investigated more advanced methods for generating synthetic samples of original data, to increase and balance the original sample size of the training dataset. Recently, generative adversarial networks (GANs) have been proposed as a method to generate synthetic images to improve the existing oversampling techniques [7]. GANs, which are DL algorithms based on game theory, have been applied to several computer vision tasks such as image denoising, reconstruction and as mentioned, synthetic data generation [8,9]. Briefly, GANs consists of two competing actors: a generator and a discriminator. They are used to generate synthetic images/samples and "judge" the quality of the generated images, respectively. The equilibrium is reached when the synthetic (*i.e.*, fake) samples cannot be distinguished from the real distribution [10].

While many studies demonstrated the potential of GANs to generate synthetic images, the generated images/samples have not been evaluated by radiologists, and this limits the acceptance and use of GANs in a clinical setting. In fact, generated images/samples should be representative of the "real" population. However, by only focusing on evaluating at the "human-level" the appropriateness of synthetic samples, it is not possible to draw any conclusion whether the introduction of synthetic samples in the training samples will improve the detection performances of CAD systems. In principle, it is expected that adding as many synthetic samples as possible to the original data will lead to a CAD system with better detection performances. It is important to notice that generating synthetic samples via GAN is in itself a learning procedure, where the original data is used to train the networks to generate the synthetic samples. The ratio between original data available and the quality of generated samples is not clear yet.

In this study, we investigated the following research questions:

i. Is it possible to use a GAN model to generate synthetic GGNs on low-dose screening CT scans that are undisguisable by clinicians from the real samples?

ii. How much labelled data is needed to generate synthetic GGNs of sufficient quality to train a CAD for pulmonary nodule detection achieving the same level of performance of a large amount of labelled data?

To answer these questions, we developed an optimised GAN model with dual discriminators to generate GGNs.

Methods

Study population

A total of 216 subjects were selected from The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) database for this study [11]. In this database, the nodules were classified into five grades by four radiologists: 1 = ground glass opacity (GGO[1]); 2 = intermediate between 2 and 3; 3 = part solid; 4 = intermediate between 4 and 5; 5 = solid. We chose 340 GGN nodules of grades 1 or 2 that were annotated by at least two radiologists for our study. To ensure data quality, further confirmation was performed by a radiologist (author Z. Z), with five years of experience in lung CT, to verify that all the nodules were GGNs.

Image preprocessing

In the preprocessing methods, first, the two-dimensional slices with annotation as GGN from the CT volume were extracted. Second, in order to avoid interference from external tissues of the lung, we first cropped the lungs from the tissue and background with a seed-filling algorithm, which starts from an inner point of the polygon area and draws points with the given grey level from inside to outside until the boundary is found. Third, the cropped images were padded by 0 in the background to keep every image having the same sizes (512 × 512) in the dataset. Fourth, we normalised the data to the range 0-1, as is the standard practice in computer vision. Fifth, we erased the nodules from the original position and saved them as region of interest (ROI) for the training set. In general, each training batch contained two images: the original image as the target image, which serves as the ground truth for the generator (as shown in Figs. 1 and 2), and another image is the input image, in which stripped the nodule area, *i.e.*, the ROI region was processed as blank for the input image. As shown in Fig. 1, the network generates the nodule from the input image. In addition, after generation, there are two discriminators (whole image discriminator and ROI discriminator) to evaluate the quality of the whole image and the ROI where the nodule is.

---

[1] GGO is defined as a type of GGNs showing a misty increase in lung attenuation without obstructing the underlying vascular markings; GGOs can also be called as “pure GGNs”, i.e. GGNs showing solely a GGO component.

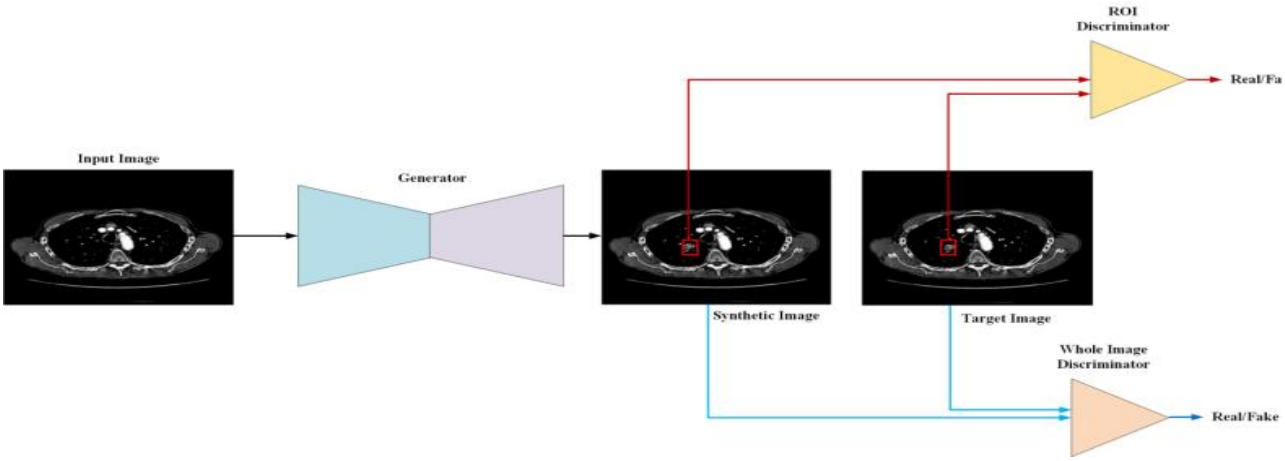


Fig. 1. The pipeline for training the model. First. the generator synthetizes ground glass nodules from the background according to the input image. Second. the region of interest (ROI) discriminator (red line) and the whole image discriminator (blue line) extract features from the ROI and whole image to classify the synthetic image and the target whether the synthetic image is real.

Construction of the DL model

The super-resolution generative adversarial network (SRGAN) was used as the backbone of the generator [12]. SRGAN compares the features difference in the model between a pair of data and train the discriminators to improve the realism of the recovered images. Both the whole image discriminator and ROI discriminator are based on a ResNet [13] which is a widely used classical classification networks combined by residual blocks with different input sizes and depths of the network. The structure of the network is shown in Fig. 2.

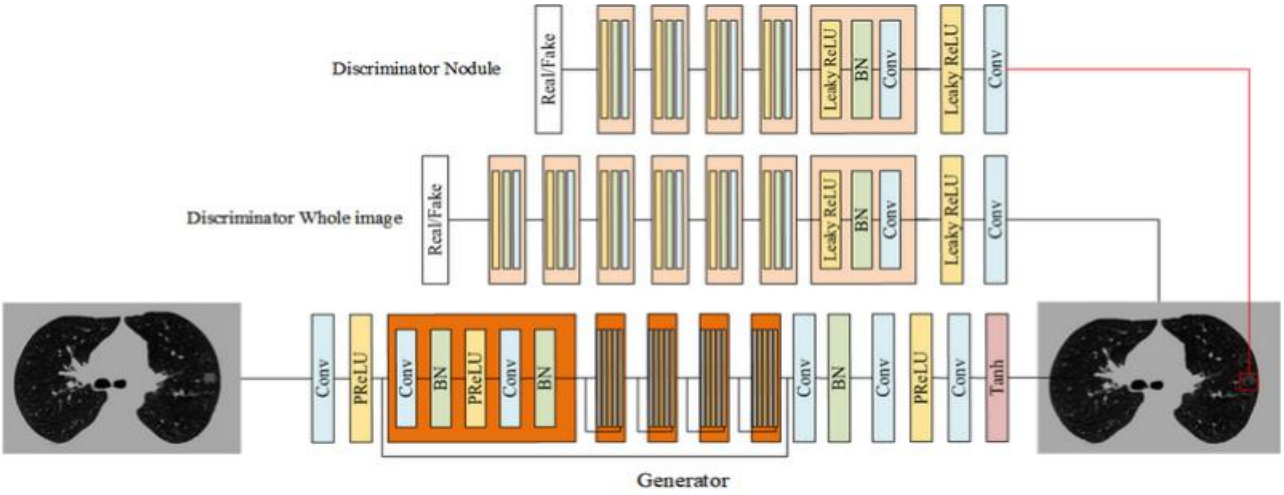


Fig. 2. The structure of the network. The generator creates the synthetic ground glass

nodule at the position where the mask in the input. The generator is composed of convolutional layers with a kernel size of 3 × 3, the batch normalisation and the "parametric rectified linear unit" (PReLU) activation function. The discriminator was composed of convolutional layers with a kernel size of 3 × 3, the batch normalisation, and the leaky PReLU activation function.

For training the network, the loss function was as follows:

$$L_{D2SRGAN} = (L_{ssim} + L_{adversarial})_{whole\ image} + (L_{ssim} + L_{adversarial})_{ROI\ image} \quad (1)$$

$$L_{adversarial} = \sum_{n=1}^{N} -logD(G(x)) \quad (2)$$

$$L_{ssim}(x, y) = 1 - \frac{(2\mu_x\mu_y + C_1) + (\sigma_{xy} + C_2)}{(\mu_x{}^2\mu_y{}^2 + C_1)(\sigma_x{}^2\sigma_y{}^2 + C_2)} \quad (3)$$

The L_{ssim} can be used to compare the similarity between two images. In this loss function, the whole image is separated into two parts to calculate the loss function respective. G and D represent the generator and discriminator, *x* is the input of the generator. μ_x and μ_y represent the average of input *x*,*y* respectively. σ_x and σ_y represent the standard deviation of input *x*,*y* respectively. σ_{xy} is the covariance of *x* and *y*. C1 and C2 are constants to avoid system errors caused by the denominator being zero.
All images were loaded with an unchanged original size of 512 × 512. The input size of the discriminator for the whole image and the ROI image were 512 × 512 and 32 × 32, respectively. An Adam optimizer was used to train both the generator and the discriminator with a learning rate of 0.0001. This model was trained using an NVIDIA Tesla V100 SXM2 32 GB graphics processing unit.

Evaluation of model performance

We evaluated the model performance using both subjective (visual Turing test, VTT) and objective (radiomics) approaches. VTT is an assessment method that evaluation the ability of a human or doctors to identify attributes and relationships from images [14]. Subjective evaluations were performed by two radiologists (authors R.M. and H.G.) and two pulmonologists (authors L.H. and

J.S.),, who all had more than five years of experience in lung CT imaging and on a daily basis evaluate chest CT scans. One hundred images (50 real and 50 synthetic GGNs) were divided into four batches and converted to a DICOM (Digital Imaging and COmmunications in Medicine) file with 25 slices of images, and each physician was randomly assigned to one of these batches. The physicians categorised the real and synthetic GGNs into four classes based on this categorical scale: confidently fake; leaning fake; leaning real; and confidently real.

To perform an objective evaluation, radiomic features were calculated from the original and generated data. Radiomics refers to the extraction of quantitative information from medical images by computing the statistical, morphological and texture features. The following feature categories were extracted using the open source Pyradiomics package (version 3.0.1) with default values: first order statistics (n = 18); grey level cooccurence matrix (n = 24); grey level dependence matrix (n = 14); grey level run length matrix (n = 16), grey level size zone matrix (n = 16); and neighbouring grey tone difference matrix (n = 5) [15-17].

The Kolmogorov-Smirnov test was used for the analysis of whether the distribution of radiomics features were similar between the real and synthetic GGNs. We considered significant p values lower than 0.05.

The results of the subjective and objective evaluations were analysed using the area under the curve (AUC) at receiver operating characteristic analysis. For the subjective evaluation, we took into consideration the VTT results. For the objective evaluation, to compare the classification ability of radiomics and radiologist, a logistic regression model was build based on radiomic features to classify both real and synthetic GGNs. The same dataset was used for the physician evaluations and the radiomics logistic regression model, with a 4-fold cross-validation.

In addition, we also investigated whether the synthetic GGNs can improve the performance of a CAD algorithm trained for recognizing GGNs from all types of nodules in the LIDC-IDRI dataset and how the performance changed as a function of labelled data used in the training.

As a CAD, we used a basic ResNet as the D L classification network with a cross-entropy loss function. First, we separated the dataset into 10 training subsets and an independent test set. We trained the classification network on portions of the original data ranging from 10% to 100% of the real data and we separately inferred on the test set. Then, we trained the classification network on the original data added systematic data generated by the GAN network trained in 10% to 100% of real data.

Results

Examples of synthetic GGNs generated in different parts of the lungs with different surrounding tissues are shown in Fig. 3. Nodules classified as fake (Fig. 3b) show more unnatural characteristics in terms of intensity and morphology than nodules classified as “real” (Fig. 3a), specifically, "fake nodules" have very high fixed gray values and regular shapes such as rectangles.

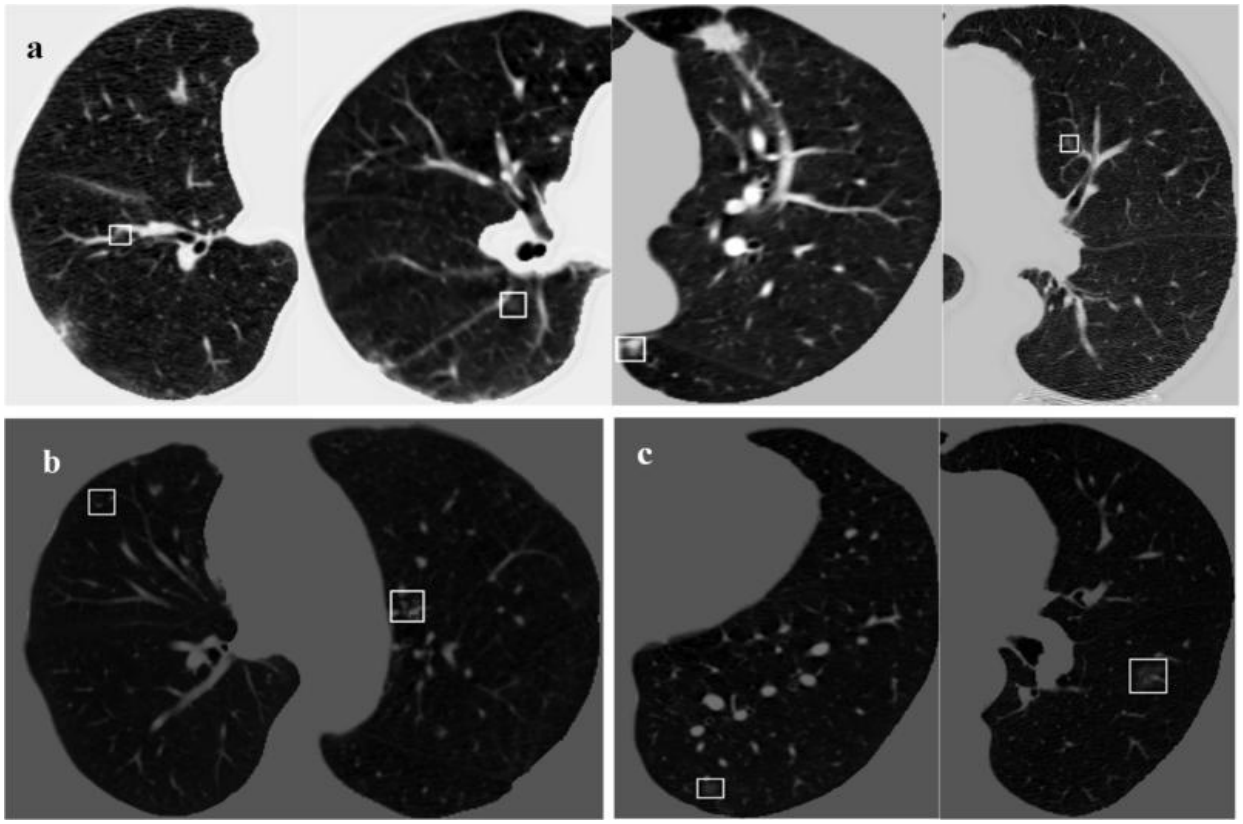


Fig. 3. Examples of synthetic ground glass nodules (GGNs), the GGNs were categorized by physicians to four categories: confidently fake; leaning fake; leaning real; and confidently real. a Synthetic GGNs classified as “real” by clinicians. b Synthetic GGNs with less convincing generated lesions (classified as “leaning fake”). c A real GGNs in the original LIDC-IDRI dataset.

VTT results

Fig. 4 presents the combination of the classification results for the four clinicians. : Of 51 synthetic GGNs, 19 (37%) were classified as real by clinicians; 8/51 (16%) were classified as confidently real and 11/51 (22%) were classified as leaning real.

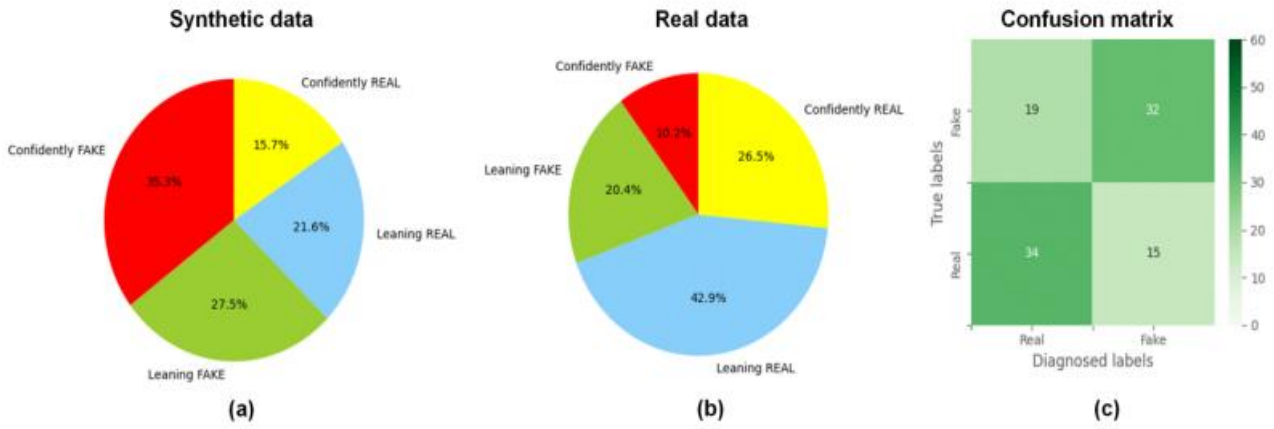


Fig. 4. Visual Turing test results. a, b Prediction distribution in synthetic and real ground glass nodules. c Confusion matrix for the prediction.

Radiomics

Of a total of 93 features, 58 (62.4%) showed no significant difference ($p \geq 0.052$) between synthetic and real GGNs, and the detailed results are provided in Table 1. Fig. 5a shows the comparison of the distribution of radiomic features between real and synthetic GGNs, the histogram shows the counts of specific feature values, and the differences (p-values) in the extracted radiomic features between real and synthetic GGNs were calculated. The receiver operating characteristic curves constructed based on the results of VVT by clinicians and logistic regression model developed by radiomics features are shown in Fig. 5b. We observed a similar classification performance of clinicians (0.68) and radiomics (0.66), with no-significantly different (p=0.23). However, the clinicians achieve significant great performance accuracy around 0.74, better than the 0.62 radiomics accuracy ($p < 0.001$). The clinicians achieves better ability to diagnosis whether the data is synthetic than radiomics.

Table 1. Comparison between real and deep learning-generated radiomic features (*p*-values according to the Kolmogorov-Smirnov test)

Class	Feature name	*p*-value
Grey level co-occurrence matrix (GLCM)	Inverse difference moment	0.984025
Grey level size zone matrix (GLSZM)	zone percentage	0.934856
Grey level dependence matrix (GLDM)	Small dependence emphasis	0.932657
Grey level co-occurrence matrix (GLCM)	Inverse difference	0.926064
First order	Robust mean absolute deviation	0.903346
GLSZM	Small area low grey level emphasis	0.860311
Grey level run length matrix (GLRLM)	Run percentage	0.827381
GLRLM	high grey level run emphasis	0.729491
GLSZM	Grey level non-uniformity normalised	0.696774
GLRLM	Long run emphasis	0.676057
GLCM	Sum entropy	0.658063
GLRLM	Long run high grey level emphasis	0.652292
GLRLM	Run entropy	0.652292
First order	Entropy	0.643479
GLCM	Inverse variance	0.616719
GLRLM	Short run high grey level emphasis	0.582172
GLDM	high grey level emphasis	0.574195
GLCM	Joint energy	0.570327
GLCM	Joint entropy	0.570327
GLRLM	Run length non-uniformity normalised	0.570327
GLRLM	Short run emphasis	0.570327
First order	90 percentile	0.541180
GLDM	Small dependence low grey level emphasis	0.512551
First order	Interquartile range	0.498064
GLCM	Inverse difference normalised	0.456086
GLDM	Large dependence emphasis	0.450880

GLDM	dependence variance	0.445137
GLSZM	Low grey level zone emphasis	0.445137
First order	Mean absolute deviation	0.414534
GLCM	Autocorrelation	0.407415
GLDM	Dependence non-uniformity normalised	0.403944
First order	Mean	0.389392
GLRLM	Run variance	0.375333
GLRLM	Grey level non-uniformity normalised	0.324190
GLCM	Maximum probability	0.307686
Neighbouring grey tone difference matrix (NGTDM)	Strength	0.272504
GLCM	Cluster tendency	0.267111
GLCM	Inverse difference moment normalised	0.264157
GLDM	dependence entropy	0.261878
GLRLM	Short run low grey level emphasis	0.227646
First order	Minimum	0.212067
GLSZM	Large area high grey level emphasis	0.202291
First order	Root mean squared	0.186989
GLSZM	Large area emphasis	0.178996
GLDM	Grey level variance	0.170028
GLCM	Joint average	0,160908
GLCM	Sum average	0,160908
First order	uniformity	0,133892
GLDM	Small dependence high grey level emphasis	0,124894
GLSZM	Zone variance	0,119210
First order	Variance	0,108119
GLCM	Sum squares	0,108119
GLSZM	High grey level zone emphasis	0,105973
GLDM	Large dependence low grey level emphasis	0.082337
GLSZM	Size zone non-uniformity normalised	0.074667
GLSZM	Small area emphasis	0.073186

GLSZM	Large area low grey level emphasis	0.069577
GLRLM	Grey level variance	0.066007
GLCM	Informational measure of correlation 2	0.052283
GLRLM	Low grey level run emphasis	0.045409
GLSZM	Small area high grey level emphasis	0.044462
GLCM	Cluster prominence	0.022046
GLSZM	Grey level variance	0.021275
NGTDM	Contrast	0.020502
First order	10^{th} percentile	0.015568
GLDM	Low grey level emphasis	0.014150
GLCM	Difference entropy	0.011605
GLSZM	Zone entropy	0.010051
GLRLM	Long run low grey level emphasis	0.008825
GLCM	Informational measure of correlation 1	0.006491
GLCM	Difference average	0.005938
GLCM	Maximal correlation coefficient	0.005586
GLDM	Large dependence high grey level emphasis	0.003520
First order	Maximum	0.002755
GLCM	Cluster shade	0.002638
First order	Range	0.001136
First order	Median	0.000355
GLCM	Contrast	0.000251
GLDM	Dependence non-uniformity	0.000230
GLSZM	Size zone non-uniformity	7.60E-05
NGTDM	Busyness	6.60E-05
GLCM	Correlation	2.40E-05
GLSZM	Grey level non-uniformity	1.40E-05
NGTDM	Complexity	1.40E-05
GLCM	Difference variance	5.00E-06
NGTDM	Coarseness	0.000000
First order	Skewness	0.000000
First order	Energy	0.000000
First order	Total energy	0.000000

First order	Kurtosis	0.000000
GLRLM	Run length non-uniformity	0.000000
GLDM	Grey level non-uniformity	0.000000
GLRLM	Grey level non-uniformity	0.000000

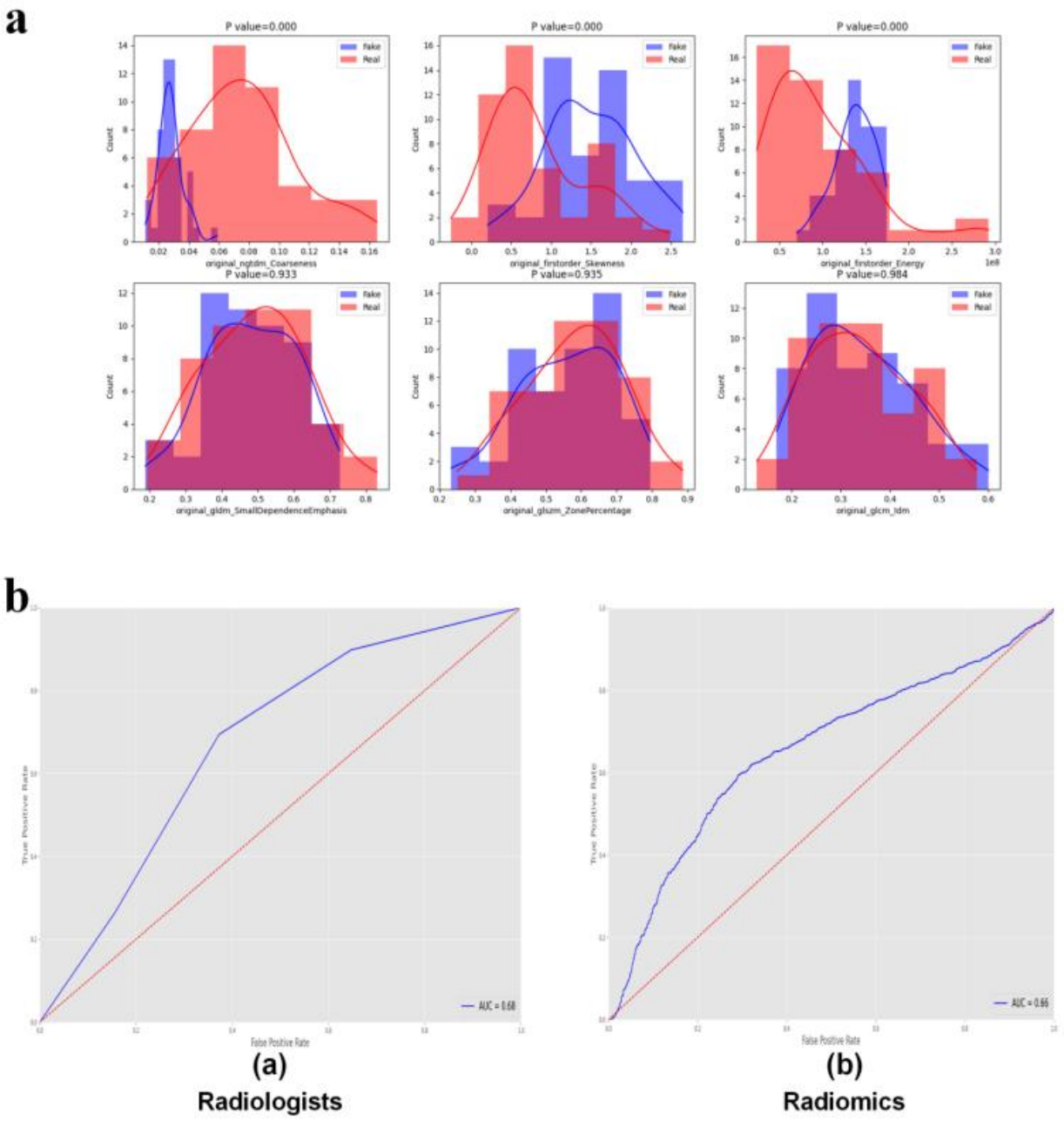


Fig. 5. a Examples for the comparison of radiomics features distribution between real and fake ground glass nodules (GGNs). The comparison of radiomics features distribution extracted from synthetic and real images with minimum three p values shows in the upper row. The comparison of radiomics features distribution extracted from synthetic and real

images with maximum three P values shows in the lower row
b. c. Receiver operating characteristic curve of the prediction of distinguishing real and fake GGNs. by radiologists (a) and by the logistic regression model (b).

DL classification network

The results of the D L classification network trained using decreasing portions of the dataset are shown in Fig. 6. When the dataset is 90%, the precision (*i.e.*, positive predictive value) was similar between the two groups. However, when the dataset decreased to 50%, the performance of the real data only group significantly decreased. On the other hand, synthetic GGNs can increase precision in training the DL network. When the sample decreased to 10%, the real data has better performance than by adding synthetic data. From Fig. 6b, the recall (*i.e.*, sensitivity) of GGN was decreasing when decreasing the dataset both in real data only and real data with GAN groups. However, in most cases, models trained on datasets with synthetic data performed better than models without the addition of synthetic data..

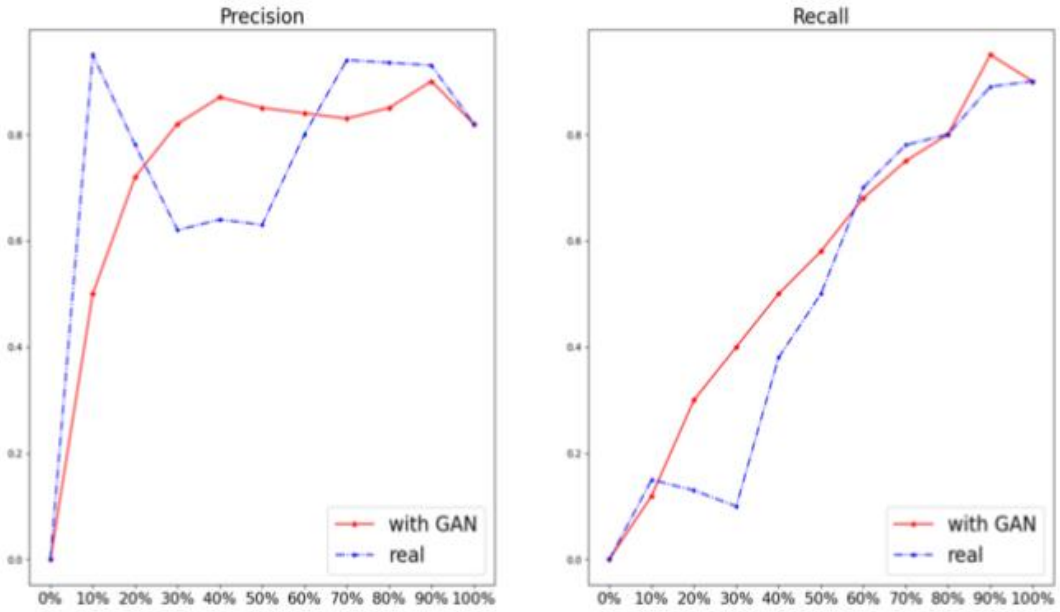


Fig. 6. Comparison precision (i.e.. positive predictive value) and recall (i.e.. sensitivity) between real and added synthetic dataset in different percentages of the training set. The blue and the red lines present the performance of the deep learning classification model trained by real data and the real data plus synthetic data, respectively. The horizontal axis label is the percentage of training data in the dataset. The vertical axis label is the score of precision and the recall with the range from 0 to 1.

Discussion

In the present study, we applied a GAN-based model with double discriminators to generate GGN in low-dose CT scans. We benchmarked the performance of the model using a qualitative (VTT with clinicians) and a quantitative approach (radiomics).

To our knowledge, only one previous study proposed the use of GANs to generate lung lesions and performed a VTT [18], which showed that 67% and 100% of the fake nodules were marked as real by two radiologists, respectively. Differences exist between this study and our study: in the VVT of the cited study [16], the radiologists reviewed the generated lesions, but the surrounding tissues or the entire lungs were not included in the field of view. Moreover, the surrounding tissues and the lung background that has relationship with nodules were not considered when training and generating the nodules. Conversely, we generated GGNs from the whole lung to use the anatomical dependence with the background tissue [19]. However, the relatively small size of our study compared to the previous research [18] probably influenced the results of the visual Turing test.

Based on our VTT evaluation, we have shown that GAN-generated lung lesions have the potential to be very consistent with real lesions. This gives us the opportunity to use GAN-generated data to solve real-world problems, such as using the generated data to train and test junior doctors, especially for hospitals that do not have large cohort datasets, long-time established picture archiving and communication systems, as privacy-preserving synthetic open datasets for research purposes.

More than half of the radiomic features were not statistically different between D L-generated and real nodules, proving that the generated GGNs are acquiring or learning detailed features from the real sample. Furthermore, these consistent radiomic features cover all classes, which could support the conclusion that the proposed approach mimics different aspects of real nodules. Conversely, one-third of the features in this study showed significant differences in the distribution between the generated and real GGNs. Based on the radiomics results and the clinicians' opinion, we think that the low complexity of the generated GGNs is the main reason for the discrepancy between the generated and real GGNs. For example, the *p*-value of the radiomic features *coarseness* (which can measure the spatial change rate) and *complexity* (which can measure the non-uniformity of local grey levels) between real and synthetic GGNs, are close to 0, supporting our hypothesis. We hypothesize the following explanations: i) the data source is derived from public databases that

have low resolution and lots of noise; ii) we did not optimise the training process by specifically including radiomics features in the loss function.
Based on the radiomics results, we built a "radiomics physician" to discriminate between real and generated GGNs, which interestingly is generally consistent with the discriminatory ability of real physicians. It is worth noting that the "radiomics physician" model was trained based on a sample of 100 cases, and the physicians have more than five years of experience. Overall, it is a challenging task to discriminate between real and generated GGNs for "radiomics physicians" and real physicians.
Finally, we wanted to test how data augmentation with GAN will affect the detection accuracy of a CAD system. Fig. 6 shows that adding synthetic GGNs to the original dataset improves the performance of our D L CAD system. However, there was no significant contribution when the size of the training dataset is under 10% and over 70% of the original sample size. We hypothesise that when the training data is under 10%, there is an insufficient number of samples to train the GAN. A GAN trained on only a few samples cannot synthesise the rich diversity and complexity of real GGNs. Based on the results (Fig. 6), we conclude that the performance of the D L model increases with the sample size in certain ranges of real data samples. However, as shown in Fig. 6, the performance of the D L model cannot be improved after a threshold value larger than the sample size, which is the plateau of the model. Specifically, for effective dataset size to train a GAN, around 50% of training data which include around 100 samples of GGN has the biggest increase in accuracy of the classification model when synthetic GGN are added. Overall, from our experiment, we found that:

i. synthetic data has the ability to increase the performance of a DL model unless only a few training samples can be used;
ii. from the perspective of cost and effectiveness, around 100 samples are sufficient to develop a GAN model that can generate realistic GGNs to significant improve the performance of the detection GGN model.

This study has some limitations. First, we used a public dataset for training the model, but we want to extend the work to other datasets. In future studies, we will add high-resolution data from our centre for model enhancement. Second, we only focused on GGNs, because of their lower incidence compared to other types of nodules. However, the dimension and density variation of the included GGNs is limited, which has the potential risk of obtaining optimistic radiomic assessment results. We will perform transfer learning to generate lung nodules

and tumours in the future based on the model in this study. Furthermore, the diagnosis of malignant GGN is a challenging task for clinical practice. However, in this study, we did not generate benign or malignant GGN. To address this issue, we are collecting data from the real world with follow-up endpoints and trying to generate qualitative GGN, especially malignant GGN.

Third, we generated only two-dimensional samples, However, generating three-dimensional (3D) images is costly for model training, first, because 3D GANs have a larger number of parameters which need more training data and also have a significantly higher requirement in hardware when the input data has large scale such as CT images. In the future work we will consider the model compression to decrease the requirement of hardware and the size of dataset for training the 3D GAN. We tried to perform our visual Turing tests by getting closer as much as possible to a real clinical scenario. Nevertheless, it was out of the scope of this study to integrate our D L models within the clinical workstations available to our radiologists. As proof-of-concept, we proposed to our radiologists the generated and real pulmonary nodules as two-dimensional axial CT images in the standard lung window. Future work will include the production of the generated nodules in standard DICOM formats in all the 3D projections. We are also investigating the possibility to invest in the development of a cloud-based platform to homogenise visual Turing tests for similar experiments. In addition, we did not evaluate the morphological features between the generated and real GGNs.

Fourth, we have not discussed the trend of data requirement for different task, such as what happens when the quality of data is decreased, how many data points need to be added when the target size us increased and whether different sources such as CT and magnetic resonance imaging influence the dataset requirements. In the future work, we will design experiments to figure out the connection between the data requirement and different tasks.

Fifth, according to the results of the radiomics part, there are still considerable differences between the real and generated GGO, and more than one-third of the radiomic feature values were different, which may be a reflection that the GAN method proposed in this study is not optimal. Based on this result, there is still much potential for improvement of our algorithm, with a particular focus on improving the level of complexity of the textures.

Sixth, we did not conduct interobserver and intraobserver testing and the degree of disagreement between different readers was not assessed. On the other hand, in our experience, the differences between the readers (physicians) included in this study were limited to the same broad category, i.e., real or fake.

For example, nodules labelled as “confidently real” by one physician have the possibility of being labelled as “leaning real” instead of “confidently/leaning fake” by other physicians.
Finally, despite the GANs are an elegant data generation mechanism gaining more and more popularity in the medical field, most of them still present a high level of complexity compared for example to traditional D L algorithms such as convolutional neural networks. For example, there is no consensus on the most appropriate metric to be used to stop the training at the best point (global minimum of the loss function). This will sometimes lead to a not satisfactory quality of the generated data. Especially when dealing with medical images, the risk of introducing novel, undesired artifacts and blurry the images is not negligible.
In conclusion, in this study, we used GANs to generate GGN and validated these by four physicians and radiomics approaches, showing that GAN methods have great potential for augmentation of the original dataset.

References

1. Zhou XR (2020) Automatic segmentation of multiple organs on 3D CT images by using deep learning approaches. Adv Exp Med Biol 1213:135–147. doi:10.1007/978-3-030-33128-3_9
2. Mastouri R, Khlifa N, Neji H, Hantous-Zannad S (2020) Deep learning-based CAD schemes for the detection and classification of lung nodules from CT images: a survey. J Xray Sci Technol 28:591–617. doi:10.3233/XST-200660
3. Setio AAA, Traverso A, de Bel T, et al (2017) Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Medi Image Analysis 42:1–13. doi:10.1016/j.media.2017.06.015
4. Kaggle Data Science Bowl (2017). https://www.kaggle.com/c/data-science-bowl-2017.
5. Bowles C, Chen L, Guerrero R, et al (2018) Gan augmentation: augmenting training data using generative adversarial networks. .arXiv preprint arXiv:1810.10863, 2018.
6. Migliore M, Fornito M, Palazzolo M, et al (2018) Ground glass opacities management in the lung cancer screening era. Ann Transl Med. 2018 Mar;6(5):90. doi: 10.21037/atm.2017.07.28. PMID: 29666813; PMCID: PMC5890046.
7. Zhang H, Hu X, Ma D, Wang R, Xie X (2020) Insufficient data generative model for pipeline network leak detection using generative adversarial networks. IEEE Trans Cybern. 2022 Jul;52(7):7107-7120. doi: 10.1109/TCYB.2020.3035518. Epub 2022 Jul 4. PMID: 33296325.
8. Bera S, Biswas PK (2021) Noise Conscious Training of Non Local Neural Network Powered by Self Attentive Spectral Normalized Markovian Patch GAN for Low Dose CT Denoising. IEEE transactions on medical imaging 40 (12):3663-3673. doi:10.1109/tmi.2021.3094525
9. Do WJ, Seo S, Han Y, Ye JC, Choi SH, Park SH (2020) Reconstruction of multicontrast MR images through deep learning. Medical physics 47 (3):983-997. doi:10.1002/mp.14006
10. Jiang Y, Chen H, Loew M, Ko H (2021) COVID-19 CT Image Synthesis With a Conditional Generative Adversarial Network. IEEE journal of biomedical and health informatics 25 (2):441-452. doi:10.1109/jbhi.2020.3042523
11. Armato SG 3rd, McLennan G, Bidaut L, et al (2011) The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics 38 (2):915-931. doi:10.1118/1.3528204

12. Ledig C, Theis L, Huszár F, et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. pp 4681-4690
13. He K, Zhang X, Ren S, Sun J Deep (2016) residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770-778
14. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen XJA inips (2016) Improved techniques for training gans. arXiv:1606.03498v1
15. de Farias EC, di Noia C, Han C, Sala E, Castelli M, Rundo L (2021) Impact of GAN-based lesion-focused medical image super-resolution on the robustness of radiomic features. Scientific reports 11 (1):21361. doi:10.1038/s41598-021-00898-z
16. Lambin P, Leijenaar RTH, Deist TM, et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nature reviews Clinical oncology 14 (12):749-762. doi:10.1038/nrclinonc.2017.141
17. Tixier F, Jaouen V, Hognon C, Gallinato O, Colin T, Visvikis D (2021) Evaluation of conventional and deep learning based image harmonization methods in radiomics studies. Physics in medicine and biology 66 (24). doi:10.1088/1361-6560/ac39e5
18. Chuquicusma MJ, Hussein S, Burt J, Bagci U How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), 2018. IEEE, pp 240-244
19. Xu Z, Wang X, Shin H-C, Roth H, Yang D, Milletari F, Zhang L, Xu D Tunable CT Lung Nodule Synthesis Conditioned on Background Image and Semantic Features. In, Cham, 2019. Simulation and Synthesis in Medical Imaging. Springer International Publishing, pp 62-70

Chapter 4: CycleGAN Clinical Image Augmentation Based on Mask Self-attention Mechanism

Adapted from Junzhuo Liu, Zhixiang Wang*, Ye Zhang, Alberto Traverso, Andre Dekker, Zhen Zhang and Qiaosong Chen. CycleGAN Clinical Image Augmentation Based on Mask Self-Attention Mechanism. IEEE Access 2022, 10, 105942–105953. https://doi.org/10.1109/ACCESS.2022.3211670.*

** indicates equal contributions*

Abstract

With the development of society and the advancement of science and technology, artificial intelligence has also emerged as the times require. In computer vision, deep learning based on convolutional neural networks(CNN) achieves state-of-the-art performance. However, the massive data requirements of deep learning have long been a pain point in the field, especially in the medical field, where it is often difficult (and sometimes impossible) to obtain enough training data for some specific tasks. To overcome insufficient and unbalanced data, in this paper, we focus on the generation and balance of data on radiation- induced pneumonia, an extremely rare disease with a low incidence. As a result, datasets on this disease are extremely sparse and unevenly distributed. To address the above problems, the predecessors' method is often to use generative models to generate data as a complement of the fewer samples to achieve a balanced distribution of data samples. Among various generative models, CycleGAN is widely used in medical image generation due to its cycle consistency to achieve style migration without changing the basic content. However, the original CycleGAN method has many shortcomings, especially in Few-shot and the data unevenly distributed, its performance will be greatly reduced. To make the generated data samples retain the original structure and have finer and clearer details, this paper proposes a mask-based self-attention CycleGAN data augmentation method. A self-attention branch is added to the generator and two different loss functions named Self-Attention Loss and Mask Loss are designed. To stabilize the training process, spectral normalization is introduced to improve the discriminator and MS-SSIM and L1 joint loss are used to improve the original identity loss. The ResNet18 is used to complete classification experiments on the radiation-induced pneumonia dataset and the COVID-19 dataset respectively. Four classification performance indicators: the area under the ROC curve (AUC), Accuracy (ACC), Sensitivity (SEN), and Specificity (SPE) are calculated to verify the effectiveness and generalization of our method. Compared with the original CycleGAN and traditional data augmentation, the classifier trained by data augmentation using our method has outstanding performance in multiple classification indicators and has better classification performance. Experimental results show that our method solves the problem of insufficient samples and data imbalance

in the pneumonia dataset by generating high-quality pneumonia images. Code is available at https://github.com/ngfufdrdh/CycleGAN-lung.

INDEX TERMS cycle generative adversarial networks, medical data augmentation, deep learning

Introduction

IIn the process of modern medical diagnosis, medical experts often use medical images to assist diagnosis, such as CT, MRI, etc. Medical images are used to reflect the internal structure of the patient's body, assist doctors to determine the possible lesions of the patient, and greatly improve the efficiency, accuracy, and reliability of clinical diagnosis. At the same time, medical images comprehensively display the subtle structure of the human body, which helps to assist doctors in detecting early lesions. Deep learning can promote the development of medical image-assisted diagnosis[1, 2]. However, there are still multiple challenges in applying deep learning to medical image-assisted diagnosis. One of the biggest challenges is the collection of medical data. In supervised learning, the training of deep learning models requires a large number of data samples, such as the CheX- pert dataset[3] contains 224,316 chest radiographs, and the fastMRI dataset[4] consists of 167,375 slices. However, the probability of side effects for most treatments is relatively low. For example, radiation pneumonitis (RP) is common radiotherapy toxicity, which accounts for only 8-25% of pa- tients receiving radiotherapy[5]. The limited positive sample size (patients suffering from RP) is an obstacle for deep learn- ing methods. In previous studies on predicting RP, sample sizes ranged from 100 to 400, with positive samples ranging from 20 to 150[5]. This reveals a paradoxical point that currently exists across disciplines, where there is a strong need for deep learning in the medical field and great difficulty in collecting data available for model building, especially for low-incidence samples.

Using traditional data augmentation methods, such as translation, rotation, cropping, denoising (or adding noise), color change, etc., can alleviate this problem, and to some extent, it can also increase the generalization ability of deep learning models. [6] proposes a cervical cancer prediction model (CCPM), which innovatively combines the outlier detection methods density-based spatial clustering of ap- plications with noise(DBSCAN) and iForest. For the data imbalance problem, the data oversampling methods SMOTE and SMOTETomek balanced data are used. Moreover, the classifier uses random forest for predicting cervical cancer based on risk factors to improve the prediction performance. There is also research about traditional methods in [7, 8]. However, the shortcomings of traditional data augmentation are also obvious in some fields. The changes to the original data samples are not always effective. Sometimes it will bring side effects, that is, it will interfere with the

normal training process, and in most cases, the performance improvement of the model is limited. This is a common issue in medical image processing.

To overcome this challenge, in recent years, GANs[9] have been widely used in the augmentation of medical image data[10–12], by generating medical images through random noise to increase data samples. Medical images generated by GANs can be used as a complement to the original data to improve the performance of deep learning models. The

function of G is to receive random noise z and generate data samples like real data samples. The D has access to the real and synthetic data instances and tries to tell the difference be- tween them. In this way, a dynamic game process is achieved. However, for medical image augmentation of rare diseases, due to the scarcity of training data, GANs have shortcomings such as non-convergence, gradient disappearance, training crash, instability, and Uncontrollability[13]. CycleGAN[14] is often used to solve this problem. CycleGAN is composed of two generators and discriminators. Based on ensuring the consistency of structure and content, it realizes the mutual conversion between the two image domains instead of gener- ating them from noise. Since the generated images are trans- formed from other images, images generated by CycleGAN tend to have better quality and detail in the case of few-shot, which is significant for medical clinical diagnosis. However, in some cases, CycleGAN has difficulty in distinguishing the region of interest from the background of images, often caus- ing unnecessary over-transfer to the background, and there is still instability during the training process[15]. This may corrupt previous images and reverse the effects in subsequent downstream tasks. At worst, it will mislead model training with disastrous consequences.

To overcome the shortcomings of CycleGAN while retaining its advantages, in this paper, a mask-based self-attention CycleGAN data augmentation method is proposed. Compare with the original CycleGAN, a mask self-attention module is designed to make the model notice potential lesions image regions faster and quickly distinguish other regions. Two dif- ferent loss functions named Attention Loss and Mask Loss, based on the masked self-attention module, are designed to guide the generation of images. At the same time, MS-SSIM and L1 joint loss is utilized to improve CycleGAN Identity Loss and introduce Spectral Normalization to the discriminator to stabilize its training process.

Our contributions can be summarized as follows:

1.Unlike the simple fusion of the self-attentive mechanism in the CycleGAN backbone network adopted in a large number of previous research, our work takes into full consideration the medical facts and the shortcomings of the original CycleGAN. We creatively design an attention branching module and two loss functions for CycleGAN using the lung image Mask and the self-attention mechanism. Our method is more applicable to the generation of pneumonia images in few-shot scenarios.

2.Radiation-induced pneumonia, a rare disease, has been associated with few studies. Sample sparsity and category imbalance are prevalent in studies about the prediction of radiation pneumonia. Our method is utilized to convert from negative pneumonia lung slices to realistic positive pneumonia lung slices, which is positive for solving the above problems and applying artificial intelligence to a wide range of studies on radiation-induced pneumonia.

3.Our method is quantitatively compared with traditional data augmentation and the original CycleGAN method on the RP dataset and the COVID-19 dataset. Experimental results show that our method outperforms both.

The structure of this paper is shown as follows:

The second section introduces the existing work related to this paper and its shortcomings while pointing out the purpose of our method. Section III presents the details of the proposed method, including the network model architecture design, the loss function, the evaluation methods for the effectiveness of the method, and the training details. Section IV presents the dataset. Section V presents our experimental results. Section VI presents the discussion and limitations of our method. The last part is the conclusion.

Methods

RELATED WORK

A. CYCLEGAN IN MEDICAL IMAGE DATA AUGMENTATION

GANs are the most widely used image generation model, composed of a generator and a discriminator. The generator learns the characteristics of the image and tries to fit the noise distribution to the real data distribution. The discriminator distinguishes whether the input data is real or fake data generated by the generator. As the two networks are trained alternately, the generated data of the generator will gradually approach the real data and the discriminator will reach a balance with the generator. However, GAN suffers from pattern collapse during training and often generates poor quality images, especially in the case of complex images. These problems are related to random noise as model input. CycleGAN, following the idea of GAN, solves the above problems by taking the image from another domain instead of random noise as input. It consists of two generators and two discriminators, to realize interconversion between two image domains. CycleGAN employs cyclic consistency to ensure that the basic structure of the input image can be preserved when transforming image domains into one another, thereby avoiding excessive changes. CycleGAN has shown superiority in a variety of computer vision applications, including image denoising[16, 17], image defogging[18–21], and super-resolution[22–24]. In addition, CycleGAN is widely used in the field of medical image data augmentation due to its unique model structure and excellent image generation. Daniel[25]et al. used several CycleGAN networks to generate virus samples from x-ray image samples that do not contain COVID-19, improving the classification performance. In a similar vein, Tatiana[26] et al. Xu[27] et al. proposed a semi-supervised attention-guided CycleGAN to generate realistic tumor MRI images by adding tumor lesions from the original normal images, which improved the performance of the ResNet18 based MRI image classification model. Thomas[28] et al. used CycleGAN to eliminate staining variants from histopathological images, which improved the segmentation performance and robustness of the tissue segmentation algorithm compared to the traditional staining transform. Tmenova[29] et al. considered both the complex physiology of arteries and the vascular texture and used CycleGAN in the generation of vascular maps as a means of data augmentation in learning tasks.

B. ATTENTION

The attention mechanism was first used in natural language processing, achieving superior in machine translation tasks. A slew of recent studies had attempted to integrate CycleGAN with attentional mechanisms for image-to-image transformation. The goal of these studies is to improve the model's performance by capturing the important regions and features in the image using attention's perceptual role. Mejjati[15] et al. used an unsupervised attention mechanism to solve the problem of difficulty in focusing attention on individual objects without altering the background in image transformation, resulting in more realistic images than other methods. Liu[30] et al. introduced multi-scale spatial atten- tion and channel attention to improve CycleGAN in spatial and channel dimensions for synthesizing high-quality remote sensing images, which improved the performance of aircraft detection models in remote sensing images. [31] proposed Augmented CycleGAN network, which achieves transferring the makeup style between two low-resolution images by identifying salient pixel regions on low-resolution images in multiple scales, and using channel attention to determine the most effective attention map.
Numerous studies have demonstrated that attentional mechanisms have an excellent ability to capture critical information. Therefore, the attention mechanism is used in this paper for enabling the model to capture critical characteristics of medical images more quickly and guiding the generation of more detailed images.

PROPOSED METHOD

A. MASK AND SELF-ATTENTION

The original image binary mask and self-attention mechanism are used to construct the CycleGAN branch module. Its function is to make the model pay attention to the key image areas, distinguish other areas quickly, and reduce the interference of the background area to the model in the early stage of training. The mask is made by generating a binary map of the subject region from the input image, where the lung slice part takes a value of 1, and the rest of the position is set to 0. The attention mechanism adopts a normalization-based attention module, the Normalization-based Attention Module (NAM)[32], which is different from other attention mechanisms because it does not require additional calculations and parameters in convolution and full connection. It applies a weight sparsity penalty to attention modules, making them more computationally efficient while retaining similar performance,

thereby suppressing less salient features. It adopts a scaling factor from batch normalization (BN) as shown in equation (1). The scaling factor measures the variance of channels and indicates their importance.

$$B_{out} = BN(B_{in}) = \gamma \frac{B_{in} - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} + \beta \tag{1}$$

$$\text{Loss} = \sum_{(x,y)} l(f(x, W), y) + p\sum g(\gamma) + p\sum g(\lambda) \tag{2}$$

In equation (1), μB and σB are the mean and standard deviation of the mini batch, respectively; γ and β are trainable affine transformation parameters (scale and shift)[33]. To suppress the less salient weights, it adds a regularization term into the loss function, as shown in equation (2)[34], where x denotes the input; y is the output; w represents network weights; l (.) is the loss function; g (.) is the l1 norm penalty function; p is the penalty that balances g (γ) and g (λ). The NAM attention mechanism was added into the middle part of the entire attention module. We performed simple feature extraction on the mask part first, then passed the feature map to the NAM module, and upsampled it back to the original input image size again. Using NAM can significantly improve the image conversion efficiency and stabilize the training process.

B. SPECTRAL NORMALIZATION

Mode collapse and training instability are common problems in the training process of GANs. WGAN[13] uses weight clipping to make the discriminator satisfy the Lipschitz constraint, and proposes Wasserstein loss instead of KullbackLeibler divergence and Jensen-Shannon divergence to measure the distance between distributions. However, there are still problems in the way of weight clipping in WGAN. Therefore, WGAN-GP[35] proposes the gradient penalty to satisfy the Lipschitz constraint, solves the problem of gradient explosion or disappearance caused by weight clipping in WGAN, and improves the feature expression ability of the model, but the training process is still unstable. SNGAN[36] proposes Spectral Normalization to normalize the discriminator parameters, which improves the training stability of the

discriminator. In this study, Spectral Normalization is also used to improve the original CycleGAN discriminator to stabilize the training process.

C. NETWORK ARCHITECTURE

Figure 1 shows the main pipeline of our method, which builds on the original CyeleGAN. It has been improved to be more suitable for our task. Specifically, the attention branch is added to CycleGAN. The input of the attention branch is the binary mask image of the input image. The output attention feature map is respectively combined with the output image of the generator and input image to perform dot product operation. Attention Loss can be obtained by subtracting the results of the two different dot product results. The Mask Loss is obtained by the ratio of the output feature map of the attention module to the binary mask map. The generator and discriminator adopt the structure suggested by the original CycleGAN.

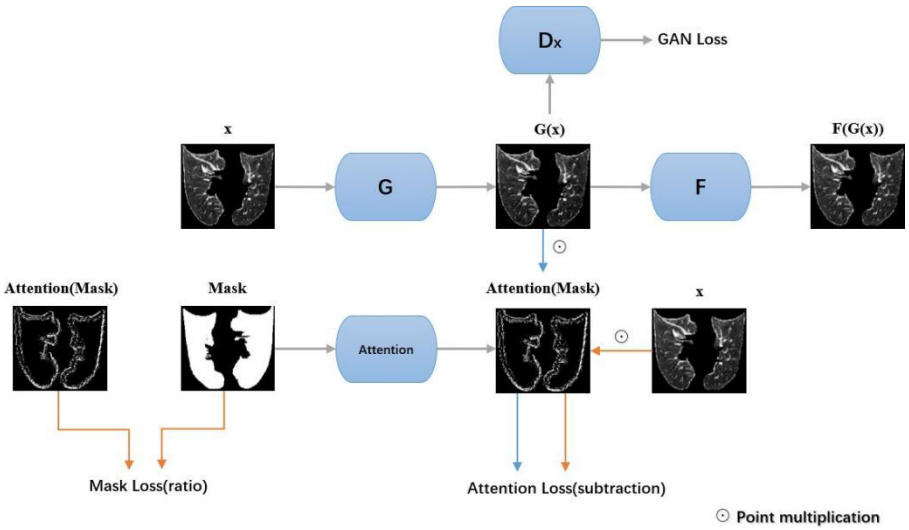


FIGURE 1: In the network pipeline, G is used to convert pneumonia negative images to positive, and F is used to convert pneumonia positive to negative. The input to the generator branch is the pneumonia negative image x, and the input to the self-attention branch is the binary mask of image x.

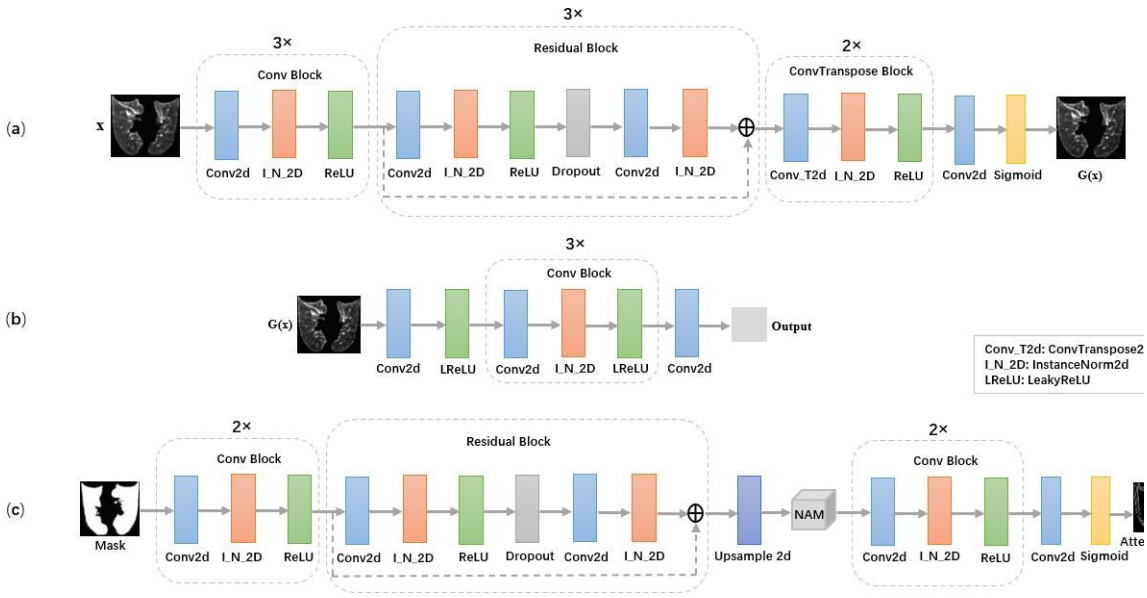


FIGURE 2: Architectures of network used in the proposed method. (a) Architecture of generator. (b) Architecture of discriminator. (c) The specific architecture of the self-attention branch. It consists of a downsampling module, a residual module, a NAM attention mechanism module, and an upsampling module.

The residual blocks in the generator are reduced to 3 on the radiation pneumonitis dataset. On the COVID-19 radiography dataset, the generator maintains the full structure with 6 residual blocks due to the more complex images. In the discriminator, spectral normalization is applied to improve the original convolution. The architectures of the generator and discriminator are shown in Figure 2 (a) and Figure 2 (b).

Figure 2 (c) shows the specific architecture of the attention branch. The attention branch consists of a downsampling module and an upsampling module. The binary mask image is input to the attention branch, and the number of channels is expanded by the convolution, which is to make full use of the channel attention in the attention mechanism. To prevent the gradient from vanishing, a residual structure is added before the attention module. The feature map by the NAM attention mechanism is passed through the upsampling module and the number of channels is gradually compressed to obtain the output. Figure 3 shows the results of the mask through the self-attentive branch. The effective pixel points of Attention(Mask) are mainly distributed at the edges of the lung slices. To make the results clearer, a threshold cutoff is applied to Attention(Mask). The cutoff value is set to 0.5.

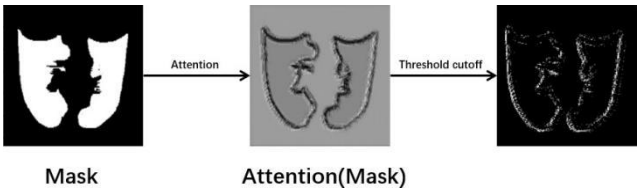


FIGURE 3: Self-Attention Results

D. LOSS FUNCTION

CycleGAN follows the design idea of GANs, and Adversarial Loss is one of the main goals of optimization. Its function is to gradually optimize the generator and discriminator in the iterative learning process to achieve dynamic balance so that the generator can generate as realistic as pictures of the target domain. Its loss function is expressed as follows in equation (3):

$$\begin{aligned} \mathrm{Loss}_{\mathrm{GAN}} &= \mathrm{L}_{\mathrm{GAN}}\,(\mathrm{G}, \mathrm{D}_{\mathrm{Y}}, \mathrm{X}, \mathrm{Y}) + \mathrm{L}_{\mathrm{GAN}}\,(\mathrm{F}, \mathrm{D}_{\mathrm{X}}, \mathrm{X}, \mathrm{Y}) \\ &= \mathrm{E}_{\mathrm{y}\sim\mathrm{pdata}}\,(\mathrm{y})\,[\log \mathrm{D}_{\mathrm{Y}}\,(\mathrm{y})] + \mathrm{E}_{\mathrm{x}\sim\mathrm{pdata}\,(\mathrm{x})}\,[\log\,(1 \\ &\quad - \mathrm{D}_{\mathrm{Y}}\,(\mathrm{G}(\mathrm{x}))] + \mathrm{E}_{\mathrm{x}\sim\mathrm{pdata}\,(\mathrm{x})}\,[\log \mathrm{D}_{\mathrm{X}}\,(\mathrm{x})] \\ &\quad + \mathrm{E}_{\mathrm{y}\sim\mathrm{pdata}\,(\mathrm{y})}\,[\log\,(1 - \mathrm{D}_{\mathrm{X}}\,(\mathrm{F}\,(\mathrm{y}))] \end{aligned} \tag{3}$$

However, just optimizing the Adversarial Loss often leads to mode collapse in the training process, because the input and output of the generator differ greatly in structure and content. Therefore, CycleGAN proposes two new loss functions, Cycle Consistency Loss and Identity Mapping Loss. Cycle Consistency Loss ensures that the output image of the generator is different in style from the input image, but the content is the same. Identity Mapping Loss, which is implemented using the mean absolute error (L1 loss), can assist the generator to convert more accurately, and ensure that the color tone of the generated image remains unchanged. Their loss function equation is as follows:

$$\begin{aligned} Loss_{cycle} &= E_{x\sim p_{data}(x)}[||F(G(x)) - x||_1] \\ &\quad + E_{y\sim p_{data}(y)}[||F(G(y)) - y||_1] \end{aligned} \tag{4}$$

$$Loss_{identity} = E_{y\sim p_{data}(y)}[||G(y)) - y||_1] + E_{x\sim p_{data}(x)}[||F(x) - x||_1] \quad (5)$$

To avoid excessive transfer on the input and output hierarchy and make the generated images more realistic, multi- level structure similarity(MS-SSIM)[37] is introduced in the identity mapping loss. So the new identity mapping loss is realized by L1 and MS-SSIM joint loss. MS-SSIM and L1 joint loss measures the difference between the generated images and the original images more effectively so that the model has a better conversion performance. Not only that, the generated images not only keep the content consistent with the original images but also have better quality and details. The implementation of MS-SSIM and L1 joint loss is referenced to[38].

In this study, two new loss functions, Self-Attention Loss and Mask Loss are proposed. The main function of Self- Attention Loss is to help the model to distinguish the image edge position information faster through the mask sub-attention module and locate the inner area of the mask. The specific calculation method is as follows: the input image and the output image of the generator are respectively dotted with the output feature map of the mask attention module, and then the result after the dot product is subtracted to obtain the loss value. In the process of model optimization, its loss value should be as low as possible. Its specific equation is as follows:

$$Loss_{attention} = E_{x\sim P_{data}(x)}[||F(G((x)).mask^* - x.mask^*||_1] - E_{y\sim P_{data}(y)}[||F(G((y)).mask^* - y.mask^*||_1] \quad (6)$$

The calculation method of Mask Loss is the ratio of the output feature map of the self-attention module to the original mask. Since the effect of the mask module is inversely proportional to the length of the training time, its value will gradually decrease during the training process. Its specific equation is as follows:

$$\text{Loss}_{\text{mask}} = E_{x\sim p_{\text{data}}(x)}\left[\left\|\frac{}{mask^*}\right\|\right] + E_{y\sim p_{\text{data}}(y)}\left[\left\|\frac{y}{mask^*}\right\|\right] \quad (7)$$

Therefore, throughout the training process, our overall optimization target is as follows:

$$\begin{aligned} \text{Loss} = \text{Loss}_{\text{GAN}} + \text{Loss}_{\text{cycle}} + \text{Loss}_{\text{identity}} \\ + \text{Loss}_{\text{attention}} + \text{Loss}_{\text{mask}} \end{aligned} \tag{8}$$

E. TRAINING PROCESS

On the RP dataset, to make the conversion ability of the model better, all the data was used to train the CycleGAN model, including its independent test set. To demonstrate the robustness and generalization of the method proposed in this study, experiments were conducted on the COVID- 19 dataset. 100 positive data samples and 300 negative data samples were sampled for model training, and other data were used as experimental test sets. During the training process, since we improved the generator structure, added a new mask part as input, and introduced a new loss function as an evaluation indicator for auxiliary model training, our training process is quite different from the original Cycle- GAN. The biggest difference is in the convergence speed of the model. Our method converges faster and the conversion ability of the model is better. The model is trained on a single GPU GTX 3080ti. The specific values of the main hyperparameters are: Epochs=100, Lr=0.00015, lambda_GAN=1.0, lambda_cyc=10.0, lambda_id=5.0, lambda_attention=1, lambda_Iou=1. Among them, lambda_GAN is the weight parameter of the adversarial loss, lambda_cyc is the weight parameter of the cycle consistency loss, and lambda_id is the weight parameter of the Identity Loss. It is worth noting that lambda_attention and lambda_Iou are the weight parameters of our newly added loss function, respectively. At the same time, Adam optimizer with 0.5 and 0.999 is used to train our CycleGAN.

EVALUATE METHOD

A. CONTRAST CHART

For radiographic pneumonia images, lesioned lung sections contain subtle characteristics that are difficult to detect intuitively compared to healthy lung

sections. Comparison charts are used to quantitatively demonstrate the differences between the pneumonia images generated by our method and the original images. The contrast chart is made by subtracting the generated image from the original image, and a specified threshold is set to emphasize the contrast difference more, same with the COVID-19 dataset.

B. CLASSIFICATION MODEL BASED ON RESNET18

ResNet18[39] is used as a classification model to evaluate the quality of the generated images. The modified CycleGAN generates pneumonia-positive images from negative images to complement the original data. ResNet18 is trained with a balanced data sample class and tested for its classification ability. ResNet18 is trained using data generated by our method with the original CycleGAN, conventional data augmentation (random horizontal flip, random rotation, and random scaling), and data without any processing to show the superiority of our suggested model. To evaluate the classification ability of ResNet18, the area under the ROC curve (AUC), accuracy (ACC), sensitivity (SEN), and specificity (SPE) are compared.

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \tag{9}$$

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

$$\mathrm{Spe} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{11}$$

Where TP is the number of correctly classified positive cases, FN is the number of incorrectly classified positive cases, TN is the number of correctly classified negative cases, and FP is the number of incorrectly classified negative cases.

C. ABLATION EXPERIMENT

An ablation experiment investigates the performance of the method by removing certain components to understand the contribution of the component to the overall method. To further show the effectiveness of the approach and to specif- ically evaluate the impact of our improvements on the base CycleGAN, an ablation experiment was completed on the RP dataset and COVID-19 radiography dataset respectively.

DATASET

Experiments on private and public datasets were done to show the robustness of our method.

A. RADIATION PNEUMONITIS DATASET

This is a real-world dataset containing 300 patients, of which 66 patients had RP (22%). In this dataset, the patient lung slices are stored as 3-dimensional image data. Considering the problem of computational memory consumption, the original 3D lung slice data was sliced into many images, and then the longitudinal center slice images are taken as the representative data sample of the patient. These images were normalized before being fed into the neural network.

B. COVID-19 RADIOGRAPHY DATASET

COVID-19 Radiography Dataset[40–42] is made by a team of researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh along with their collaborators from Pakistan and Malaysia in collaboration with medical doctors, which includes 3616 COVID-19 positive cases along with 10,192 Normal lung X-ray images and corresponding lung masks. In this study, 100 COVID-19 positive cases and 300 Normal lung X-ray images are used.

Results

A. CONTRAST CHART

The samples generated by our method are compared with the original samples and their comparison graphs are shown to demonstrate the effectiveness of our method. Figure 4 and Figure 5 show some representative examples generated by our method on the RP dataset and the COVID-19 radiography dataset respectively.

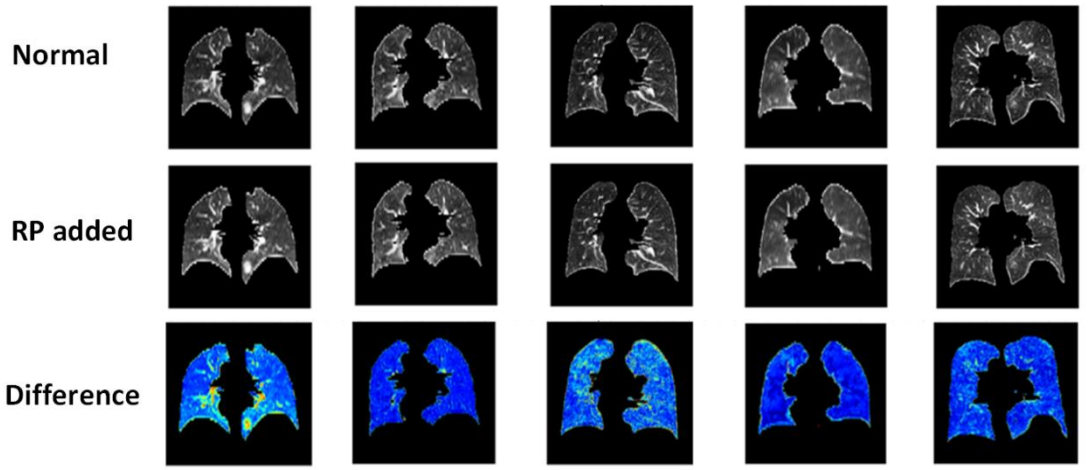


FIGURE 4: Results on radiation-induced pneumonia generation. The difference and heatmap between Original and Generated images are shown in the figure.

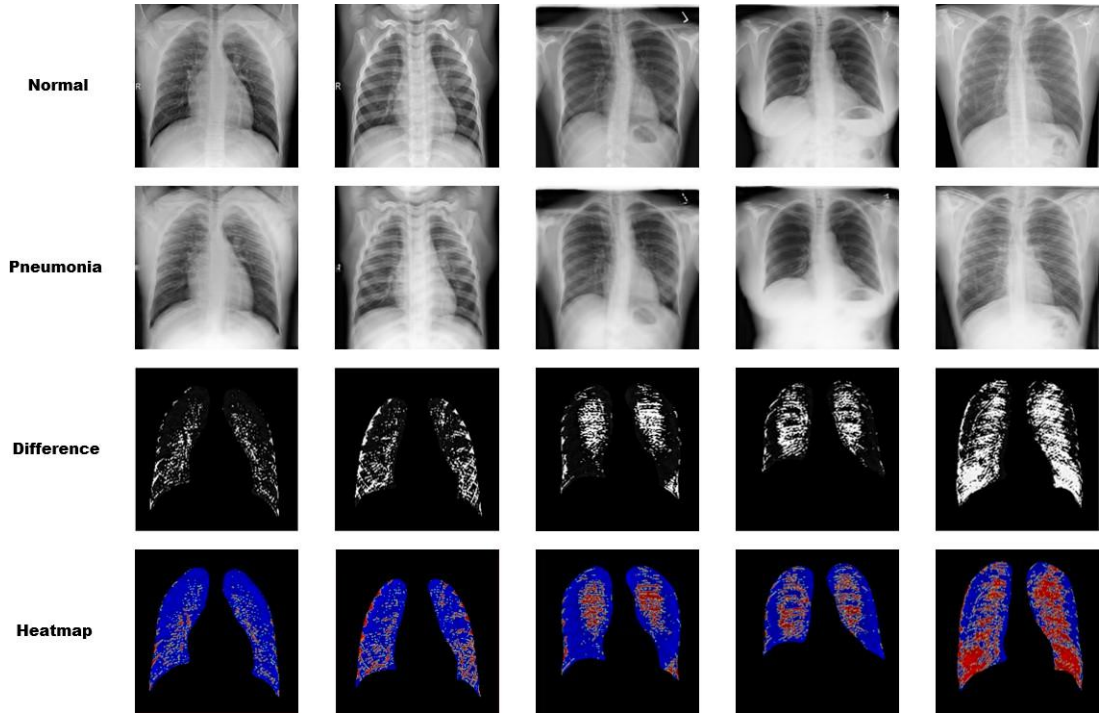


FIGURE 5: Results on COVID-19 generation. The difference and heatmap between Original and Generated images are shown in the figure. Compared to normal lung images, the generated images of COVID-19 are more fibrotic overall and more blurred in the lungs.

B. CLASSIFICATION EXPERIMENTS

The ResNet18 is used to train and calculate the AUC, ACC, SEN, and SPE on the RP dataset and the COVID-19 radiography dataset respectively, to verify the method's effectiveness. The result of the RP dataset is shown in Table 1. The values of SEN are low, 0.27 and 0.18 respectively when ResNet18 is trained and tested directly using raw data and traditional data augmentation. CycleGAN can significantly improve the model's classification ability, resulting in an improvement in SEN to 0.55 and a considerable improvement in AUC. Our method shows desirable improvements in AUC, ACC and SEN compared to CycleGAN. Our method has lower SPE values since the model is more likely to predict the test sample as negative when the positive sample is not representative. This phenomenon is evident when original data and traditional data augmentations are used.

TABLE 1: AUC, ACC, SEN, SPE on radiation pneumonitis dataset

methods	AUC	ACC	SEN	SPE
w/o DA	0.55(0.52-0.57)	0.63(0.60-0.65)	0.27(0.26-0.29)	0.79(0.77-0.81)
w/Traditional	0.57(0.56-0.59)	0.66(0.64-0.69)	0.18(0.17-0.19)	0.87(0.85-0.89)
w/CycleGAN	0.64(0.61-0.66)	0.63(0.60-0.64)	0.55(0.49-0.56)	0.67(0.65-0.68)
w/Ours	0.67(0.66-0.70)	0.66(0.64-0.67)	0.73(0.71-0.77)	0.63(0.60-0.64)

The result of the COVID-19 radiography dataset is shown in Table 2. There is no significant improvement in AUC when using the original data, traditional data augmentation, and original CycleGAN data augmentation. The AUC, ACC, and

SEN are improved when our method is used for data augmentation. It is worth noting that the SEN is the most significant improvement.

TABLE 2: AUC, ACC, SEN, SPE on COVID-19 radiography dataset

methods	AUC	ACC	SEN	SPE
w/o DA	0.88(0.88-0.88)	0.85(0.85-0.85)	0.52(0.52-0.52)	0.97(0.96-0.97)
w/Traditional	0.89(0.88-0.89)	0.87(0.87-0.87)	0.60(0.59-0.60)	0.96(0.96-0.96)
w/CycleGAN	0.89(0.89-0.89)	0.88(0.88-0.88)	0.63(0.63-0.63)	0.97(0.97-0.97)
w/Ours	0.92(0.92-0.92)	0.89(0.89-0.89)	0.68(0.67-0.68)	0.97(0.97-0.97)

Table 3 shows the comparison of our method with CovidGAN[43]. According to our survey, CovidGAN is the most relevant work to our study and one of the most state-of- the-art works. However, there are some differences between our work and CovidGAN, specifically in the number of data samples and the sample imbalance. The total number of samples in our training set is 400 while the number of training samples in CovidGAN is 932. The fewer training sets make it difficult to fit the model to the distribution of the data. Our training set contains 100 COVID-19 positive samples and 300 normal samples. The training set of CovidGAN contains 331 COVID-19 positive samples and 601 normal samples. Compared with CovidGAN, we use far fewer positive samples than negative samples, which can create difficulties for feature extraction of the model. CovidGAN achieves 0.95 Accuracy while our method is 0.89. Although our results are slightly lower than those of CovidGAN, with a much smaller amount of data and severe class imbalance, this result which benefits from the adopted CycleGAN generation paradigm and attention branching is not only acceptable but also surprising.

TABLE 3: Comparison with CovidGAN				
Methods	COVID-19	Normal	Ratio	Accuracy
CovidGAN	331	601	1: 1.815	0.95
Ours	100	300	1: 3	0.89

COVID-19: The number of COVID-19 samples Normal: The number of normal samples
Ratio: COVID-19/ Normal

C. ABLATION EXPERIMENT

The ablation experiments are completed on the two datasets separately to specifically evaluate the improvement of the model by the two added loss functions. The results are shown in Table 4 and Table 5. The results using spectral normalization and MS-SSIM and L1 joint loss are shown in the first row. The results with the addition of Attention Loss are shown in the second row. The results of the complete model are shown in the last row.

TABLE 3: Ablation Experiment on radiation pneumonitis dataset				
Loss	AUC	ACC	SEN	SPE
$L_{GAN} + L_{cycle} + L_{identity}$	0.64(0.63-0.66)	0.62(0.61-0.64)	0.57(0.55-0.58)	0.68(0.66-0.69)
$L_{GAN} + L_{cycle} + L_{identity} + L_{attention}$	0.65(0.64-	0.64(0.63-	0.69(0.69-	0.65(0.64-

	0.67)	0.65)	0.72)	0.66)
$L_{GAN} + L_{cycle} + L_{identity} + L_{attention} + L_{mask}$	0.67(0.66-0.70)	0.66(0.64-0.67)	0.73(0.71-0.77)	0.63(0.60-0.64)

TABLE 4: Ablation Experiment on COVID-19 radiography dataset

Loss	AUC	ACC	SEN	SPE
$L_{GAN} + L_{cycle} + L_{identity}$	0.91(0.91-0.91)	0.88(0.88-0.88)	0.60(0.60-0.60)	0.97(0.97-0.97)
$L_{GAN} + L_{cycle} + L_{identity} + L_{attention}$	0.91(0.91-0.91)	0.88(0.88-0.89)	0.66(0.66-0.66)	0.97(0.97-0.97)
$L_{GAN} + L_{cycle} + L_{identity} + L_{attention} + L_{mask}$	0.92(0.92-0.92)	0.89(0.89-0.89)	0.68(0.67-0.68)	0.97(0.97-0.97)

D. COMPUTATIONAL COMPLEXITY

Table 6 shows the number of parameters about the important structure of our model. On the radiation pneumonitis dataset, a single generator contains 3 residual blocks. Our improved generator parameters are 4.99 M, which is higher than the original CycleGAN generator of 0.71 M. On the Covid-19 radiography dataset, due to more complex images, the generator contains 6 residual blocks and the number of params is 8.53 M. The number of params of the generator we used is 2.76 M.

Models	Params
Original CycleGAN Generator(3 residuals blocks)	4.28M
Our Generator(3 residuals blocks)	4.99M
Original CycleGAN Generator(6 residuals blocks)	7.83M
Our Generator(6 residuals blocks)	8.53M
Discriminators	2.76M

Discussion

In this paper, a mask-based self-attention CycleGAN data augmentation method is proposed. According to the char- acteristics of our task, two different loss functions named Attention Loss and Mask Loss are designed. Spectral Normalization is introduced to improve the discriminator and the original Identity Mapping Loss function is replaced by MS- SSIM and L1 joint loss to replace.

Although we did a technical study, it still has important clinical utility. In previous studies, researchers used Gaussian Mixed Model or traditional augmentation methods to augment the data set. This is a simple and effective way to expand the training set. However, some augmentation methods are not suitable and reasonable based on clinical perspectives in real-world application settings. For example, elastic twist affects the anatomical structure of clinical images and adding Gaussian noise impacts texture features. These are important empirical reference features for the diagnosis of physicians, especially radiologists. The method in this study avoids these shortcomings and is more suitable for real-world studies.

In Section 6.1, the positive image data transformed from the negative data is shown. The images generated by the model have better quality and clarity while maintaining the same structure and content as the original images. From the comparison of the original images and the generated images, it can be concluded that the position changed by the model is mainly concentrated in the inner part of the lung slice, which is consistent with the medical facts, and shows that our model can accurately realize the transformation of the image domain.

Section 6.2 shows the results of the classification experiments. Considering that large and deep classification models have a greater risk of overfitting on few-shot, ResNet18 is used to complete the classification experiments. From the changes of SEN on the radiation pneumonitis data, it can be seen that the ResNet18 model trained by only using the original data and traditional data augmentation has a poor classification ability, and the entire data samples are often regarded as negative samples. This is why the SEN value is low and the SPE value is high. After using CycleGAN and our method for data augmentation, the SEN value has been significantly improved. Although its SPE value will decrease, the improvement of the AUC value indicates that the classifier has better classification ability. Regarding the change of SEN value on the COVID-19 radiography dataset, it can be noticed that the use of different data augmentation methods to expand the positive sample data set can increase the

SEN value to varying degrees. However, our method not only significantly outperforms other methods in the performance of SEN value but also has a small improvement in AUC, which indicates that the classifier trained by our method has better classification ability. In other words, the sample distribution of the data generated by our method is closer to the sample distribution of the original data, whether on the RP dataset or the COVID-19 radiography dataset. The comparison of our method with CovidGAN demonstrates the dramatic impact of the sample size and whether the categories are balanced in the experiment. It is obvious that our dataset is more challenging. Our dataset contains only 100 COVID-19 samples, so 200 COVID-19 samples have to be generated as added data to train the classification network. This means that the quality of the generated samples determines the final classification results. Therefore, although our accuracy is slightly lower than CovidGAN, this result is meaningful.

The experimental result in Section 6.3 shows that only using Spectral Normalization and MS_SSIM and L1 joint Loss function for data sample expansion can improve the classifier performance in a small range compared to the original CycleGAN data augmentation method. After adding Self-Attention Loss on its basis, it is obvious that SEN is worth improving, which means that the classifier has better classification ability for positive samples, indicating that our method is effective in data augmentation for positive samples. Ablation experiments on the COVID-19 dataset further verify the effectiveness of our proposed method.

Section 6.4 shows the computational complexity of the model. Two points are of interest. First, the generator containing self-attentive branches only increases the number of parameters by 0.71M compared to the original CycleGAN. In fact, successive stacks of residual blocks contribute most of the complexity of the generator. Second, although we did not optimize the number of parameters of the model more in this research, our model still has low complexity.

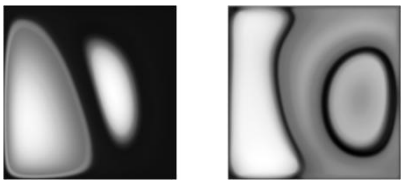


FIGURE 6: Results on radiation-induced pneumonia generation with WGAN-GP.

Figure 6 shows the results of radiation-induced pneumonia image generation using WGAN-GP. Even after training, the generated images still largely follow

the initial random distribution. This is due to the difficulty of providing sufficient information for network optimization with sparse training data. CycleGAN, which uses style migration, has a large amount of target distribution and structural information in its input. Therefore, it is difficult to generate medical images from noise on Few-shot compared to CycleGAN.

Limitations and future work

In this paper, a mask-based self-attention CycleGAN data augmentation method is proposed. Our method has a better performance on both the RP dataset and the COVID-19 radiography dataset. However, during our experiments, we found that there are still many problems that need to be solved. The first is the instability of training CycleGAN on few-shot, which is difficult to avoid during GANs training, although the stability has been substantially improved compared to the original CycleGAN. Second, the CycleGAN has limitations, particularly in its generative capability. Cycle- GAN can capture the main features of the image style and transform the original input image to another style, while it is difficult to generate minute details that the original input image itself does not have. The use of Mask and the new loss function allows the model to accurately capture features at the marginal locations of the lungs that are more likely to distinguish between negative and positive pneumonia. However, such processing makes it difficult to focus on the interior of the lung, producing more improvements in the generation of internal images. Therefore, the generation of internal lung details relies more on the original loss function of CycleGAN. Finally, our model is not optimized for complexity or the number of parameters. Although a certain model complexity is positive for the generation of high- quality images, a lower complexity is essential for the practical application of the method. In the future, we will continue to investigate CycleGAN-based medical data augmentation methods to propose effective solutions to the above problems.

Conclusions

This paper proposes a mask-based self-attentive CycleGAN data augmentation method for overcoming the problem of medical image sample imbalance. We use Spectral Normalization for the discriminator, introduce MS-SSIM and L1 joint loss to stabilize the training process, and design two loss functions named Self-Attention Loss and Mask Loss to guide training. Experiments on the RP

dataset and COVID- 19 radiography dataset and excellent performance on several classification metrics have demonstrated that our method can generate delicate images that better match the target domain than traditional data augmentation and original CycleGAN data augmentation methods. Therefore, our method can be utilized as a general data augmentation method to assist in overcoming the sample imbalance problem in medical image datasets.

References

[1]P. N. Srinivasu, J. G. SivaSai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, "Classification of skin disease using deep learn- ing neural networks with mobilenet v2 and lstm," Sensors, vol. 21, no. 8, p. 2852, 2021.

[2]A. Vulli, P. N. Srinivasu, M. S. K. Sashank, J. Shafi, J. Choi, and M. F. Ijaz, "Fine-tuned densenet-169 for breast cancer metastasis prediction using fastai and 1-cycle policy," Sensors, vol. 22, no. 8, p. 2988, 2022.

[3]J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute,

H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 590–597.

[4]J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno et al., "fastmri: An open dataset and benchmarks for accelerated mri," arXiv preprint arXiv:1811.08839, 2018.

[5]T. J. Bledsoe, S. K. Nath, and R. H. Decker, "Radiation pneumonitis," Clinics in chest medicine, vol. 38, no. 2, pp. 201–208, 2017.

[6]M. F. Ijaz, M. Attique, and Y. Son, "Data-driven cervical can- cer prediction model with outlier detection and over-sampling methods," Sensors, vol. 20, no. 10, p. 2809, 2020.

[7]S. Dash, S. Verma, S. Bevinakoppa, M. Wozniak, J. Shafi, and M. F. Ijaz, "Guidance image-based enhanced matched filter with modified thresholding for blood vessel extraction," Symmetry, vol. 14, no. 2, p. 194, 2022.

[8]S. Dash, S. Verma, M. S. Khan, M. Wozniak, J. Shafi, and

M. F. Ijaz, "A hybrid method to enhance thick and thin vessels for blood vessel segmentation," Diagnostics, vol. 11, no. 11, p. 2017, 2021.

[9]I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde- Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[10]J. Liu, C. Shen, T. Liu, N. Aguilera, and J. Tam, "Active appearance model induced generative adversarial network for controlled data augmentation," in

International Conference on Medical Image Computing and Computer-Assisted Interven- tion. Springer, 2019, pp. 201–208.

[11]H. Rashid, M. A. Tanveer, and H. A. Khan, “Skin lesion clas- sification using gan based data augmentation,” in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019, pp. 916– 919.

[12]D. Srivastav, A. Bajpai, and P. Srivastava, “Improved classi- fication for pneumonia detection using transfer learning with gan based synthetic image augmentation,” in 2021 11th In- ternational Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2021, pp. 433–437.

[13]M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein gan,” 2017.

[14]J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image- to-image translation using cycle-consistent adversarial net- works,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[15]Y. Alami Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim, “Unsupervised attention-guided image-to-image trans- lation,” Advances in neural information processing systems, vol. 31, 2018.

[16]T. Kwon and J. C. Ye, “Cycle-free cyclegan using invert- ible generator for unsupervised low-dose ct denoising,” IEEE Transactions on Computational Imaging, vol. 7, pp. 1354– 1368, 2021.

[17]J. Gu and J. C. Ye, “Adain-based tunable cyclegan for efficient unsupervised low-dose ct denoising,” IEEE Transactions on Computational Imaging, vol. 7, pp. 73–85, 2021.

[18]W. Zheng, L. Yan, W. Zhang, C. Gou, and F.-Y. Wang, “Guided cyclegan via semi-dual optimal transport for photo- realistic face super-resolution,” in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 2851–2855.

[19]R. E. Rivadeneira, A. D. Sappa, and B. X. Vintimilla, “Ther- mal image super-resolution: A novel architecture and dataset.” in VISIGRAPP (4: VISAPP), 2020, pp. 111–119.

[20]H. Liu, J. Liu, S. Hou, T. Tao, and J. Han, “Perception con- sistency ultrasound image super-resolution via self-supervised cyclegan,” Neural Computing and Applications, pp. 1–11, 2021.

[21]H. Ji, Z. Gao, X. Liu, Y. Zhang, and T. Mei, “Small object de- tection leveraging on simultaneous super-resolution,” in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 803–810.

[22]J. Zhao, J. Zhang, Z. Li, J.-N. Hwang, Y. Gao, Z. Fang,

X. Jiang, and B. Huang, “Dd-cyclegan: Unpaired image de- hazing via double-discriminator cycle-consistent generative adversarial network,” Engineering Applications of Artificial Intelligence, vol. 82, pp. 263–271, 2019.

[23]Y.-F. Chen, A. K. Patel, and C.-P. Chen, “Image haze re- moval by adaptive cyclegan,” in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2019, pp. 1122–1127.

[24]D. Engin, A. Genç, and H. Kemal Ekenel, “Cycle-dehaze: Enhanced cyclegan for single image dehazing,” in Proceed- ings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 825–833.

[25]D. I. Morís, J. J. de Moura Ramos, J. N. Buján, and M. O. Hortas, “Data augmentation approaches using cycle-consistent adversarial networks for improving covid-19 screening in portable chest x-ray images,” Expert Systems with Applica- tions, vol. 185, p. 115681, 2021.

[26]T. Malygina, E. Ericheva, and I. Drokin, “Data augmentation with gan: Improving chest x-ray pathologies prediction on class-imbalanced cases,” in International conference on anal- ysis of images, social networks and texts. Springer, 2019, pp. 321–334.

[27]Z. Xu, C. Qi, and G. Xu, “Semi-supervised attention-guided cyclegan for data augmentation on medical images,” in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019, pp. 563–568.

[28]T. de Bel, J.-M. Bokhorst, J. van der Laak, and G. Lit- jens, “Residual cyclegan for robust domain transformation of histopathological tissue slides,” Medical Image Analysis, vol. 70, p. 102004, 2021.

[29]O. Tmenova, R. Martin, and L. Duong, “Cyclegan for style transfer in x-ray angiography,” International journal of com- puter assisted radiology and surgery, vol. 14, no. 10, pp.1785–1794, 2019.

[30]W. Liu, B. Luo, and J. Liu, “Synthetic data augmentation using multiscale attention cyclegan for aircraft detection in remote sensing images,” IEEE Geoscience and Remote Sensing Let- ters, vol. 19, pp. 1–5, 2021.

[31]D. Organisciak, E. S. Ho, and H. P. Shum, “Makeup style transfer on low-quality images with weighted multi-scale at- tention,” in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 6011–6018.

[32]Y. Liu, Z. Shao, Y. Teng, and N. Hoffmann, “Nam: Normalization-based attention module,” arXiv preprint arXiv:2111.12419, 2021.

[33]S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in International conference on machine learning. PMLR, 2015, pp. 448–456.

[34]Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, “Learning efficient convolutional networks through network slimming,” in Proceedings of the IEEE international confer- ence on computer vision, 2017, pp. 2736–2744.

[35]I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” Advances in neural information processing systems, vol. 30, 2017.

[36]T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spec- tral normalization for generative adversarial networks,” arXiv preprint arXiv:1802.05957, 2018.

[37]Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2. Ieee, 2003, pp. 1398–1402.

[38]H. Zhao, O. Gallo, I. Frosio, and J. Kautz, “Loss functions for image restoration with neural networks,” IEEE Transactions on computational imaging, vol. 3, no. 1, pp. 47–57, 2016.

[39]K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770– 778.

[40]“Covid19 radiography database,” https://www.kaggle.com/datasets/tawsifurrahman/covid19- radiography-database.

[41]M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar,

M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal,

N. Al Emadi et al., “Can ai help in screening viral and covid- 19 pneumonia?” IEEE Access, vol. 8, pp. 132 665–132 676, 2020.

[42]T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz,

S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier,

M. S. Khan et al., “Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images,” Computers in biology and medicine, vol. 132, p. 104319, 2021.

[43]A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, “Covidgan: data augmentation using aux- iliary classifier gan for improved covid-19 detection,” Ieee Access, vol. 8, pp. 91 916–91 923, 2020.

Chapter 5: Radiomics and dosiomics signature from whole lung predicts radiation pneumonitis: a model development study with prospective external validation and decision-curve analysis

Adapted from: Zhen Zhang, Zhixiang Wang, Meng Yan, Jiaqi Yu, Andre Dekker, Lujun Zhao, Leonard Wee. Radiomics and Dosiomics Signature from Whole Lung Predicts Radiation Pneumonitis: A Model Development Study with Prospective External Validation and Decision-Curve Analysis. International Journal of Radiation Oncology, Biology, Physics 2022, S0360-3016(22)03189-3. https://doi.org/10.1016/j.ijrobp.2022.08.047

Abstract

Purpose Radiation pneumonitis (RP) is one of the common side effects of radiotherapy in the thoracic region. Radiomics and dosiomics quantifies information implicit within medical images and radiotherapy dose distributions. In this study we demonstrated the prognostic potential of radiomics, dosiomics, and clinical features for RP prediction.

Materials and methods Radiomics, dosiomics, dose-volume histogram (DVH) metrics, and clinical parameters were obtained on 314 retrospectively-collected and 35 prospectively-enrolled patients diagnosed with lung cancer between 2013 to 2019. A radiomics risk score (R-score) and dosiomics risk score (D-score) and DVH-score were calculated based on logistic regression after feature selection. Six models were built using different combinations of R-score, D-score, DVH-score, and clinical parameters to evaluate their added prognostic power. Over-optimism was evaluated by bootstrap resampling from the training set, and the prospectively-collected cohort was used as the external test set. Model calibration and decision-curve characteristics of the best-performing models were evaluated. For ease of further evaluation, nomograms were constructed for selected models.

Results A model built by integrating all of R-score, D-score, and clinical parameters had the best discriminative ability with area under the curves (AUCs) of 0.793 (95%CI 0.735-0.851), 0.774 (95%CI 0.762-0.786), and 0.855 (95%CI 0.719-0.990) in the training set, bootstrapping set, and external test set, respectively. The calibration curve image showed good agreement between the predicted and actual values with a slope of 1.21 and an intercept of - 0.04. The decision curve image showed positive net benefit for the final model based on the nomogram.

Conclusion Radiomics and dosiomics features have potential to assist with the prediction of RP, and the combination of radiomics, dosiomics, and clinical parameters led to the best prognostic model in the present study.

Keywords: Radiomics; Dosiomics; Lung cancer; Radiation Pneumonitis

Introduction

Radiotherapy (RT) plays a crucial role in the management of lung cancer (LC) [1], especially for locally advanced and unresectable cases [2, 3]. Advances in thoracic RT have led to steadily improving prognosis for LC patients, but RT-related side effects remain a treatment-limiting concern [4-6]. Radiation pneumonitis (RP) is a common adverse effect that degrades patients' quality of life and can be fatal in severe cases. To date, there is no highly effective cure for RP [7], thus prevention of RP remains one of the top clinical priorities during RT dose planning [8, 9]. Robust and reproducible predictive models that could estimate the risk of developing RP after lung RT would be of immense clinical value. Such estimates could be incorporated into treatment planning and informed shared decision-making consultations (such as a choice between starting prophylactic medication or active vigilance).

Studies to date suggest a number of clinical factors, such as smoking status, pre-existing lung disease [10], pre-existing cardiac disease [11], and chemotherapy [12], may affect an individual's pre-disposition to develop RP. Although these parameters may indicate towards susceptibility, RP remains a disease exhibiting strong inter-person variability [13]; this heterogeneity does not appear to be sufficiently well represented in conventional clinical factors. Single-nucleotide polymorphism (SNPs) [14] and plasma cytokines [15, 16] can also be indicative of heterogeneity, and several studies have revealed significant associations between SNPs and the occurrence of RP [17], which suggests the feasibility of genetic and molecular biomarkers. However, some biomarkers may be subject to vagaries of limited spatial sampling and are only available through invasive means.

Radiomics is the high-throughput extraction of quantitative handcrafted features from medical images. Image-based radiomics has the potential to characterize heterogeneity within the entire pre-RT lung parenchyma and, in the case where suitable repeated imaging could be available, to be able to quantify parenchymal changes during a course of RT in a non-invasive manner. It has been demonstrated that radiomics features are associated with genetic heterogeneity (radiogenomics) [18]. There have been several studies that demonstrate the potential of radiomics to predict RP [19-21], but building predictive models only from an image perspective may not be sufficient.

Physicians routinely modify treatment strategies based on the patient's condition. For example, some patients with pre-existing lung disease diagnosed by imaging may be prescribed a relatively low dose thereby reducing the chance of developing RP and weakening the predictive power of radiomics. Therefore, there is a need to incorporate prescription dose information into predictive model.

In a different context, the occurrence of RP is strongly related to RT dose, and therefore a number of studies have used dose-volume histogram (DVH) metrics, such as mean lung dose (MLD) [22] and volume of the lung receiving 20 Gy (V20) [23], to predict RP. DVH parameters are not able to fully describe the immense spatial heterogeneity of dose distribution, which may be realized through intensity modulated radiation delivery (i.e., IMRT and/or VMAT) [23, 24]. Dosiomics, conceived as using radiomics tools to characterize spatial heterogeneity of RT dose (as opposed to image voxel intensities) provides a greater depth of information in contrast to traditional DVH measures [25, 26].

Previous works [19-21, 25, 26] attempt to predict RP solely on the basis of medical (tomographic) imaging alone, or on the basis of dose information, and those results show that it is highly unlikely to be clinically sufficient by relying exclusively on either the imaging features (radiomics) or the dose-volume parameters. There is a lack of studies combining radiomics and dosiomics to predict RP in lung cancer, furthermore, studies using rigorous and rational steps of selecting handcrafted features are needed. A large sample-based, prospective study is also required to assess the objective predictive power of the models.

In this study, we extracted radiomics features in RT planning CT and dosiomics features in 3D dose grids from the RT treatment planning system (TPS) and performed objective and rigorous feature selection. We evaluated the performance of clinical parameters, radiomics features, dosiomics features, and DVH metrics, singly as well as in combination, to predict RP after RT to the chest area. We evaluated the prediction models in terms of discriminative performance and model calibration using a prospectively collected dataset. Moreover, decision-curve analysis was used to investigate the potential clinical relevance of such models if implemented in routine practice. A nomogram was provided to facilitate future independent validation of our work in other clinical settings.

Methods

1. Study design

This study was designed as a Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) type 3 study comprising model development and independent validation [27]. This study was registered on artificial intelligence in biomedical research platform (AIMe, ID: mn9jLf) [28]. The overarching study flow is illustrated in Figure 1.

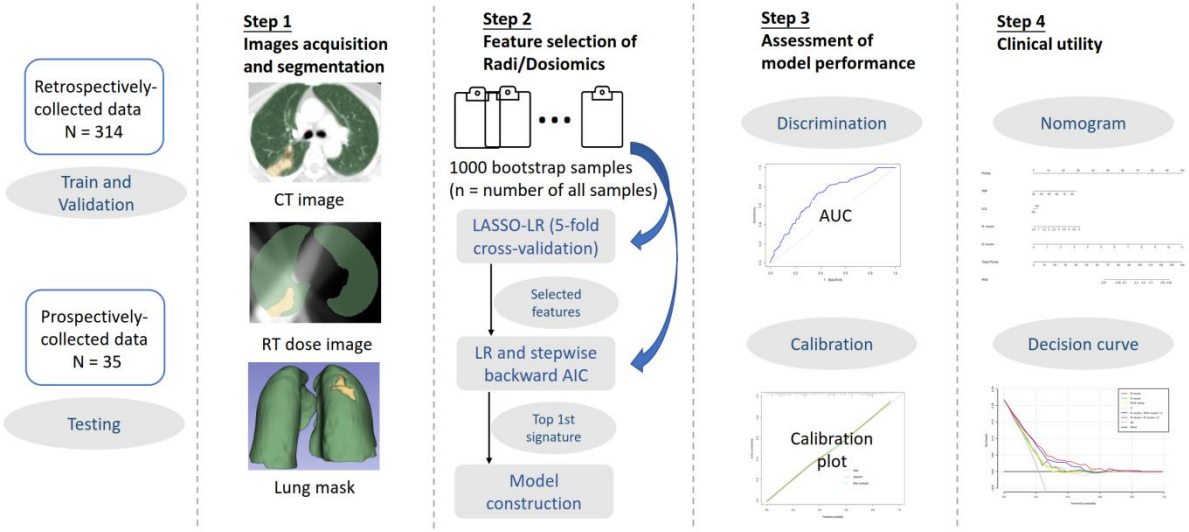


Figure 1. Analysis flowchart. Step 1, The radiomics and dosiomics features of the lung tissue region were extracted. Step 2, 1000 unique bootstrap samples were taken from all samples, features were selected by correlation, least absolute shrinkage (LASSO) embedded with logistic regression (LR) and Akaike information criterion (AIC) for modeling. Step 3, The model performance was evaluated using discrimination and calibration. Step 4, Clinical applications were evaluated using nomogram and decision curves.

2. Patients

A single-institutional model development cohort of 314 subjects was retrospectively extracted from institutional records after ethics board approval (IRB/bc2021135), comprising patients diagnosed with LC and treated with radical (chemo)-RT, with either IMRT or VMAT techniques, at Anonymized for Review Hospital between January 2013 and December 2018. For model validation, an additional 35 patients with the same criteria were prospectively enrolled with informed consent and same ethics approval (IRB/bc2021135), who

were treated between October 2018 and March 2019 in the same institution. Detailed inclusion and exclusion criteria have been specified in the Supplementary Materials A.

3. Image acquisition and treatment planning

Intravenous contrast-enhanced planning CT scans were acquired on a single Brilliant (Philips Medical Systems; Best, The Netherlands) multislice scanner with a standardized protocol: 120 kVp, 100 mAs, 3 mm slice thickness, 512 x 512 image matrix, 50 cm fields of view, 0.977 mm pixel spacing and vendor's default convolution kernel. Experienced radiation oncologists delineated the LC gross tumor volume (GTV) and malignant lymph nodes in the Pinnacle TPS (Philips Radiation Oncology Systems; Fitchburg, Wisconsin, United States), with image fusion against complementary imaging studies whenever available (such as positron emission tomography).

The GTV was isotropically expanded by 5 mm, as well as subclinical microscopic malignant lesions to derive the clinical target volume (CTV). The planning target volume (PTV) was an additional 5 mm isotropic expansion around the CTV. Dosimetrist were instructed to cover at least 95% of the PTV with the prescribed RT dose. Delineations conformed to the guidelines set by the Radiotherapy and Oncology Group (RTOG). The relevant dose constraints were as follows: MLD < 20 Gy, V20 < 30%, and volume of the lung receiving 5 Gy (V5) < 60%. All patients were nominally prescribed 2 Gy per fraction once daily. Radiation oncologists determined the total prescribed dose based on each patient's overall physical condition and best achievable normal tissue constraints. The actual total RT dose delivered ranged between 50 to 70 Gy. The dose grid resolution is 4 mm, and the dose calculation algorithm is Collapsed Cone Convolution [29, 30]. The planning CT series with associated RT structure delineations and RT planned radiotherapy 3D dose grids were exported from Pinnacle in the standard DICOM format.

4. Lung segmentation and RP grading

We extracted radiomics features and dosiomics features from the region corresponding to total (left plus right) lung. To ensure consistency of lung segmentation, we quality assured the lung delineations for each subject using a deep-learning automatic lung contouring tool based on retraining of the published model. The original and automatically generated lung outlines were inspected and then manually edited by a single experienced radiation

oncologist (author MY). Two other radiation oncologists (author JQY and ZZ) subsequently independently reviewed the lung organ segmentation, and any disputes were resolved by direct consultation among all three authors.

The primary outcome RP was defined, in accordance with the Common Terminology Criteria for Adverse Events (CTCAE) v5.0, as symptomatic radiation pneumonitis of CTCAE grade 2 or higher within 6 months after the end of RT [12, 16]. Monitoring of RP was based on the combination of clinical examination, reported symptoms, outpatient medical records, laboratory tests, chest X-ray, and visual inspection of follow-up CTs, which were all performed at intervals of 1, 3, and 6 months after completion of RT, and then every 6 months thereafter.

5. Radiomics and dosiomics features extraction

A total of 103 handcrafted radiomics features were extracted from DICOM CT and RT Structures using the "O-RAW" package [31] (based on Pyradiomics v3.7 [32]). These features comprised 17 intensity histogram features, 13 morphological (shape) features, and 73 textural features. No digital image filters were applied during pre-processing. Most of the hand-crafted features conformed to the Image Biomarker Standardization Initiative (IBSI) [33]; specific divergences from the IBSI at the time of writing have been reported according to the PyRadiomics documentation. Radiomics extraction settings are the same as for a previous publication [31], and our PyRadiomics parameters setting file has been provided in the Supplementary Materials B. For dosiomics features, DICOM RT Dose files were first converted as NRRD images using 3D Slicer [34], and then the same feature extraction procedure in PyRadiomics was applied for the total lung region. Additionally, voxel-wise values in the "dose images" were scaled to represent the absolute physical dose in units of Gray (Gy). Isotropic spatial resampling (1 mm) was applied on the CT images and dose images prior to feature extraction as recommended by previous studies [35].

6. Feature selection

An overview of multi-step feature selection and model construction is given in Figure 1. The clinical parameters for modeling were evaluated by using univariate and multivariate analyses for twelve clinical parameters with predictive potential. Feature selection for the radiomics model and the dosiomics model were performed separately, and has been adapted from the

feature pooling and signature pooling method used by Compter et al. [36]. In brief, the selection process was as follows:

(i) A thousand unique bootstrap samples (with replacement) were drawn from the whole training cohort. Within each bootstrap sample, we first minimized the number of strong pairwise normalized (Z-score, (original value-mean value)/standard deviation) feature correlations greater than 0.90 or less than -0.90. A least absolute shrinkage (LASSO) loop with 20-times repeated 5-fold cross-validation embedded with a logistic regression (LR) supervised classifier was used to select features. From each of the 1000 bootstraps, we ranked each individual feature according to how frequently it was retained by the LASSO-LR.

(ii) We arbitrarily selected some of the top most frequently-appearing individual features from the above table. From this small subset of selected features, we built a multivariable LR model on each of the same aforementioned bootstraps samples with stepwise backwards elimination using the Akaike information criterion (AIC) as metric. From each of these 1000 bootstraps, we tabulated how many times each combination of one or more features (i.e., potential signatures) was retained by the stepwise LR.

(iii) We arbitrarily selected the top most frequently-appearing signature arbitrarily selected to build the final multivariable LR model. The coefficients of the final model were fitted using the original non-bootstrapped development cohort.

7. Model construction

The clinical model was presented as a multivariable LR model. To this, we added an aggregated Radiomics Risk Score (R-score) and an aggregated Dosiomics Risk Score (D-score), separately. The R-score was defined as the linear predictor (LP) of the multivariable LR radiomics model, and likewise the D-score was defined as the LP of the multivariable dosiomics model. For combined models, we assessed the combinations of the clinical factors together with either, or both, of the R-score and D-score.

V20 and mean lung dose (MLD) were used to build DVH model, and details of feature selection and model construction are provided in the Supplementary Materials C. To address the issue of imbalanced data, we performed the Synthetic Minority Oversampling Technique (SMOTE) approach in the training

set. We also examined the Pearson correlation between the R-score and clinical parameters, and between the D-score and dose-volume histogram metrics (dosimetrics).

8. Model validation – internal and external

We estimated the over-optimism in model development using the method recommended in the TRIPOD guidelines; for each of the 1000 abovementioned pre-defined bootstraps, we fitted the LR model coefficients on each bootstrap, and then computed its Area under the curve (AUC) of receiver operating characteristic curve (ROC) using the original non-bootstrapped development cohort. From these 1000 bootstraps, we computed the average AUC and its 95% confidence interval.

As external validation, we evaluated the aforementioned models using the prospectively-registered cohort of 35 subjects. Processing of these 35 subjects followed exactly the same procedure as for the model development cohort, and none of these subjects were used in any way during model construction.

The well-established calibration curve technique was used to assess model goodness of fit (i.e., the extent of concordance between the predicted and observed values) again using a bootstrap of 1000 repetitions. To facilitate clinical use and support fully independent validation of our model, a simple nomogram was generated for the R-score, D-score, and the selected clinical parameters. Lastly, we tried to discuss the potential clinical utility of our model using decision curve analysis (DCA) [37].

9. Statistical analyses

Baseline patient characteristics for continuous variables are presented as mean ± standard deviation. For univariate ranking of clinical predictors, Pearson $X2$ tests and exact Fisher tests were used for categorical variables and logistic regression for continuous variables. For significance of clinical factors, a two-sided hypothesis test at the $\alpha = 0.05$ confidence level was assumed. Significant characteristics were subsequently combined in multivariable logistic regression.

All data had been collated and standardized using the Statistical Package for Social Science program (SPSS for Windows, version 27.0; SPSS Inc, Chicago, IL). Feature selection, model construction, model performance assessment and decision-curve analysis were all performed in R software (version 4.0.5).

Results

1. Patient characteristics and incidence of RP

The case mix of patients and treatments studied in this model are reported in Table 1. Univariate analysis showed statistically significant differences in interstitial lung disease (ILD), concurrent chemoradiotherapy (CCRT), and age between patients with and without RP. The overall incidence of CTCAE grade 2 or higher for RP was 21.5% (75 of 349), 21% (66 of 314) in the retrospective data set, and 25.7% (9 of 35) in the prospective validation set. Multivariable analysis indicated that ILD (OR 2.471; 95%CI 1.037-5.888, p = 0.041) and age (OR 1.051; 95%CI 1.012-1.085, p =0.008) were independent factors associated with RP. A forest plot for the coefficients in the multivariable LR model is shown in Figure 2.

Table 1 Patient Characteristics

Characteristics	All retro pts n (%)	Without RP2 Mean ± SD	With RP2 Mean ± SD	P*	Pros pts n (%)
Age median	61 (30-85)	61 (30-85)	63 (44-79)	0.005	62 (34-75)
Gender				0.523	
Male	238 (75.8%)	186 (78.2%)	52 (21.8%)		23 (65.7%)
Female	76 (24.2%)	62 (81.6%)	14 (18.4%)		12 (34.3%)
Smoking				0.569	
Yes	244 (77.7%)	191 (78.3%)	53 (21.7%)		26 (74.3%)
No	70 (22.3%)	57 (81.4%)	13 (18.6%)		9 (25.7%)
KPS				0.725	
≤80	132 (42.0%)	103 (78.0%)	29 (22.0%)		13 (37.1%)
>80	182 (58.0%)	145 (79.7%)	37l (20.3%)		22 (62.9%)
Diabetes				0.609	
Yes	34 (10.8%)	28 (82.4%)	6 (17.6%)		2 (5.7%)
No	280 (89.2%)	220 (78.6%)	60 (21.4%)		33 (94.3%)
ILD				0.015	
Yes	25 (8.0%)	15 (60.0%)	10 (40.0%)		9 (25.7%)
No	289 (92.0%)	233 (80.6%)	56 (19.4%)		26 (74.3%)
Pathology				0.656	
LUSC	86 (27.4%)	65 (75.6%)	21 (24.4%)		8 (22.9%)
LUAD	73 (23.2%)	59 (80.8%)	14 (19.2%)		10 (28.6%)

SCLC	155 (49.4%)	124 (80.0%)	31 (20.0%)		17 (48.5%)
Induc chemo				0.739	
Yes	287 (91.4%)	226 (78.7%)	61 (21.3%)		31 (88.6%)
No	27 (8.6%)	22 (81.5%)	5 (18.5%)		4 (11.4%)
CCRT				0.047	
Yes	93 (29.6%)	168 (76.0%)	53 (24.0%)		8 (22.9%)
No	221 (70.4%)	80 (86.0%)	13 (14.0%)		27 (77.1%)
Conso chemo				0.116	
Yes	179 (57.0%)	147 (82.1%)	32 (17.9%)		19 (54.3%)
No	135 (43.0%)	101 (74.8%)	34 (25.2%)		16 (45.7%)
PGTV(Gy)	59.274±2.977	59.204±3.063	59.539±2.634	0.415	60.200±2.870
Smoking index	661.540±571.430	641.840±550.543	735.600±643.084	0.237	668.600±550.412

Abbreviations: Retro = retrospective; Pts = patients; Pros = prospective; LUSC = lung squamous cell carcinoma; LUAD = lung adenocarcinoma; SCLC = small cell lung cancer; IMRT = intensity-modulated radiotherapy; VMAT = volumetric modulated arc therapy; chemo = chemotherapy; KPS = Karnofsky performance score; Induc chemo = induction chemotherapy; CCRT = concurrent chemoradiotherapy; Conso chemo = consolidation chemotherapy; PGTV = planning gross tumor volume.

*The differences in characteristics were evaluated by logistic regression for continuous variables or Pearson X2 test and exact Fisher test for categorical variables

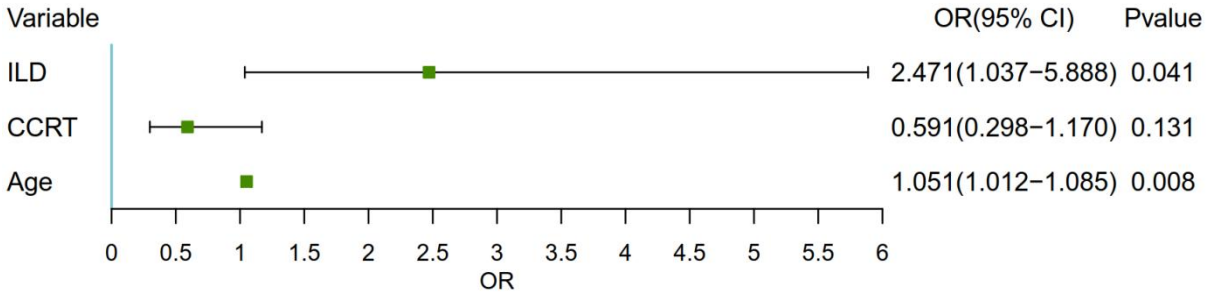


Figure 2. Multivariate analysis forest plot by logistic regression. Characteristics with statistically significant univariate analysis were subjected to multivariate analysis, with ILD and age as independent predictors of RP. Abbreviations: OR = Odds ratio; ILD = Interstitial lung disease; CCRT = Concurrent chemoradiotherapy.

2. Feature selection and risk scores

By inspecting the frequency ranking of individual features, we noted that a threshold frequency of around 600 yielded us 11 radiomics features and 12 dosiomics features. Subsequently, we derived a final radiomics signature comprising of 7 features for the R-score, and a final dosiomics model of 6 features for the R-score. Detailed tables and graphs from the feature selection process, along with the names and definitions of the selected features, are provided in the Supplementary Materials D.

The R-score and the D-score were calculated based on the coefficients weighted by LR. The formula of R-score and D-score are provided in the Supplementary Materials D. For ease of computing the R-score and D-score, a simple calculator has been provided and can be found here: only for Windows or MacOS operating systems, (https:// https://github.com/Radiologyzz/Calculator.git). Instructions for using the calculator are given in the Supplementary Materials E.

Examples of low and high R-score and D-score are given in Figure 3. In this example, ILD was evident in the patient with high R-score. The lung tissue of the patient with high D-score received higher dose of radiation than the patient with low D-score (the same prescription dose for both patients). The results showed no significant correlation (>0.8) by Spearman's analysis between R-score and clinical parameters, D-score and dosimetrics, respectively (Supplementary material F Figure 3). However, there were slight differences in the distribution of R-score for the population with and without ILD, and more noticeable differences in the distribution of D-score for the population with different MLD (Supplementary material F Figure 4).

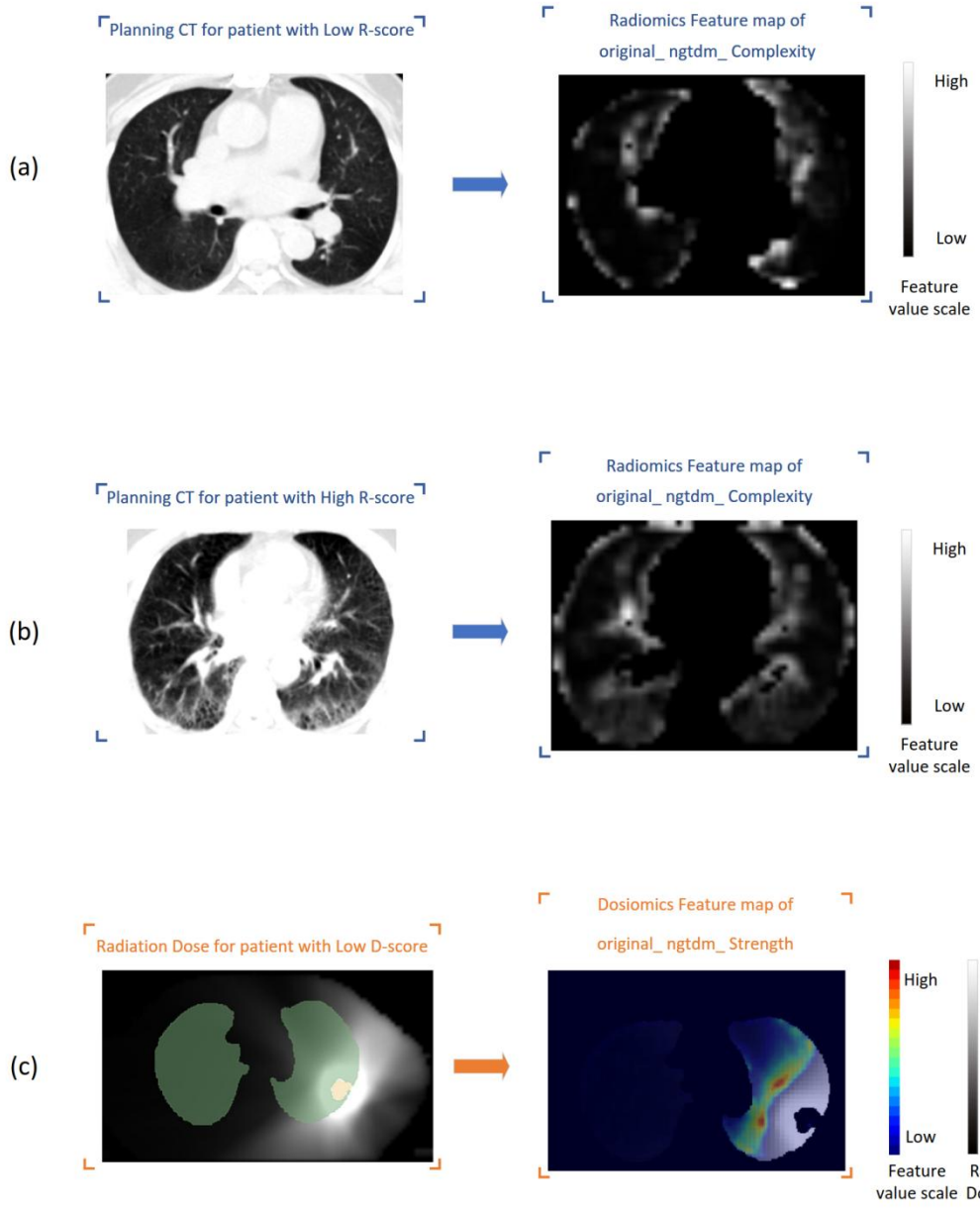

(a)
Planning CT for patient with Low R-score
Radiomics Feature map of
original_ ngtdm_ Complexity
High
Low
Feature
value scale
(b)
Planning CT for patient with High R-score
Radiomics Feature map of
original_ ngtdm_ Complexity
High
Low
Feature
value scale
(c)
Radiation Dose for patient with Low D-score
Dosiomics Feature map of
original_ ngtdm_ Strength
High
Low
Feature
value scale
High
Low
Radiation
Dose scale
(d)
Radiation Dose for patient with High D-score
Dosiomics Feature map of
original_ ngtdm_ Strength
High
Low
Feature
value scale
High
Low
Radiation
Dose scale

Figure 3. (a) The left image is the planning CT image of a patient with a low Radiomics risk score (R-score). The right image is the radiomics feature (original_ngtdm_Complexity) map of CT image at roughly the same level as shown on the left. Feature values are indicated from dark to light.
(b) The left image is the planning CT image of a patient with a high R-score. The right image is the radiomics feature (original_ngtdm_Complexity) map of CT image at roughly the same level as shown on the left.
(c) The left image is the radiation dose (RD) image of a patient with a low Dosiomics risk score (D-score). The right image is the dosiomics feature (original_ngtdm_Strength) map of RD image at roughly the same level as shown on the left. Feature values are represented by rainbow color bar, i.e., from blue to red. The irradiation dose is indicated from dark to light.
(d) The left image is the radiation dose (RD) image of a patient with a high D-score. The right image is the dosiomics feature (original_ngtdm_Strength) map of RD image at roughly the same level as shown on the left.

3. Comparison of discrimination performance of different models
Prediction performance was quantified as AUC for six models and is summarized in Table 2. Other possible combinations of models are provided in the Supplementary material G. The model that yielded the highest AUC was the combination of R-score, D-score, and clinical parameters. The discrimination performances were 0.793 (95%CI 0.735-0.851) and 0.855 (95%CI 0.719-0.99), in the training and prospective validation sets, respectively. As the estimate of the degree of over-optimism (i.e., over-fitting) during model construction, our bootstrap-based validation yielded an AUC of 0.774 (95%CI 0.762-0.786).

Table 2 Discrimination ability of different models according to area under the curve (AUC) with 95%CI provided between parentheses.

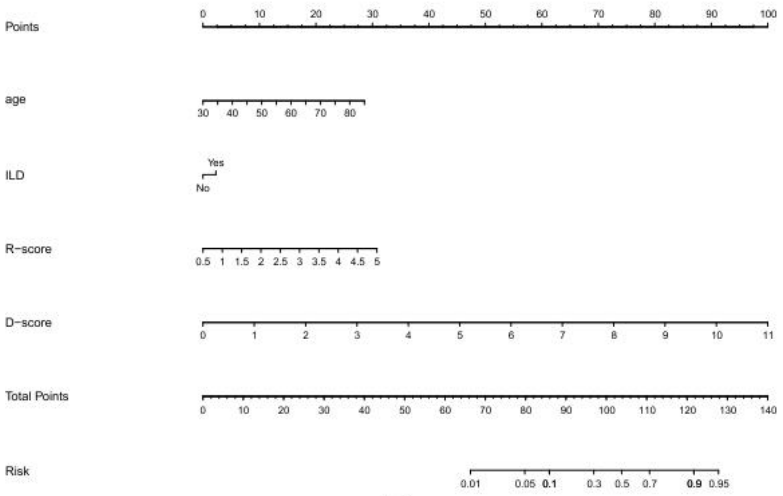
Model	Train (95%CI)	Validation by bootstrapping (95%CI)	Testing (95%CI)
R-score	0.676 (0.606-0.745)	0.619 (0.592-0.646)	0.671 (0.558-0.899)
D-score	0.728 (0.665-0.790)	0.687 (0.667-0.706)	0.684 (0.573-0.883)
DVH-score	0.637	0.628	0.661

	(0.570-0.705)	(0.613-0.642)	(0.551-0.856)
Clinical parameters	0.664 (0.594-0.735)	0.654 (0.628-0.680)	0.709 (0.509-0.91)
R-score + DVH-score + C	0.728 (0.674-0.803)	0.719 (0.703-0.736)	0.782 (0.686-0.832)
R-score + D-score + C	0.793 (0.735-0.851)	0.774 (0.762-0.786)	0.855 (0.719-0.990)

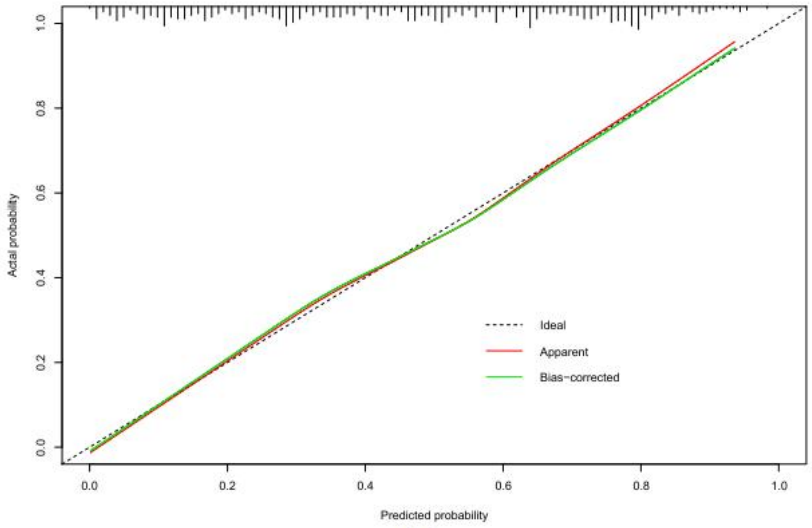
Abbreviations: R = radiomics risk score; D = dosiomics risk score; DVH = dose-volume histogram; C = clinical parameters.

4. Model calibration and decision curve analysis

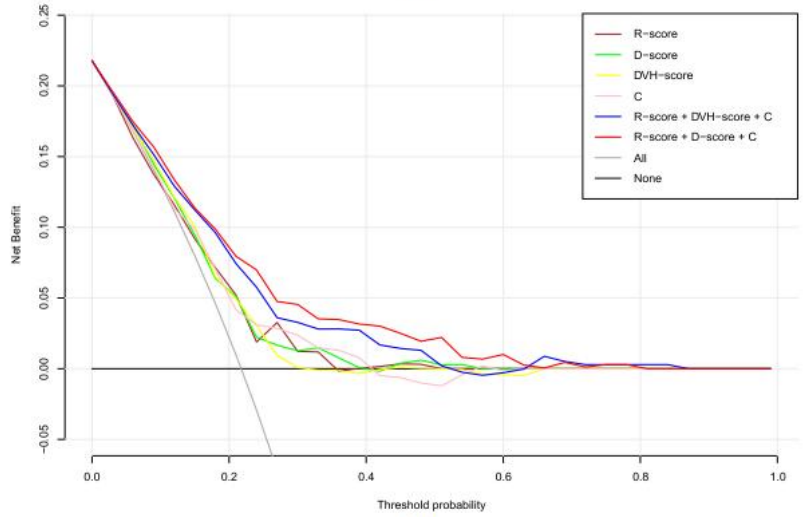
A nomogram based on clinical parameters, R-score, and D-score was constructed and is shown in Figure 4a. The calibration curve of nomogram validated by bootstrap resampling is displayed in Figure 4b, which illustrates good agreement between the predicted probabilities of RP versus the actual observed probabilities. The calibration curve of prospective validation set is provided in the Supplementary Material H with a slope of 1.21 and an intercept of - 0.04. DCA (Figure 4c) showed that the prediction model with the combination of R-score, D-score and clinical parameters has the best positive net benefits at threshold probabilities, implying that a proportion of patients could benefit from using the model to assist in clinical decision making.


Points
0 10 20 30 40 50 60 70 80 90 100
age
30 40 50 60 70 80
ILD
Yes
No
R−score
0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
D−score
0 1 2 3 4 5 6 7 8 9 10 11
Total Points
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Risk
0.01 0.05 0.1 0.3 0.5 0.7 0.9 0.95

(a)


Actal probability
Predicted probability
0.0 0.2 0.4 0.6 0.8 1.0
Ideal
Apparent
Bias−corrected

(b)


Net Benefit
Threshold probability
−0.05 0.00 0.05 0.10 0.15 0.20 0.25
0.0 0.2 0.4 0.6 0.8 1.0
R−score
D−score
DVH−score
C
R−score + DVH−score + C
R−score + D−score + C
All
None

(c)

Figure 4. (a) Nomogram predicting the occurrence of symptom RP. Abbreviations: ILD: Interstitial lung disease; R-score = Radiomics risk score; D-score = Dosiomics risk score. (b) Calibration curve with a bootstrap resampling validation of prediction model combining radiomics risk score, dosiomics risk score, and clinical parameters. Dashed line indicated the ideal model in which predicted and actual probabilities were perfectly identical; Red line indicated actual performance with apparent accuracy; Green line indicated bootstrap corrected estimate of the calibration curve. (c) Decision curve analysis of prediction models. The color lines represent the DCA of different prediction models, the horizontal black line represents the hypothesis that no patients receive interventions, the oblique gray line represents the hypothesis that all patients receive the interventions. Abbreviations: R-score = Radiomics risk score; D-score = Dosiomics risk score; DVH-score = dose-volume histogram score; C. = clinical parameters.

Discussion

Identifying patients at higher risk of developing RP following thoracic irradiation remains an important and topical clinical question, as this adverse event directly affects patient prognosis and reduces quality of life. Patients with RP are a highly heterogeneous group, hence this study evaluated non-invasive methods (radiomics and dosiomics) using only pre-treatment information to characterize individual differences. In this study, the dosiomics features were shown to have stronger predictive power than the conventional DVH parameters, and the combination of a radiomics signature, a dosiomics signature, and two clinical factors were found to be predictive of RP. The results demonstrated that all three types of data appear to carry complementary information relevant to the risk of developing RP. To facilitate further clinical evaluation, we provided a nomogram and discuss the potential clinical benefits of applying the RP predictive model.

Several studies to date have been conducted to predict RP by extracting handcrafted radiomics features from CT. Cunliffe et al. [38] explored the correlation between radiomics and RP and found that 12 radiomics features extracted from CT images of patients with esophageal cancer changed over time in association with the development of RP (AUC=0.78), however, this study focuses on measurement and assessment rather than prediction. Krafft et al. [21] performed an in-depth study for lung cancer and concluded that the best predictive power (AUC=0.68) was achieved when combining radiomics, clinical and dosimetric parameters to build the model. Similar findings were obtained in a study of esophageal cancer by Du et al [20]. They developed a model combining radiomics, clinical and dosimetric parameters by studying 96 patients with esophageal cancer (AUC=0.91). Although these studies included small sample sizes, they inspired us that the combination of handcrafted radiomics features and dosimetric parameters can improve the predictive power of the model. For dosiomics, several studies have demonstrated its potential to predict radiotherapy-related endpoints, including prognosis [39-41] and treatment efficacy [42, 43], but there are very few studies using handcrafted dosiomics to predict side effects. A recent study published by Takanori et al. [25] used a combination of dosiomics and dose-volume indices to predict the occurrence of RP and concluded that dosiomics has the ability to predict RP. Liang et al. [26] conducted a study on dosiomics prediction of RP and

confirmed that dosiomics predictive ability was superior to both dosimetric and NTCP predictors (AUC of 0.78 compared to 0.68 and 0.74), which gives us an idea that dosiomics relative to dosimetrics perhaps possessing more dimensional information.

Based on the results of this study (Table 2) we conclude that the predictive power and stability (with narrower 95%CI) of the model based on dosiomics features is stronger than the model based on dosimetrics. The correlation analysis between dosimetric and D-score showed that they are correlated, where D-score correlates with V30, V25, and V20 between 0.7 and 0.8 (Supplementary material F Figure 3b). Although both dosiomics and dosimetric are quantitative values obtained by calculating from 3D dose distributions, dosiomics obtains more detailed information from texture analysis of the dose distribution, while dosimetric obtains information based on dose-volume histograms. The shape features, which measure the dose delivery from another perspective, may also give a stronger predictive power to the dosiomics. Combining the results of this study and the published dosiomics studies to date, we suggest that neither can replace the other. Inspired by radiomics studies, we resampled the RD images to 1 mm. Different dose grids affect dosiomics feature values [29], however, the utility of resampling RD images, more specifically, whether resampling improves the reproducibility and stability of dosiomics features, requires more research. Placidi et al. conducted a multi-institutional basic study on dosiomics features, which concluded that dosiomics is a tool with predictive potential suitable for multi-institutional studies by analyzing the reproducibility, stability, and sensitivity of dosiomics features [29]. Our results also demonstrate that dosiomics have predictive potential and therefore it is worthwhile to investigate dosiomics more extensively and deeply.

To the best of our knowledge, no previous published studies have combined handcrafted radiomics, dosiomics, and clinical parameters of lung cancer in various ways and compared their ability to predict RP. In this work, we have compared models with radiomics alone, and with 3D spatial dose quantitative features (dosiomics) and we then go beyond current knowledge by proposing a combined model which shows that radiomics and dosiomics are complementary thus leading to improved model performance. We implemented a careful and objective feature selection approach, with robustness as the selection principle for each step of feature selection rather than best predictive ability, which to some extent avoids the occurrence of

chance events. After this, the robust model validation approach was conducted and validated using bootstrap datasets and a prospective dataset, respectively, with over-optimism correction in both ways. Meanwhile, the number of variables in the model was controlled to avoid overfitting. The objective potential of radiomics/dosiomics for predicting RP was explored according to such a process.

We evaluated the performance of the model in three aspects, discrimination ability, calibration, and clinical application potential [44-46]. First, the differences between the training set, bootstrapping set, and test set are satisfactory in the results of discriminative validation, and the fluctuation range of 1000 repetitions is small. Based on this result, we think the model has stable prediction ability and low risk of overfitting. Second, the goodness of fit is another evaluation criterion for the prediction model. The final comprehensive model has excellent calibration, with no significant over- or under-estimation for different risk intervals. Third, a nomogram was built to assist clinical practice, and an offline calculator was provided to facilitate the calculation of R/D-score. The potential of the predictive model for clinical application was also evaluated using DCA. In Figure 4c, it can be seen that the nomogram-based prediction model has positive net benefits. In more detail, the net benefit of the prediction model is greater than the hypothesis that all patients receive RP prophylaxis or pro-active countermeasures (e.g., taking drugs to prevent RP or reducing the dose of radiotherapy) and that all patients do not receive such measures indiscriminately. It is worth noting that the net benefit of the D-score-based model is higher than that of the DVH-score-based model, implying that the model with the D-score has more potential clinical benefit. In summary, the model we developed has potential clinical utility.

In univariate analysis of clinical parameters, whether receiving CCRT had an effect on the occurrence of RP, and the incidence of RP was lower in patients who received CCRT, which is not consistent with clinical experience and with findings in most studies [12, 13, 47]. This might be a bias due to subjective clinical decision making by physicians. Patients included in our study were evaluated by physicians for risk prior to receiving CCRT, and patients with poor health status and high incidence of radiation therapy side effects in the opinion of physicians would not be given CCRT. Some patients will receive potentially lower prescription doses in the radical dose range with stricter dose constraints of the lung to ensure they can complete a full cycle of radiotherapy without

serious radiation therapy side effects. Similar views have been proposed by other researchers [8]. A negative correlation between age and CCRT can be seen in Supplementary material F Figure 3, which also illustrates the subjectivity in the setting of the CCRT protocol. Our findings suggest that ILD is a risk factor for the development of RP. Clinically, RT may lead to exacerbation of ILD and thus interfere with the diagnosis of RP [48]. Accordingly, in this study, the diagnosis of RP in patients with ILD was determined by collaboration with radiologists. And it should be noted that strictly to define, the ILD mentioned in this study is subclinical ILD, according to previous studies. [49, 50]. To investigate the effect of ILD on the model, we excluded patients with ILD in all datasets and performed the same independent validation methods as described previously. Based on the results (Supplementary Material G), we propose our hypothesis: 1. The radiomics model focuses not only on lung texture but also includes other information, as there is no significant difference between the model including or excluding patients with ILD. 2. The discrimination performance of the model built by dosiomics or DVH metrics is improved by excluding patients with ILD, as dose-based models are difficult to predict RP in patients with ILD. 3. ILD is a critical clinical predictor. In previous reports, patients with ILD have high risk of RP, and ILD has been considered a high risk factor for fatal RP [51, 52]. A number of studies have been conducted to analyze the relationship between age and RP [8, 53]. Several studies [54-56] and a meta-analysis [57] have shown that older patients have a higher risk of developing RP. However, some studies did not find an association between age and the risk of RP [58, 59]. In summary, patients who are elderly or/and have ILD should be given more attention and a more comprehensive risk assessment before receiving radiotherapy.

A current challenge in radiomics/dosiomics studies is interpretability, and we attempted to analyze the omics results from a clinical perspective. The analysis revealed no strong correlation between clinical parameters and the R-score (Supplementary material F Figure 3a). However, imaging radiomics contains a large amount of quantitative information and it may not be possible to interpret the full meaning of what it represents using a few clinical parameters. The feature maps of radiomics and dosiomics can provide the direct visualization of voxel-based feature values. As shown in Figure 3 (a) and (b), the radiomic feature "original_ngtdm_Complexity" can reflect the texture characteristics, and for ILD patients, higher voxel-based feature values were obtained compared to patients without pre-existing lung disease. The dosomic

feature "original_ngtdm_Strength" (Figure 3(c) and (d)) shows a pattern of variation from high to low dose, which is some reflection of the radiotherapy planning pattern. Feature maps of other features are provided in the Supplementary material F Figure 5. We compared the feature maps with the follow-up diagnostic CTs and found that the radiomics signature map did not match the areas of symptomatic RP. In contrast, there is a significant overlap between some dosiomics feature maps and the symptomatic RP regions (Supplementary Material F Figure 6). This is consistent with the clinical understanding that the regional localization of symptomatic RP is more closely related to the physical radiation dose distribution.

The Rad/Dosiomics features selected in this study include shape features, which give us a suggestion that the contouring of the lung tissue is important. Currently, manual segmentation is still the "gold standard", but it is time consuming. Therefore, we performed manual check to ensure the accuracy and quality of the automatic segmentation, following processing by the automatic segmentation software. We think this approach is suitable for future multi-institutional studies to assure accuracy while reducing physician workload. Since dosiomics is still relatively little studied, there are no standardized parameter settings yet. Although it has common points with imaging radiomics, some of the parameter settings are different and have a great impact on the results, so we provide the setting files in Supplementary material B, which also provides a reference for future investigators.

This present study has several limitations. First, although the sample size included in our study is relatively large for radiomics/dosiomics RP prediction study, the prospective validation sample size is too small. Our institution's prospective study is still ongoing and continues to expand the sample size. For the scope of this work, we did not yet optimize the plan based on the results of the omics model. We acknowledge that the prospective data set used in this study was derived from an observational prospective study and no interventions were implemented in those patients based on our abovementioned predictive models. By prospective inclusion, we were strictly only able to standardize the follow-up strategy, specifically, patients received regular follow-up examinations and RP grade was jointly diagnosed by the study investigators, which ensured the highest achievable accuracy and consistency of the endpoints, while giving more attention towards patients with likelihood of developing RP. At the present time, it is not yet clear which aspect

of the treatment plan to change in order to intervene correctly in the planning dosimetry process, so this requires further work. A prospectively-enrolled clinical study would be important in the clinical implementation process, this is planned for future work, but is not the principal purpose of this paper. Second, the current gold standard for predictive model validation is still multi-institutional real-world external validation. Third, we built a binary prediction model because the sample size is limited and as the dataset expands, models that can predict different grades are needed. Fourth, pneumonitis associated with immune checkpoint inhibitor (ICI) therapy is an important adverse event. However, the relationship between ICI and RP or the relationship between ICI-associated pneumonitis and radiotherapy-associated pneumonitis remains unclear. Therefore, we excluded patients treated with ICI. Fifth, most current studies comparing machine learning and deep learning conclude that deep learning has stronger predictive power. This study is a pilot study. Therefore, deep learning which is currently a "black box" is not applied, and machine learning with observable processing is chosen. Finally, individualized treatment should incorporate more multidimensional omics information, including genomics and imaging multimodality data. To address several issues above, our institution is conducting a multi-institutional study.

Conclusions

This study was a TRIPOD type 3 prediction model development study, validated using bootstrap samples and a prospective validation set. The radiomics, dosiomics signature, and clinical parameters associated with RP were selected. By comparing the performance of the models built by combining different types of parameters, the best prediction model was found with the best performance of the three types of parameters combined. Furthermore, a comprehensive nomogram was built to assist in clinical decision making and individualized treatment. In the future, a multi-institutional study is needed.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: a cancer journal for clinicians. 2021;71:209-49. doi:10.3322/caac.21660.
2. Yang W-C, Hsu F-M, Yang P-C. Precision radiotherapy for non-small cell lung cancer. J Biomed Sci. 2020;27:82. doi:10.1186/s12929-020-00676-5.
3. Vinod SK, Hau E. Radiotherapy treatment for lung cancer: Current status and future directions. Respirology. 2020;25 Suppl 2:61-71. doi:10.1111/resp.13870.
4. Luo H-S, Huang H-C, Lin L-X. Effect of modern high-dose versus standard-dose radiation in definitive concurrent chemo-radiotherapy on outcome of esophageal squamous cell cancer: a meta-analysis. Radiation Oncology. 2019;14:178. doi:10.1186/s13014-019-1386-x.
5. Ladbury CJ, Rusthoven CG, Camidge DR, Kavanagh BD, Nath SK. Impact of Radiation Dose to the Host Immune System on Tumor Control and Survival for Stage III Non-Small Cell Lung Cancer Treated with Definitive Radiation Therapy. International Journal of Radiation Oncology*Biology*Physics. 2019;105:346-55. doi:10.1016/j.ijrobp.2019.05.064.
6. Kong F-M, Ten Haken RK, Schipper MJ, Sullivan MA, Chen M, Lopez C, et al. High-dose radiation improved local tumor control and overall survival in patients with inoperable/unresectable non-small-cell lung cancer: long-term results of a radiation dose escalation study. International Journal of Radiation Oncology, Biology, Physics. 2005;63:324-33. doi:10.1016/j.ijrobp.2005.02.010.
7. Niu S, Zhang Y. Applications and therapeutic mechanisms of action of mesenchymal stem cells in radiation-induced lung injury. Stem Cell Res Ther. 2021;12:212. doi:10.1186/s13287-021-02279-9.
8. Ullah T, Patel H, Pena GM, Shah R, Fein AM. A contemporary review of radiation pneumonitis. Curr Opin Pulm Med. 2020;26:321-5. doi:10.1097/MCP.0000000000000682.
9. Käsmann L, Dietrich A, Staab-Weijnitz CA, Manapov F, Behr J, Rimner A, et al. Radiation-induced lung toxicity - cellular and molecular mechanisms of pathogenesis, management, and literature review. Radiation Oncology (London, England). 2020;15:214. doi:10.1186/s13014-020-01654-9.
10. Thomas R, Chen Y-H, Hatabu H, Mak RH, Nishino M. Radiographic patterns of symptomatic radiation pneumonitis in lung cancer patients:

Imaging predictors for clinical severity and outcome. Lung Cancer. 2020;145:132-9. doi:10.1016/j.lungcan.2020.03.023.
11. Nalbantov G, Kietselaer B, Vandecasteele K, Oberije C, Berbee M, Troost E, et al. Cardiac comorbidity is an independent risk factor for radiation-induced lung toxicity in lung cancer patients. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2013;109:100-6. doi:10.1016/j.radonc.2013.08.035.
12. Arroyo-Hernández M, Maldonado F, Lozano-Ruiz F, Muñoz-Montaño W, Nuñez-Baez M, Arrieta O. Radiation-induced lung injury: current evidence. BMC Pulm Med. 2021;21:9. doi:10.1186/s12890-020-01376-4.
13. Kong F-MS, Wang S. Nondosimetric risk factors for radiation-induced lung toxicity. Seminars in Radiation Oncology. 2015;25:100-9. doi:10.1016/j.semradonc.2014.12.003.
14. Huang Q, Xie F, Ouyang X. Predictive SNPs for radiation-induced damage in lung cancer patients with radiotherapy: a potential strategy to individualize treatment. Int J Biol Markers. 2015;30:e1-11. doi:10.5301/jbm.5000108.
15. Niu X, Li H, Chen Z, Liu Y, Kan M, Zhou D, et al. A study of ethnic differences in TGFβ1 gene polymorphisms and effects on the risk of radiation pneumonitis in non-small-cell lung cancer. Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer. 2012;7:1668-75. doi:10.1097/JTO.0b013e318267cf5b.
16. Yu H, Wu H, Wang W, Jolly S, Jin J-Y, Hu C, et al. Machine Learning to Build and Validate a Model for Radiation Pneumonitis Prediction in Patients with Non-Small Cell Lung Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2019;25:4343-50. doi:10.1158/1078-0432.CCR-18-1084.
17. Mak RH, Alexander BM, Asomaning K, Heist RS, Liu C-y, Su L, et al. A single-nucleotide polymorphism in the methylene tetrahydrofolate reductase (MTHFR) gene is associated with risk of radiation pneumonitis in lung cancer patients treated with thoracic radiation therapy. Cancer. 2012;118:3654-65. doi:10.1002/cncr.26667.
18. Bodalal Z, Trebeschi S, Nguyen-Kim TDL, Schats W, Beets-Tan R. Radiogenomics: bridging imaging and genomics. Abdominal Radiology (New York). 2019;44:1960-84. doi:10.1007/s00261-019-02028-w.
19. Wang L, Gao Z, Li C, Sun L, Li J, Yu J, et al. Computed tomography-based delta-radiomics analysis for discriminating radiation pneumonitis in

patients with esophageal cancer after radiation therapy. International Journal of Radiation Oncology, Biology, Physics. 2021. doi:10.1016/j.ijrobp.2021.04.047.
20. Du F, Tang N, Cui Y, Wang W, Zhang Y, Li Z, et al. A Novel Nomogram Model Based on Cone-Beam CT Radiomics Analysis Technology for Predicting Radiation Pneumonitis in Esophageal Cancer Patients Undergoing Radiotherapy. Front Oncol. 2020;10:596013. doi:10.3389/fonc.2020.596013.
21. Krafft SP, Rao A, Stingo F, Briere TM, Court LE, Liao Z, et al. The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis. Med Phys. 2018;45:5317-24. doi:10.1002/mp.13150.
22. Liu Y, Wang W, Shiue K, Yao H, Cerra-Franco A, Shapiro RH, et al. Risk factors for symptomatic radiation pneumonitis after stereotactic body radiation therapy (SBRT) in patients with non-small cell lung cancer. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;156:231-8. doi:10.1016/j.radonc.2020.10.015.
23. Saha A, Beasley M, Hatton N, Dickinson P, Franks K, Clarke K, et al. Clinical and dosimetric predictors of radiation pneumonitis in early-stage lung cancer treated with Stereotactic Ablative radiotherapy (SABR) - An analysis of UK's largest cohort of lung SABR patients. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;156:153-9. doi:10.1016/j.radonc.2020.12.015.
24. Bourbonne V, Da-Ano R, Jaouen V, Lucia F, Dissaux G, Bert J, et al. Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;155:144-50. doi:10.1016/j.radonc.2020.10.040.
25. Adachi T, Nakamura M, Shintani T, Mitsuyoshi T, Kakino R, Ogata T, et al. Multi-institutional dose-segmented dosiomic analysis for predicting radiation pneumonitis after lung stereotactic body radiation therapy. Med Phys. 2021;48:1781-91. doi:10.1002/mp.14769.
26. Liang B, Yan H, Tian Y, Chen X, Yan L, Zhang T, et al. Dosiomics: Extracting 3D Spatial Features From Dose Distribution to Predict Incidence of Radiation Pneumonitis. Front Oncol. 2019;9:269. doi:10.3389/fonc.2019.00269.
27. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015;162:W1-W73. doi:10.7326/M14-0698.

28. Matschinske J, Alcaraz N, Benis A, Golebiewski M, Grimm DG, Heumos L, et al. The AIMe registry for artificial intelligence in biomedical research. Nat Methods. 2021;18:1128-31. doi:10.1038/s41592-021-01241-0.
29. Placidi L, Gioscio E, Garibaldi C, Rancati T, Fanizzi A, Maestri D, et al. A Multicentre Evaluation of Dosiomics Features Reproducibility, Stability and Sensitivity. Cancers (Basel). 2021;13:3835. doi:10.3390/cancers13153835.
30. Adachi T, Nakamura M, Kakino R, Hirashima H, Iramina H, Tsuruta Y, et al. Dosiomic feature comparison between dose-calculation algorithms used for lung stereotactic body radiation therapy. Radiol Phys Technol. 2022. doi:10.1007/s12194-022-00651-9.
31. Shi Z, Traverso A, van Soest J, Dekker A, Wee L. Technical Note: Ontology-guided radiomics analysis workflow (O-RAW). Med Phys. 2019;46:5677-84. doi:10.1002/mp.13844.
32. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77:e104-e7. doi:10.1158/0008-5472.CAN-17-0339.
33. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295:328-38. doi:10.1148/radiol.2020191145.
34. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magnetic Resonance Imaging. 2012;30:1323-41. doi:10.1016/j.mri.2012.05.001.
35. Larue RTHM, van Timmeren JE, de Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncologica. 2017;56:1544-53. doi:10.1080/0284186X.2017.1351624.
36. Compter I, Verduin M, Shi Z, Woodruff HC, Smeenk RJ, Rozema T, et al. Deciphering the glioblastoma phenotype by computed tomography radiomics. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;160:132-9. doi:10.1016/j.radonc.2021.05.002.
37. Shi Z, Zhang Z, Liu Z, Zhao L, Ye Z, Dekker A, et al. Methodological quality of machine learning-based quantitative imaging analysis studies in esophageal cancer: a systematic review of clinical outcome prediction after concurrent chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2021. doi:10.1007/s00259-021-05658-9.

38. Cunliffe A, Armato SG, Castillo R, Pham N, Guerrero T, Al-Hallaq HA. Lung Texture in Serial Thoracic Computed Tomography Scans: Correlation of Radiomics-based Features With Radiation Therapy Dose and Radiation Pneumonitis Development. International Journal of Radiation Oncology*Biology*Physics. 2015;91:1048-56. doi:10.1016/j.ijrobp.2014.11.030.
39. Rossi L, Bijman R, Schillemans W, Aluwini S, Cavedon C, Witte M, et al. Texture analysis of 3D dose distributions for predictive modelling of toxicity rates in radiotherapy. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2018;129:548-53. doi:10.1016/j.radonc.2018.07.027.
40. Gabryś HS, Buettner F, Sterzing F, Hauswald H, Bangert M. Design and Selection of Machine Learning Methods Using Radiomics and Dosiomics for Normal Tissue Complication Probability Modeling of Xerostomia. Front Oncol. 2018;8:35. doi:10.3389/fonc.2018.00035.
41. Zhen X, Chen J, Zhong Z, Hrycushko B, Zhou L, Jiang S, et al. Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. Phys Med Biol. 2017;62:8246-63. doi:10.1088/1361-6560/aa8d09.
42. Buizza G, Paganelli C, D'Ippolito E, Fontana G, Molinelli S, Preda L, et al. Radiomics and Dosiomics for Predicting Local Control after Carbon-Ion Radiotherapy in Skull-Base Chordoma. Cancers (Basel). 2021;13. doi:10.3390/cancers13020339.
43. Wu A, Li Y, Qi M, Lu X, Jia Q, Guo F, et al. Dosiomics improves prediction of locoregional recurrence for intensity modulated radiotherapy treated head and neck cancer cases. Oral Oncol. 2020;104:104625. doi:10.1016/j.oraloncology.2020.104625.
44. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128-38. doi:10.1097/EDE.0b013e3181c30fb2.
45. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. JAMA. 2017;318:1377. doi:10.1001/jama.2017.12126.
46. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2016;34:2534-40. doi:10.1200/JCO.2015.65.5654.

47. Bledsoe TJ, Nath SK, Decker RH. Radiation Pneumonitis. Clinics in Chest Medicine. 2017;38:201-8. doi:10.1016/j.ccm.2016.12.004.
48. Kocak Z, Evans ES, Zhou S-M, Miller KL, Folz RJ, Shafman TD, et al. Challenges in defining radiation pneumonitis in patients with lung cancer. International Journal of Radiation Oncology, Biology, Physics. 2005;62:635-8. doi:10.1016/j.ijrobp.2004.12.023.
49. Yamaguchi S, Ohguri T, Matsuki Y, Yahara K, Oki H, Imada H, et al. Radiotherapy for thoracic tumors: association between subclinical interstitial lung disease and fatal radiation pneumonitis. Int J Clin Oncol. 2015;20:45-52. doi:10.1007/s10147-014-0679-1.
50. Doyle TJ, Hunninghake GM, Rosas IO. Subclinical interstitial lung disease: why you should care. Am J Respir Crit Care Med. 2012;185:1147-53. doi:10.1164/rccm.201108-1420PP.
51. Doi H, Nakamatsu K, Nishimura Y. Stereotactic body radiotherapy in patients with chronic obstructive pulmonary disease and interstitial pneumonia: a review. Int J Clin Oncol. 2019;24:899-909. doi:10.1007/s10147-019-01432-y.
52. Okumura M, Hojo H, Nakamura M, Hiyama T, Nakamura N, Zenda S, et al. Radiation pneumonitis after palliative radiotherapy in cancer patients with interstitial lung disease. Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology. 2021;161:47-54. doi:10.1016/j.radonc.2021.05.026.
53. Giuranno L, Ient J, De Ruysscher D, Vooijs MA. Radiation-Induced Lung Injury (RILI). Front Oncol. 2019;9:877. doi:10.3389/fonc.2019.00877.
54. Leprieur EG, Fernandez D, Chatellier G, Klotz S, Giraud P, Durdux C. Acute radiation pneumonitis after conformational radiotherapy for nonsmall cell lung cancer: clinical, dosimetric, and associated-treatment risk factors. J Cancer Res Ther. 2013;9:447-51. doi:10.4103/0973-1482.119339.
55. Dang J, Li G, Zang S, Zhang S, Yao L. Risk and predictors for early radiation pneumonitis in patients with stage III non-small cell lung cancer treated with concurrent or sequential chemoradiotherapy. Radiation Oncology (London, England). 2014;9:172. doi:10.1186/1748-717X-9-172.
56. Tsujino K, Hashimoto T, Shimada T, Yoden E, Fujii O, Ota Y, et al. Combined analysis of V20, VS5, pulmonary fibrosis score on baseline computed tomography, and patient age improves prediction of severe radiation pneumonitis after concurrent chemoradiotherapy for locally advanced non-small-cell lung cancer. Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer. 2014;9:983-90. doi:10.1097/JTO.0000000000000187.

57. Vogelius IR, Bentzen SM. A literature-based meta-analysis of clinical risk factors for development of radiation induced pneumonitis. Acta Oncologica (Stockholm, Sweden). 2012;51:975-83. doi:10.3109/0284186X.2012.718093.
58. Wen J, Liu H, Wang Q, Liu Z, Li Y, Xiong H, et al. Genetic variants of the LIN28B gene predict severe radiation pneumonitis in patients with non-small cell lung cancer treated with definitive radiation therapy. European Journal of Cancer (Oxford, England: 1990). 2014;50:1706-16. doi:10.1016/j.ejca.2014.03.008.
59. Bradley JD, Hope A, El Naqa I, Apte A, Lindsay PE, Bosch W, et al. A nomogram to predict radiation pneumonitis, derived from a combined analysis of RTOG 9311 and institutional data. International Journal of Radiation Oncology, Biology, Physics. 2007;69:985-92. doi:10.1016/j.ijrobp.2007.04.077.

Supplementary Materials

Supplementary material A

Inclusion and exclusion criteria for retrospective and prospective data

The first dataset was collected retrospectively as a training set and validation set. A total of 314 patients treated between January 2013 and December 2018 with definitive RT at Anonymized for Review hospital were considered for the retrospective dataset. The inclusion criteria were as follows: (1) Patients identified with histologically confirmed NSCLC or SCLC. (2) Diagnosed with Stage I-III NSCLC and limited-stage SCLC (American Joint Committee on Cancer, 8th edition, 2017) before RT, and patients underwent radical RT. (3) No thoracic RT or thoracic surgery prior to RT. (4) CT examinations were performed at 1, 3, and 6 months (± 15 days) after treatment at Anonymized for Review Hospital. Patients were excluded, if treatment break of more than 5 days during RT, if patients received surgical treatment within 6 months after radiotherapy, if patients received adjuvant/concurrent immunotherapy, if there was also a second primary tumor, and if the patients had a lung infection within 6 months after radiotherapy, so it was difficult to identify whether it was RP.

The second dataset was collected prospectively at the same institution as a test set. A total of 56 patients were enrolled in the study from October 2018 to March 2019. Finally, 35 patients were included in the analysis. 21 patients were excluded because did not meet the eligible criteria, fourteen of which did not follow up CT as planned, six of which did not complete radiotherapy, and one patient died two months after radiation therapy. The inclusion and exclusion criteria were the same as the retrospective dataset and these patients were followed-up every month after had received radiotherapy. The follow-up items included blood routine examination, C-reactive protein, tumor markers associated with lung cancer, chest X-rays, and patients received CT examination at 1, 3, and 6 months (± 7 days) after radiotherapy.

Patient Characteristics for prospective data

Supplementary Table 1. Patient Characteristics for prospective data

Characteristics	Pros pts	Without RP2	With RP2	P*

	n (%)	Mean ± SD	Mean ± SD	
Age median	62 (34-75)	61.5 (34-75)	62 (59-68)	0.363
Gender				1.000
Male	23 (65.7%)	17 (73.9%)	6 (26.1%)	
Female	12 (34.3%)	9 (75.0%)	3 (25.0%)	
Smoking				1.000
Yes	26 (74.3%)	19 (73.1%)	7 (26.9%)	
No	9 (25.7%)	7 (77.8%)	2 (22.2%)	
KPS				1.000
≤80	13 (37.1%)	10 (76.9%)	3 (23.1%)	
>80	22 (62.9%)	16 (72.7%)	6 (27.3%)	
Diabetes				1.000
Yes	2 (5.7%)	2 (100.0%)	0 (0%)	
No	33 (94.3%)	24 (72.7%)	9 (27.3%)	
ILD				0.192
Yes	9 (25.7%)	5 (55.6%)	4 (44.4%)	
No	26 (74.3%)	21 (80.8%)	5 (19.2%)	
Pathology				0.776
LUSC	8 (22.9%)	5 (62.5%)	3 (37.5%)	
LUAD	10 (28.6%)	8 (80.0%)	2 (20.0%)	
SCLC	17 (48.5%)	13 (76.5%)	4 (23.5%)	
Induc chemo				0.553
Yes	31 (88.6%)	22 (71.0%)	9 (29.0%)	
No	4 (11.4%)	4 (100.0%)	0 (0%)	
CCRT				0.081
Yes	8 (22.9%)	8 (100.0%)	0 (0%)	
No	27 (77.1%)	18 (66.7%)	9 (33.3%)	
Conso chemo				0.245
Yes	19 (54.3%)	16 (84.2%)	3 (15.8%)	
No	16 (45.7%)	10 (62.5%)	6 (37.5%)	
PGTV (Gy)	60.200±2.870	60.423±2.862	59.556±2.963	0.436
Smoking index	668.600±550.412	646.154±566.555	733.333±527.376	0.679

Abbreviations: Retro = retrospective; Pts = patients; Pros = prospective; LUSC = lung squamous cell carcinoma; LUAD = lung adenocarcinoma; SCLC = small cell lung cancer; IMRT = intensity-modulated radiotherapy; VMAT = volumetric modulated arc therapy; chemo = chemotherapy; KPS = Karnofsky performance score; Induc chemo = induction chemotherapy; CCRT =

concurrent chemoradiotherapy; Conso chemo = consolidation chemotherapy; PGTV = planning gross tumor volume.
*The differences in characteristics were evaluated by logistic regression for continuous variables or Pearson X2 test and exact Fisher test for categorical variables

Supplementary material B

Radiomics and dosiomics features extraction parameter settings file
Radiomics
imageType:
  Original:
    binWidth: 25
featureClass:

  shape: # Remove VoxelVolume, correlated to Volume
    - Elongation
    - Flatness
    - LeastAxisLength
    - MajorAxisLength
    - Maximum2DDiameterColumn
    - Maximum2DDiameterRow
    - Maximum2DDiameterSlice
    - Maximum3DDiameter
    - MeshVolume
    - MinorAxisLength
    - Sphericity
    - SurfaceArea
    - SurfaceVolumeRatio
  firstorder: # Remove Total Energy, correlated to Energy (due to resampling enabled)
    - 10Percentile
    - 90Percentile
    - Energy
    - Entropy
    - InterquartileRange
    - Kurtosis

- Maximum
- Mean
- MeanAbsoluteDeviation
- Median
- Minimum
- Range
- RobustMeanAbsoluteDeviation
- RootMeanSquared
- Skewness
- Uniformity
- Variance

glcm: # Disable SumAverage by specifying all other GLCM features available

- 'Autocorrelation'
- 'JointAverage'
- 'ClusterProminence'
- 'ClusterShade'
- 'ClusterTendency'
- 'Contrast'
- 'Correlation'
- 'DifferenceAverage'
- 'DifferenceEntropy'
- 'DifferenceVariance'
- 'JointEnergy'
- 'JointEntropy'
- 'Imc1'
- 'Imc2'
- 'Idm'
- 'Idmn'
- 'Id'
- 'Idn'
- 'InverseVariance'
- 'MaximumProbability'
- 'SumEntropy'
- 'SumSquares'

glrlm:

glszm:

gldm:

```
  ngtdm:

setting:
  interpolator: 'sitkBSpline'
  resampledPixelSpacing: [2, 2, 2]
  padDistance: 10    # Extra padding for large sigma valued LoG filtered images

  resegmentRange: [-3, 3]
  resegmentMode: sigma

  voxelArrayShift: 1000    # Minimum value in HU is -1000, shift +1000 to prevent negative values from being squared.
```

Dosiomics

```
imageType:
  Original:
    binWidth: 0.5
featureClass:

  shape: # Remove VoxelVolume, correlated to Volume
    - Elongation
    - Flatness
    - LeastAxisLength
    - MajorAxisLength
    - Maximum2DDiameterColumn
    - Maximum2DDiameterRow
    - Maximum2DDiameterSlice
    - Maximum3DDiameter
    - MeshVolume
    - MinorAxisLength
    - Sphericity
    - SurfaceArea
    - SurfaceVolumeRatio
  firstorder: # Remove Total Energy, correlated to Energy    (due to resampling enabled)
    - 10Percentile
    - 90Percentile
    - Energy
```

- Entropy
- InterquartileRange
- Kurtosis
- Maximum
- Mean
- MeanAbsoluteDeviation
- Median
- Minimum
- Range
- RobustMeanAbsoluteDeviation
- RootMeanSquared
- Skewness
- Uniformity
- Variance

glcm: # Disable SumAverage by specifying all other GLCM features available

- 'Autocorrelation'
- 'JointAverage'
- 'ClusterProminence'
- 'ClusterShade'
- 'ClusterTendency'
- 'Contrast'
- 'Correlation'
- 'DifferenceAverage'
- 'DifferenceEntropy'
- 'DifferenceVariance'
- 'JointEnergy'
- 'JointEntropy'
- 'Imc1'
- 'Imc2'
- 'Idm'
- 'Idmn'
- 'Id'
- 'Idn'
- 'InverseVariance'
- 'MaximumProbability'
- 'SumEntropy'
- 'SumSquares'

```
  glrlm:
  glszm:
  gldm:
  ngtdm:

setting:
interpolator: 'sitkBSpline'
  resampledPixelSpacing: [2, 2, 2]
  padDistance: 10    # Extra padding for large sigma valued LoG filtered images

  voxelArrayShift: 0    # Minimum value in HU is -1000, shift +1000 to prevent
negative values from being squared.
```

Supplementary material C

Dose-volume histogram (DVH) metrics selection and model construction

Due to the colinearity of DVH metrics, it does is not suitable to perform the same feature selection approaches as radiomics/dosiomics. Instead, the predictive model is built using the already acknowledged metrics V20 and MLD. The DVH-score was defined as the linear predictor of the multivariable LR radiomics model.

The validation method was performed in exactly the same way as for the radiomics/dosiomics model: (1) For each of the 1000 bootstraps, we fitted the logistic regression model coefficients on each bootstrap, and then computed its Area under the curve (AUC) of receiver operating characteristic curve (ROC) using the original non-bootstrapped development cohort. From these 1000 bootstraps, we computed the average AUC and its 95% confidence interval. (2) As external validation, we evaluated the DVH model using the prospectively-registered cohort of 35 subjects. Processing of these 35 subjects followed exactly the same procedure as for the model development cohort, and none of these subjects were used in any way during model construction.

Since V5 is an important predictor in the IMRT/VMAT era, we also built another DVH model by combining V5 and MLD. However, based on this dataset, the predictive power of the "V5+MLD" model is worse than that of the

"V20+MLD" model, so we used the DVH model of V20 and MLD as the comparative model in this study.

Supplementary material D

Feature selection results and graphs

The top twenty features that were screened are displayed in **Supplementary Table 1.** The features are sorted according to the number of frequencies selected and shown in the **Supplementary Figure 1**. The cut-off points were decided based on the frequency breakpoints shown in the graphs. The cut-off points for both radiomics and dosiomics features are around 600.

The three most frequently selected signatures are shown in **Supplementary Table 2**, with the highest selected frequencies of 45 and 105 for the radiomics and dosiomics signatures, respectively.

Definitions of the selected features are provided in **Supplementary Table 3.**

Supplementary Table 1a. The top twenty radiomics features that were selected

No.	Radiomics features	Frequency
1.	original_shape_Elongation	1000
2.	original_shape_Flatness	922
3.	original_shape_MinorAxisLength	871
4.	original_shape_MeshVolume	746
5.	original_firstorder_90Percentile	700
6.	original_glcm_JointEntropy	696
7.	original_ngtdm_Complexity	694
8.	original_firstorder_Median	684
9.	original_shape_Maximum2DDiameterSlice	677
10.	original_glszm_LargeAreaEmphasis	670
11.	original_shape_Maximum2DDiameterRow	663
12.	original_gldm_DependenceNonUniformityNormalized	603
13.	original_gldm_DependenceEntropy	587
14.	original_glcm_DifferenceEntropy	563
15.	original_ngtdm_Contrast	562
16.	original_glszm_SmallAreaLowGrayLevelEmphasis	542
17.	original_gldm_SmallDependenceLowGrayLevelEmphasis	537

18.	original_shape_LeastAxisLength	525
19.	original_shape_SurfaceVolumeRatio	525
20.	original_ngtdm_Strength	521

Supplementary Table 1b. The top twenty dosiomics features that were selected

No.	dosiomics features	Frequency
1.	original_shape_Elongation	1000
2.	original_glszm_LargeAreaEmphasis	864
3.	original_shape_Flatness	736
4.	original_ngtdm_Strength	715
5.	original_shape_SurfaceArea	704
6.	original_shape_MeshVolume	693
7.	original_shape_Maximum2DDiameterRow	643
8.	original_glszm_GrayLevelVariance	642
9.	original_shape_MinorAxisLength	641
10.	original_ngtdm_Coarseness	622
11.	original_ngtdm_Contrast	605
12.	original_glszm_SmallAreaLowGrayLevelEmphasis	594
13.	original_gldm_LargeDependenceEmphasis	555
14.	original_shape_LeastAxisLength	554
15.	original_glcm_DifferenceEntropy	545
16.	original_glrlm_ShortRunLowGrayLevelEmphasis	544
17.	original_glszm_ZoneEntropy	544
18.	original_glrlm_RunLengthNonUniformity	537
19.	original_glszm_SmallAreaHighGrayLevelEmphasis	526
20.	original_gldm_DependenceEntropy	523

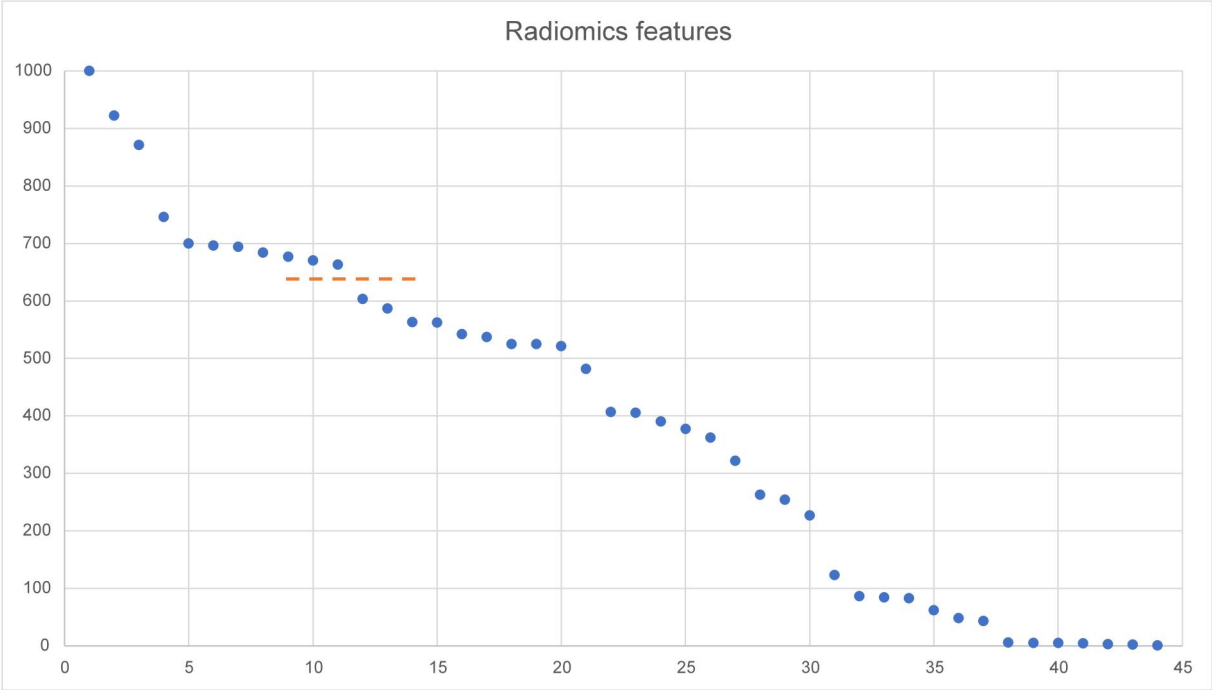
Supplementary Table 2a. The top three frequently selected radiomics signatures

No	Signature	Freq
1.	original_glcm_JointEntropy + original_ngtdm_Complexity + original_shape_Elongation + original_shape_Flatness + original_shape_Maximum2DDiameterSlice + original_shape_MeshVolume + original_shape_MinorAxisLength	45
2.	original_firstorder_90Percentile + original_firstorder_Median + original_glcm_JointEntropy + original_ngtdm_Complexity +	40

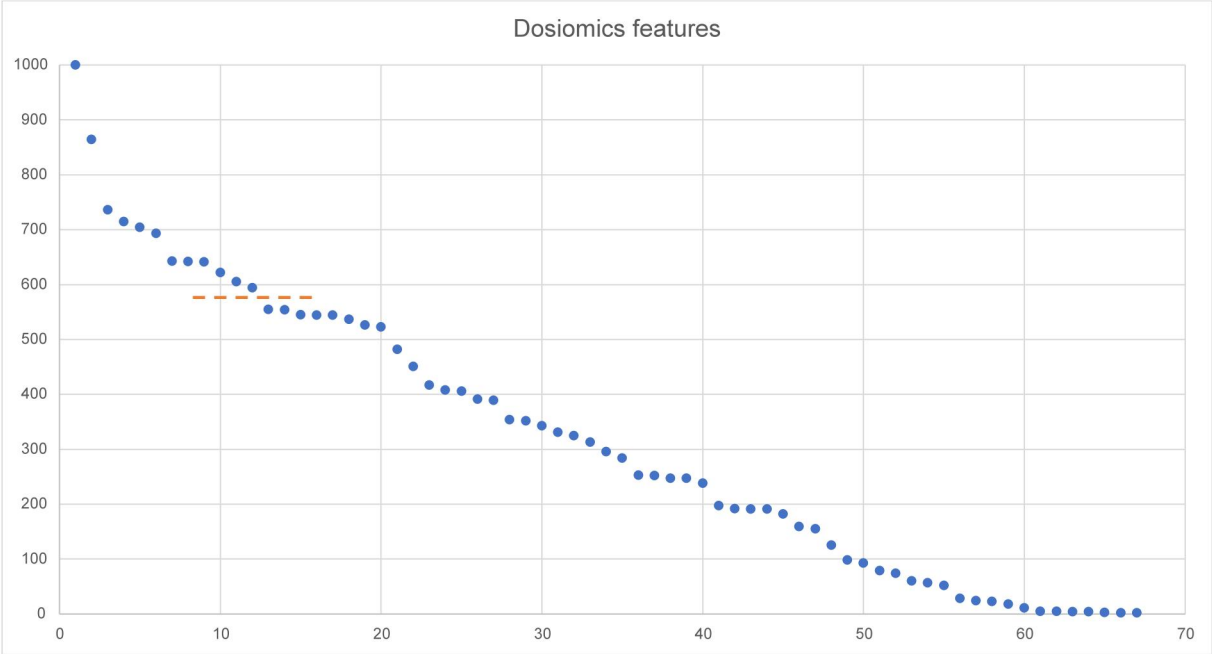
	original_shape_Elongation + original_shape_Flatness + original_shape_MeshVolume + original_shape_MinorAxisLength	
3.	original_firstorder_90Percentile + original_firstorder_Median + original_glszm_LargeAreaEmphasis + original_ngtdm_Complexity + original_shape_Elongation + original_shape_Flatness + original_shape_Maximum2DDiameterRow + original_shape_MeshVolume + original_shape_MinorAxisLength	36

Supplementary Table 2b. The top three frequently selected dosiomics signatures

No	Signature	Freq
1.	original_glszm_GrayLevelVariance + original_glszm_LargeAreaEmphasis + original_ngtdm_Contrast + original_ngtdm_Strength + original_shape_MeshVolume + original_shape_SurfaceArea	105
2.	original_glszm_GrayLevelVariance + original_glszm_LargeAreaEmphasis + original_ngtdm_Contrast + original_ngtdm_Strength + original_shape_MeshVolume	70
3.	original_glszm_GrayLevelVariance + original_glszm_LargeAreaEmphasis + original_ngtdm_Contrast + original_ngtdm_Strength + original_shape_MeshVolume + original_shape_MinorAxisLength + original_shape_SurfaceArea	43



(a)



(b)

Supplementary Figure 1. (a) The radiomics features are sorted according to the number of frequencies selected. (b) The dosiomics features are sorted according to the number of frequencies selected.

Supplementary Table 3. Definitions of the selected features.

Feature	Definition
original_glcm_JointEntropy	Joint entropy is a measure of the randomness/variability in neighborhood intensity values.
original_ngtdm_Complexity	An image is considered complex when there are many primitive components in the image, i.e. the image is non-uniform and there are many rapid changes in gray level intensity.
original_shape_Elongation	Elongation shows the relationship between the two largest principal components in the ROI shape. For computational reasons, this feature is defined as the inverse of

	true elongation.
original_shape_Flatness	Flatness shows the relationship between the largest and smallest principal components in the ROI shape. For computational reasons, this feature is defined as the inverse of true flatness.
original_shape_Maximum2DDiameterSlice	Maximum 2D diameter (Slice) is defined as the largest pairwise Euclidean distance between tumor surface mesh vertices in the row-column (generally the axial) plane.
original_shape_MeshVolume	The volume of the ROI V is calculated from the triangle mesh of the ROI.
original_shape_MinorAxisLength	This feature yield the second-largest axis length of the ROI-enclosing ellipsoid and is calculated using the largest principal component λminor.
original_glszm_GrayLevelVariance	GLV measures the variance in gray level intensities for the zones.
original_glszm_LargeAreaEmphasis	LAE is a measure of the distribution of large area size zones, with a greater value indicative of more larger size zones and more coarse textures.
original_ngtdm_Contrast	Contrast is a measure of the spatial intensity change, but is also dependent on the overall gray level dynamic range. Contrast is high when both the dynamic range and the spatial change rate are high, i.e. an image with a large range of gray levels, with large changes between voxels and their neighbourhood.

original_ngtdm_Strength	Strength is a measure of the primitives in an image. Its value is high when the primitives are easily defined and visible, i.e. an image with slow change in intensity but more large coarse differences in gray level intensities.
original_shape_SurfaceArea	To calculate the surface area, first the surface area of each triangle in the mesh is calculated (1). The total surface area is then obtained by taking the sum of all calculated sub-areas (2).

Radiomics (R)-score and Dosiomics (D)-score

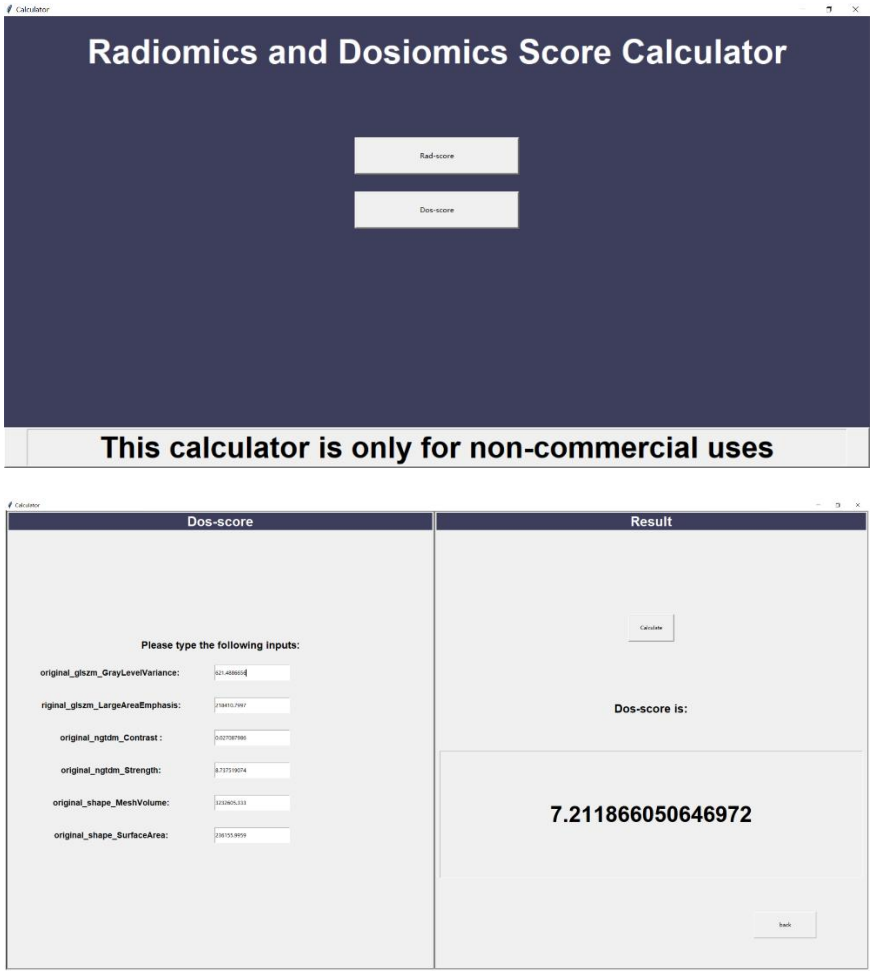
We added a constant offset in order to return strictly positive scores.
The R-score was calculated as follows: -1.383 + 1.067*original_glcm_JointEntropy - 0.370*original_ngtdm_Complexity + 1.605*original_shape_Elongation - 0.635*original_shape_Flatness + 0.398*original_shape_Maximum2DDiameterSlice + 1.557* original_shape_MeshVolume - 2.148*original_shape_MinorAxisLength + 4.
The D-score: -1.522 - 0.616*original_glszm_GrayLevelVariance - 0.868*original_glszm_LargeAreaEmphasis + 0.878*original_ngtdm_Contrast + 0.922*original_ngtdm_Strength + 1.457*original_shape_MeshVolume - 0.625*original_shape_SurfaceArea + 9.

Supplementary material E

Instructions for R-score and D-score calculator

Selecting either Radiomics risk score (R-score) or Dosiomics risk score (D-score), then enter the feature values into the corresponding input boxes and click the “Calculate” button to get the scores.

**This calculator can only be used for research purposes, not for commercial use.*

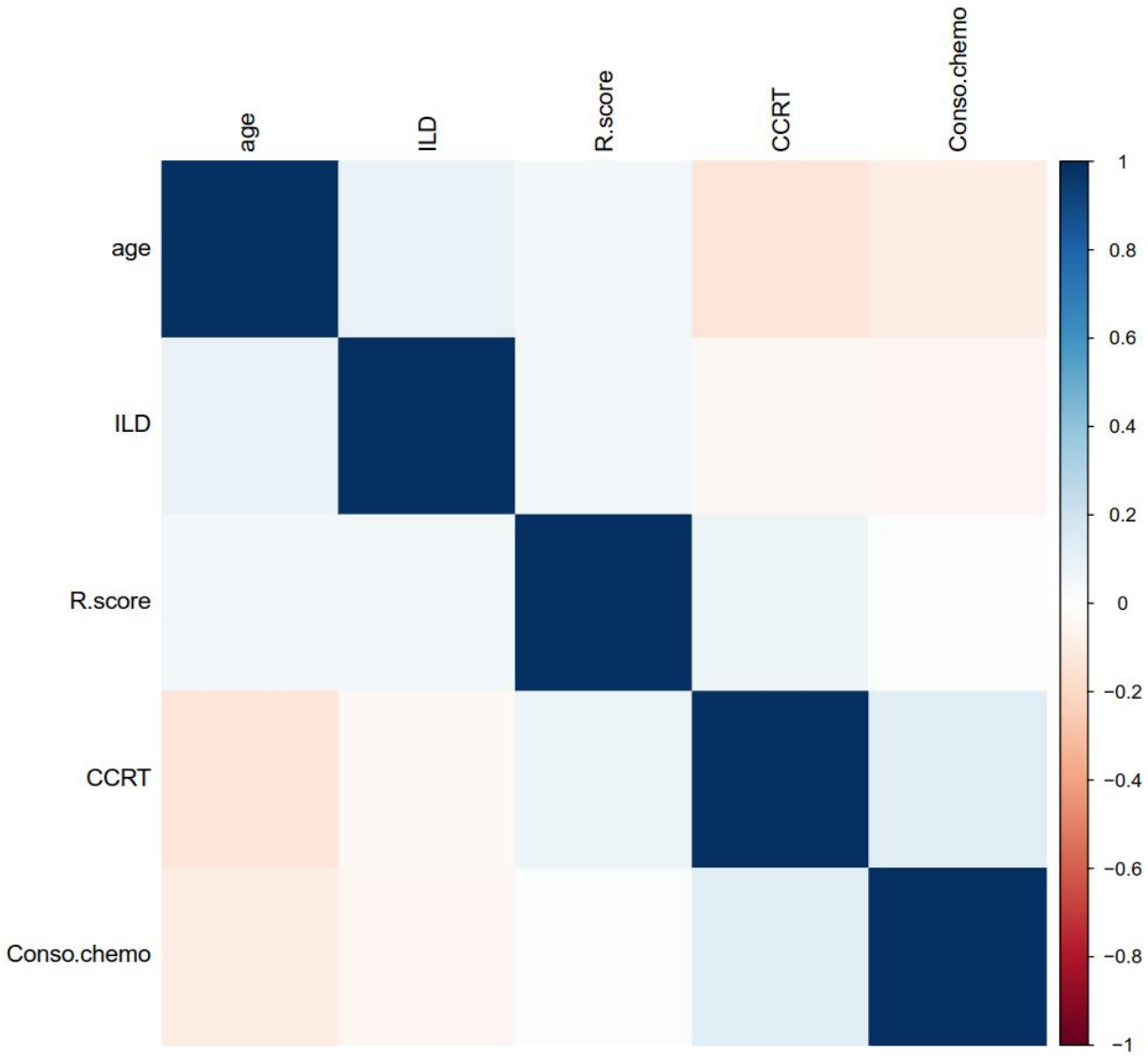


Supplementary Figure 2. The operator interface of the calculator.

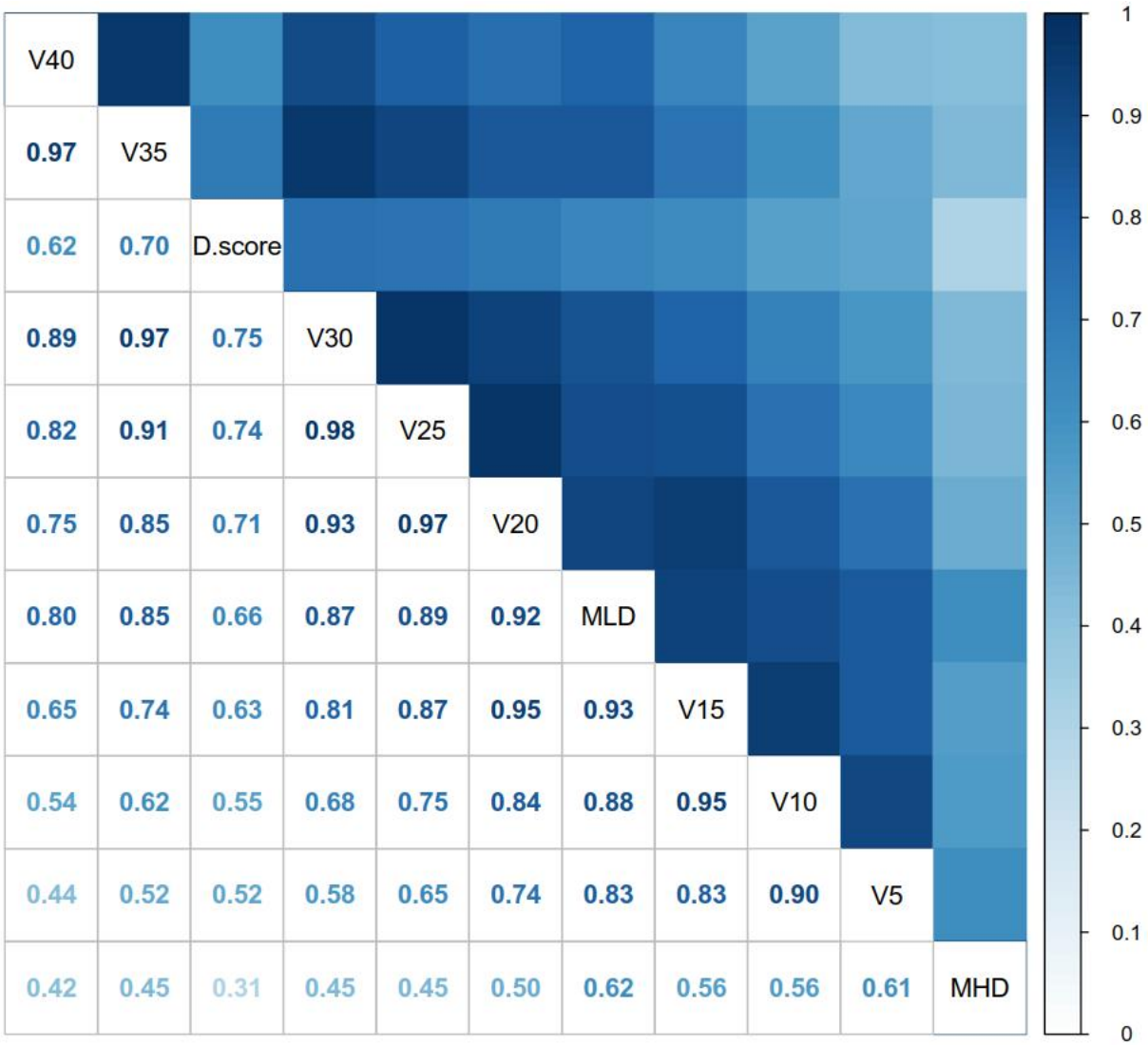
Supplementary material F

Correlations between different parameters

The correlation between the different parameters was calculated (Spearman correlation, R version 4.0.5). The results showed no significant correlation (>0.8) between radiomics risk score (R-score) and clinical parameters, dosiomics risk score (D-score) and dosimetrics.



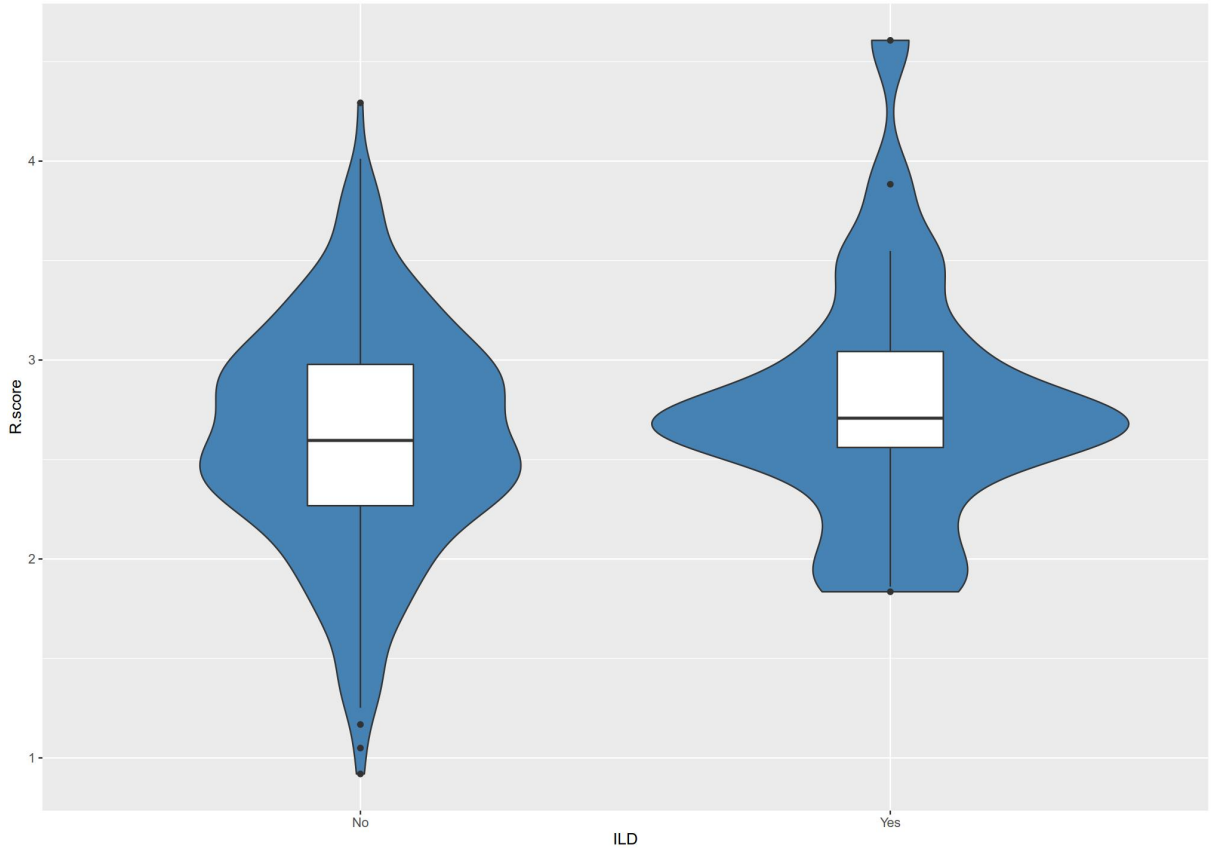
(a)



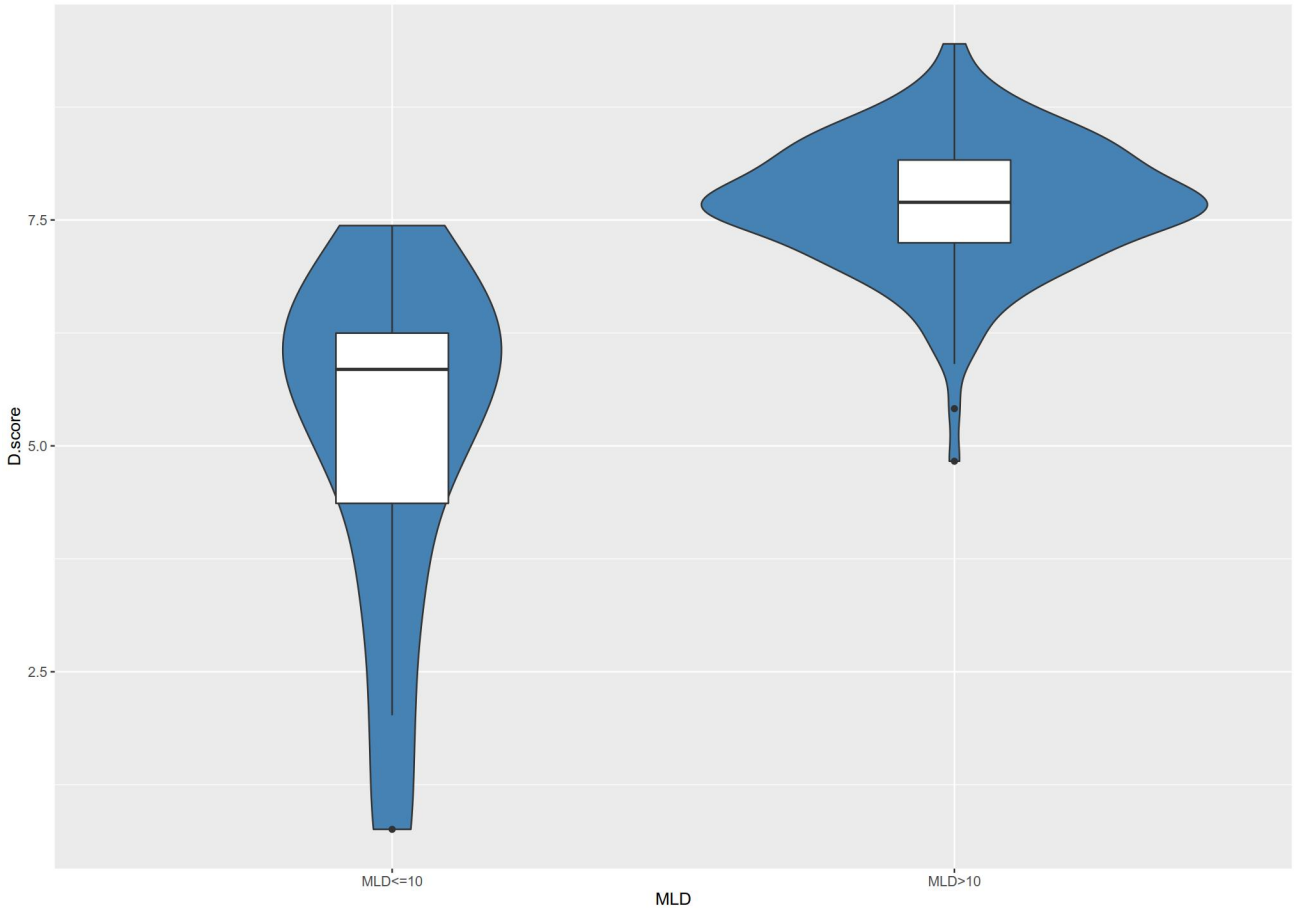
(b)

Supplementary Figure 3. (a) Correlations between R-score and clinical parameters. (b) Correlation between D-score and dosimetrics. *Abbreviations*: CCRT = concurrent chemoradiotherapy; Conso chemo = consolidation chemotherapy; R-score = radiomics risk score; D-score = dosiomics risk score; MLD = mean lung dose; MHD = mean heart dose.

Distribution of R-score and D-score



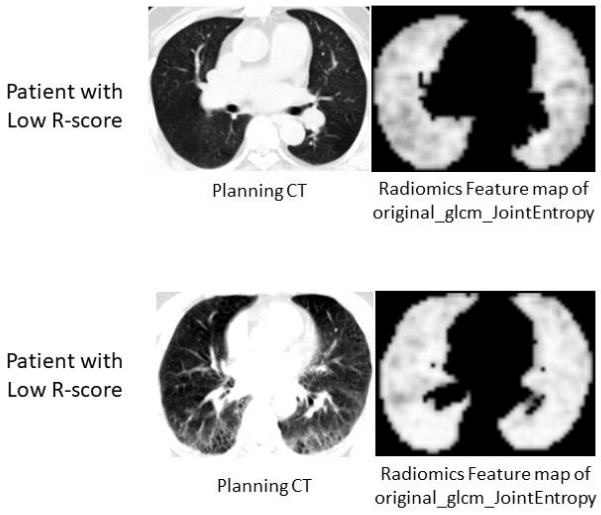
(a)



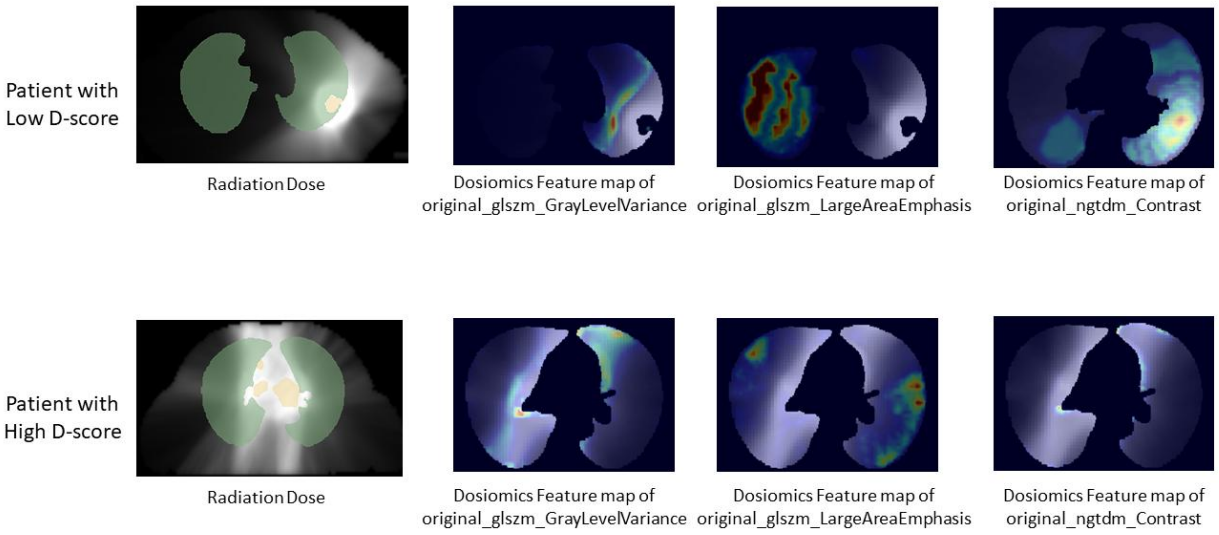
(b)

Supplementary Figure 4. (a) Distribution of radiomics risk score (R-score) in patients with and without interstitial lung disease (ILD). (b) Distribution of dosiomics risk score (D-score) among patients with mean lung dose (MLD) greater than 10Gy and less than or equal to 10Gy.

Feature maps



(a)



(b)

Supplementary Figure 5. (a) Radiomics feature map of feature "original_glcm_JointEntropy" for patient with low radiomics risk score (R-score) and patient with high R-score. (b) Dosiomics feature map of feature "original_glszm_GrayLevelVariance", "original_glszm_LargeAreaEmphasis" and "original_ngtdm_Contrast" for patient with low dosiomics risk score (D-score) and patient with high D-score.

Supplementary material G

Discrimination ability of different combination of Radiomics score, Dosiomics score and clinical parameters

Model	Train (95%CI)	Validation by bootstrapping (95%CI)	Testing (95%CI)
R-score + D-score	0.735 (0.673-0.796)	0.729 (0.720-0.736)	0.739 (0.553-0.926)
R-score + C	0.717 (0.652-0.782)	0.701 (0.683-0.719)	0.771 (0.585-0.962)
D-score + C	0.770 (0.710-0.830)	0.755 (0.744-0.765)	0.756 (0.559-0.954)

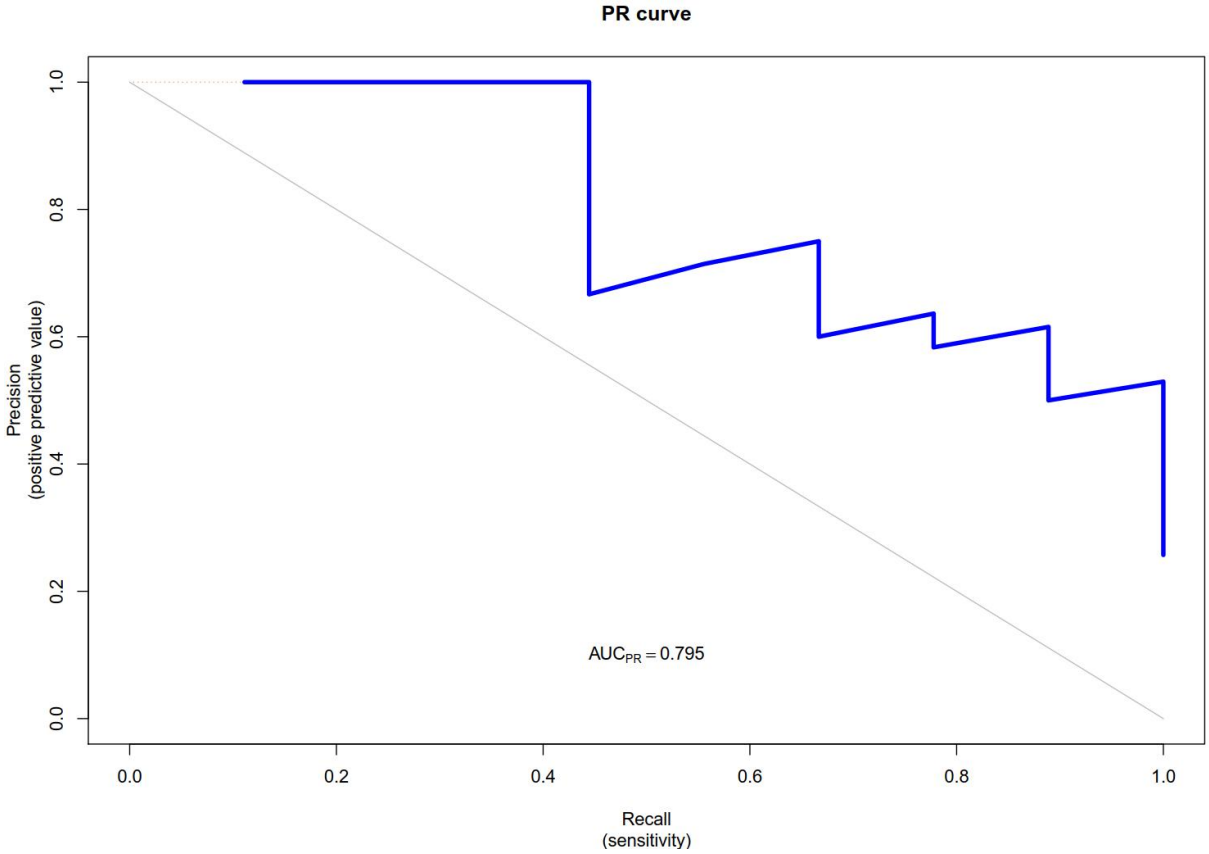
Abbreviations: R = radiomics risk score; D = dosiomics risk score; C = clinical parameters.

Discrimination ability of different models without patients with interstitial lung disease (ILD)

Model	Testing (95%CI)	Testing without patient with ILD (95%CI)
R-score	0.671 (0.558-0.899)	0.714 (0.348-1.000)
D-score	0.684 (0.573-0.883)	0.800 (0.613-0.987)
DVH-score	0.661 (0.551-0.856)	0.752 (0.505-1.000)
Clinical parameters	0.709 (0.509-0.91)	0.629 (0.392-0.865)
R-score + D-score + C	0.855 (0.719-0.990)	0.914 (0.785-1.000)

Abbreviations: ILD = interstitial lung disease.

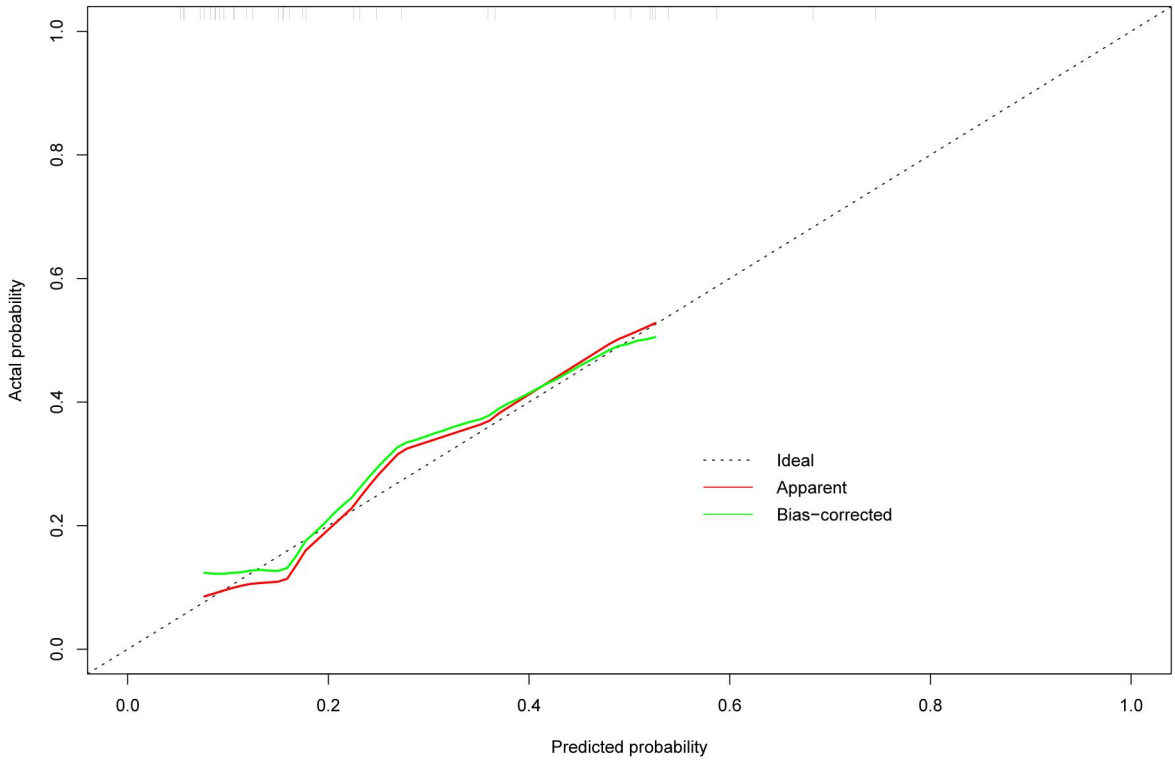
Precision Recall (RP) curve of the model combing R-score, D-score and Clinical parameters on the test set



Supplementary Figure 6. Precision Recall (RP)-curve

Supplementary material H

Calibration curve of prospective validation set with a bootstrap resampling method



Supplementary Figure 7. Calibration curve of prospective validation set with a bootstrap resampling method. Dashed line indicated the ideal model in which predicted and actual probabilities were perfectly identical; Red line indicated actual performance with apparent accuracy; Green line indicated bootstrap corrected estimate of the calibration curve.

Chapter 6: Clinical analysis and Artificial Intelligence Survival Prediction of Serous Ovarian Cancer Based on Preoperative Circulating Leukocytes

Adapted from Ying Feng, **Zhixiang Wang***, Ran Cui, Meizhu Xiao, Huiqiao Gao, Huimin Bai, Bert Delvoux, Zhen Zhang, Andre Dekker, Andrea Romano, Shuzhen Wang, Alberto Traverso, Chongdong Liu and Zhenyu Zhang. Clinical analysis and artificial intelligence survival prediction of serous ovarian cancer based on preoperative circulating leukocytes. J Ovarian Res 15, 64 (2022). https://doi.org/10.1186/s13048-022-00994-2*

** indicates equal contributions*

Abstract

Circulating leukocytes are an important part of the immune system. The aim of this work is to explore the role of preoperative circulating leukocytes in serous ovarian carcinoma and investigate whether they can be used to predict survival prognosis. Routine blood test results and clinical information of patients with serous ovarian carcinoma were retrospectively collected. And to predict survival according to the blood routine test result the decision tree method was applied to build a machine learning model.

The results showed that the number of preoperative white blood cells ($p = 0.022$), monocytes ($p < 0.001$), lymphocytes ($p < 0.001$), neutrophils ($p < 0.001$), and eosinophils ($p < 0.001$) and the monocyte to lymphocyte (MO/LY) ratio in the serous ovarian cancer group were significantly different from those in the control group. These factors also showed a correlation with other clinicopathological characteristics. The MO/LY was the root node of the decision tree, and the predictive AUC for survival was 0.69. The features involved in the decision tree were the MO/LY, differentiation status, CA125 level, neutrophils (NE,) ascites cytology, LY% and age.

In conclusion, the number and percentage of preoperative leukocytes in patients with ovarian cancer is changed significantly compared to those in the normal control group, as well as the MO/LY. A decision tree was built to predict the survival of patients with serous ovarian cancer based on the CA125 level, white blood cell (WBC) count, presence of lymph node metastasis (LNM), MO count, the MO/LY ratio, differentiation status, stage, LY%, ascites cytology, and age.

Keywords

Machine learning, leukocytes, serous ovarian cancer, recurrence, survival, and prediction.

1. Background

Ovarian carcinoma is the 5th leading cause of cancer-related deaths among women and the deadliest disease among gynecological malignancies[1,2]. Statistics from the United States show that the number of new cases of ovarian carcinoma in 2021 will be 22,530, and the number of deaths per year is estimated at approximately 13,770[1]. Ovarian cancer usually has a poor prognosis because many patients already present with advanced metastatic stages before diagnosis[3,2]. The 1-year survival rate is approximately 72%, the 5-year survival rate is 48%, and the 10-year survival rate is approximately 35%[4,5,2]. Serous carcinoma accounts for 75% of all ovarian cancers and is the most common pathological type [3,6]. Therefore, it is worthwhile to preoperatively predict the survival of serous ovarian carcinoma using clinicopathological features to guide decisions regarding surgery and postsurgical care.

Some reports have indicated that the interaction between ovarian cancer and the immune system may affect tumor growth and progression[7,8]. There is also some evidence that the inflammatory process caused by pelvic inflammatory disease may be associated with ovarian cancer[9]. Regarding the tumor evasion mechanism, tumor cells modulate the immune response for their benefit; tumor cells secrete specific cytokines that recruit and stimulate the production of myeloid-derived suppressor cells (MDSCs). They also produce TGF-β and IL-10 and inhibit T lymphocytes, macrophages and dendritic cells to create an immunosuppressive tumor microenvironment[10-12]. Due to the prominent role of the immune system in ovarian cancer, preoperative immune and inflammatory features may be suitable prognostic biomarkers. One promising characteristic is the leukocyte count.

Leukocytes, also called white blood cells (WBCs), are immune cells involved in protecting the body from disease and pathogens[13-15]. WBCs are distributed throughout the body, including the blood system and lymphatic system. WBCs account for approximately 1% of the total blood volume of healthy adults. There are five main subtypes of leukocytes: lymphocytes, monocytes, neutrophils, eosinophils, and basophils. They have a great impact on health because human immunity is based on the presence of and balance among these cell types. When an immune response occurs, as in the case of cancer, the number of WBCs will change accordingly[16,17,7,18].

Higher monocyte counts were reported to be associated with a poor prognosis in patients with endometrial cancer[19]. The lymphocyte-to-monocyte ratio

(LMR) in patients with epithelial ovarian cancer (EOC) and those with benign ovarian masses is significantly different[16]. The lymphocyte-to-monocyte ratio (LMR) has been significantly associated with the stage of EOC [20] and can provide prognostic information [21]. The monocyte-to-lymphocyte ratio has also been shown to predict shorter overall survival (OS) and progression-free survival (PFS) in EOC patients[22]. Therefore, we also paid attention to the monocyte-to-lymphocyte ratio in patients with serous ovarian cancer.

In this study, we aimed to explore the potential role of WBCs as prognostic biomarkers. Our primary objective is to investigate whether the number and proportion of circulating leukocytes in patients with serous ovarian carcinoma are different from those in normal controls (uterine prolapse patients). We also aimed to determine their association with clinicopathological characteristics, survival, and prognosis. As a secondary objective, we explored whether the test of preoperative circulating leukocytes can be used to predict the survival of ovarian serous carcinoma. To this end, machine learning in artificial intelligence (AI) [23], which is widely used in various medical fields, such as anatomy and brain-machine interfaces [24], is used to develop algorithms to predict the survival of patients with serous ovarian cancer.

2. Methods

2.1. Study subject

This study retrospectively analyzed patients with ovarian serous carcinoma who were initially treated at the Department of Obstetrics and Gynecology at Beijing Chaoyang Hospital, Capital Medical University, from July 2009 to December 2018. The case inclusion criteria were as follows: (1) surgical treatment performed at Beijing Chaoyang Hospital, (2) confirmation of ovarian serous carcinoma (serous cystadenocarcinoma or high-grade serous cystadenocarcinoma) by postoperative pathology, (3) standard platinum-based chemotherapy after the first tumor reduction surgery, and (4) complete preoperative routine blood and clinical data. The exclusion criteria were as follows: (1) presence of other types of benign and/or malignant ovarian tumors, (2) presence of primary malignant tumors of other organs, (3) no standardized chemotherapy after the first tumor reduction operation, and (4) incomplete routine blood and clinical data. The obtained data included age, BMI, childbirth history, menopause, neoadjuvant chemotherapy, surgical satisfaction, differentiation, stage based on the 2014 International Federation of Gynecology and Obstetrics (FIGO) staging system[25], ascites cytology, lymph node metastasis (LNM), recurrence, which is defined as the time from the first cytoreductive surgery to the time of ovarian cancer recurrence, death of disease (DOD) that is defined as the date from the first cytoreductive surgery to the date of the patient's death due to ovarian cancer, preoperative leukocyte count and proportion (within 90 days before the operation).

Recurrence was defined as the time from the first cytoreductive surgery to the time of ovarian cancer recurrence; death was defined as the date from the first cytoreductive surgery to the date of the patient's death due to ovarian cancer. The “normal”/control group selected and consisted of patients of a similar age who were diagnosed with uterine prolapse. The results of preoperative routine blood tests for these patients were also collected.

Ethics approval for this research was provided by the Beijing Chaoyang Hospital, Capital Medical University (approval number 2021-ke-205, study number 2012DRF30490).

2.2. Statistical analysis

Statistical analysis of the clinical data was performed with SPSS (version 23.0, IBM). Continuous data are expressed as the median and compared using the

Mann–Whitney U test. Total/differential leukocytes were divided into two groups according to the median value. Kaplan–Meier survival curves were then performed to compare overall survival (OS), which is defined as the time from the date of surgery to death (due to serous ovarian cancer), and progression-free survival (PFS), which is defined as the time from the date of surgery to recurrence, between the two groups. Significance was tested using the log-rank test, where these patients (5 patients) who had different first chemotherapy regimen were excluded. The three-dimensional (3D) histograms with three variables were constructed with Python 3.8. The significance was set at a two-sided p value < 0.05.

2.3. Machine learning

For survival prediction, we choose the machine learning-based decision tree algorithm. We divided the method into several steps, as shown in the flowchart (Figure 1).

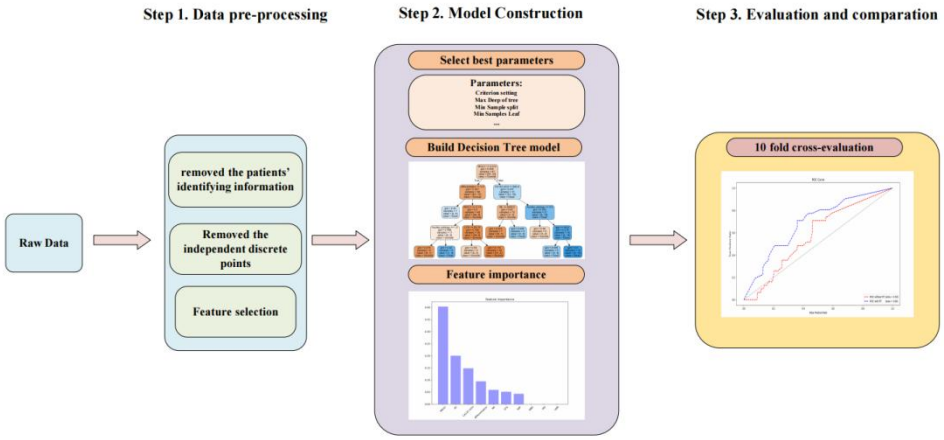


Figure 1. Flowchart for decision tree prediction. To build a machine learning model, frst, data pre-processing is required. Second, select best parameters and build the model. Third, evaluation and compare the model performance.

We implemented the algorithm in Python 3.8 and the scikit-learn 0.24 package. In the preprocessing part, we first removed the patients' identifying information. Second, we analyzed the distribution and removed the independent discrete points that were out of the value range of 5-95%. Third, we selected the features , such as stage, grade of differentiation and LNM, according to the National Comprehensive Cancer Network (NCCN) guideline[26]. Then, using 10-fold cross-validation, we separated the data into two parts: 90% of the data was used for training and 10% of the data was used

for testing. Imbalanced datasets are often handled well by decision tree classifiers[27], so we built a decision tree model and trained the model. Finally, to ensure the stability of the model, we used 10-fold cross-validation to train and test the model. To reduce the influence of imbalanced data, we used the synthetic minority oversampling technique (SMOTE) method to oversample the training set, which is an improved scheme based on a random oversampling algorithm[28]. To prove the role of circulating leukocytes in survival prediction, we performed comparisons with the same model trained by the features without circulating leukocytes.

For the decision tree learning process, these patients (93 patients) who had same first chemotherapy regimen were included. The optimal feature was selected recursively, and the training data were segmented according to the feature so that each subdataset had the best classification process. This process corresponded to the division of the feature space and the construction of the decision tree. First, the root node was constructed, and all training data were placed in the root node. An optimal feature was chosen, and the training dataset was divided into subsets according to this feature so that each subset had the best classification under the current conditions. If these subsets could be relatively correctly classified, then the leaf nodes were constructed, and these subsets were divided into the corresponding leaf nodes; if there were still subsets that could not be relatively correctly classified, then these subsets selected the new optimal feature, continued to divide it, and constructed the corresponding node. This process proceeded recursively until all the training data subsets were basically correctly classified or there were no suitable features. Finally, each subset was assigned to the leaf nodes.

3. Results

3.1. Patient clinicopathological characteristics and preoperative circulating leukocytes

A total of 98 patients with ovarian serous carcinoma who were initially treated at the Department of Obstetrics and Gynecology at Beijing Chaoyang Hospital, Capital Medical University, from July 2009 to December 2018 were included in the analysis according to the inclusion and exclusion criteria. The first chemotherapy regimen after the first surgery for all selected patients was platinum-based treatment (93 patients received 6-8 cycles of paclitaxel and cisplatin (PT), 3 patients received 8 cycles of cisplatin + adriamycin + cyclophosphamide (PAC), 1 patient received 8 cycles of paclitaxel and carboplatin, and 1 patient received 4 cycles of PT and 2 cycles of cisplatin + etoposide + ifosfamide (PEI)). The average age was 57 years old, and the mean BMI was 24.3. The pathological results revealed that 88.60% of the patients had poorly differentiated tumors (G3), 79.60% had stage III disease, 66.3% had positive ascites cytology, and 43.2% had LNM. The recurrence and mortality rates were 55.3% and 29.7%, respectively, at the time of follow-up (28 July 2019). The preoperative monocyte count and proportion in the serous ovarian cancer group (98 patients) were significantly higher than those in the control group (75 patients, $p < 0.001$ and $p < 0.001$, respectively, Table 1 and Figure 2).

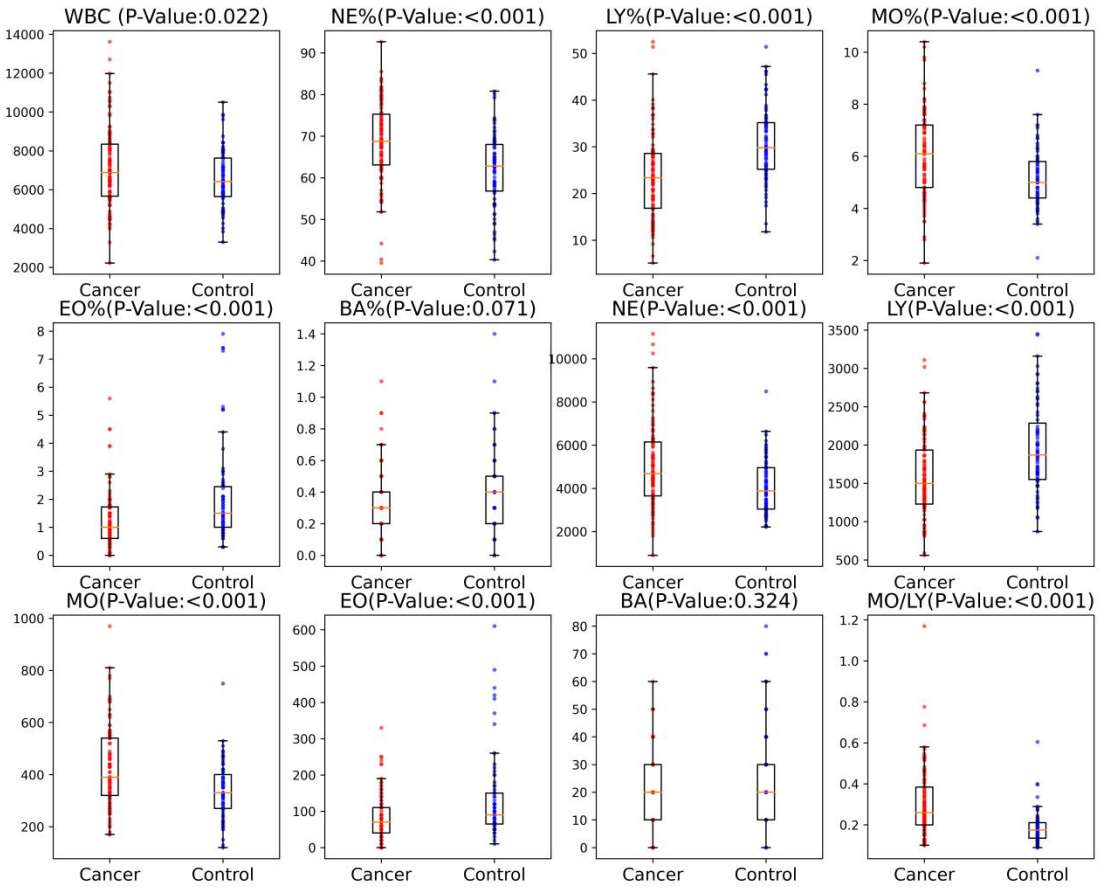


Figure 2. Boxplot distribution diagram of white blood cells. The mean±SEM of preoperative white blood cells were compared between control and serous ovarian cancer samples using a boxplot. Preoperative WBC, NE, NE%,MO, MO% and MO/LY in ovarian serous carcinoma patients (n=98) were significantly higher than those in control group (n=75), while LY, LY%, EO and EO% were significantly lower than those in control group. BA and BA% showed no difference between the two groups.

Table 1. The comparison of preoperative blood counts between serous ovarian cancer group and control normal group

Blood routine		Median		P-Value
		Serous ovarian cancer (N=98)	Control (N=75)	
WBC	10^6/L	7000	6420	0.022
NE	10^6/L	4710	3880	<0.001
LY	10^6/L	1510	1870	<0.001
MO	10^6/L	390	330	<0.001
EO	10^6/L	70	90	<0.001
BA	10^6/L	20	20	0.324
NE%		68.9	62.8	<0.001

LY%	23.2	29.8	＜0.001
MO%	6.1	5	＜0.001
EO%	1	1.5	＜0.001
BA%	0.3	0.4	0.071
MO/LY	0.2592	0.1746	＜0.001

Notes: WBC, white blood cells; NE, neutrophils; LY, lymphocytes; MO, monocytes; EO, eosinophil; BA, basophils; MO/LY, the ratio of monocytes to lymphocytes.

The monocyte-to-lymphocyte (MO/LY) ratio in the serous ovarian cancer group was also significantly higher than that in the normal control group (p <0.001). The number of white blood cells (WBCs, p =0.022), lymphocytes (LYs, p <0.001), neutrophils (NEs, p <0.001), and eosinophils (EOs, p <0.001) were also significantly different between the serous ovarian cancer group and the normal control group.

As shown in Table 2, the percentage of monocytes showed significant differences across the different disease stages (p=0.046); the more advanced the stage was, the higher the average percentage. The monocyte counts also showed similar results, with patients with LNM having more monocytes (p=0.05). The MO/LY ratio showed significant differences according to differentiation status (p=0.029), stage (p=0.007), LNM (p=0.025), and recurrence (p=0.036), with a higher ratio indicating a worse result, similar to the results for CA125. In addition, the number of NE (p =0.049) and BA (p =0.011) and the percentage of LY (LY%, p =0.036) affected LNM. LY% (p =0.048 and p =0.015, respectively) and NE% (p =0.027 and p =0.028, respectively) were significantly correlated with positive ascites cytology and recurrence.

Table 2. The relationship between preoperative blood counts and clinicopathological features in patients with serous ovarian cancer.

Characteristics (n=98)	WBC 10^6/L	NE 10^6/L	LY 10^6/L	MO 10^6/L	EO 10^6/L	BA 10^6/L	NE %	LY %	MO %	EO %	BA %	MO/LY	CA125 U/ml
Stage	p = 0.554	p = 0.260	p = 0.239	p = 0.050	p = 0.196	p = 0.053	p = 0.152	p = 0.054	p = 0.338	p = 0.347	p = 0.169	p = 0.026	p = 0.002
I + II	6430.00	4380.00	2040.00	350.00	70.00	20.00	66.30	27.80	5.50	1.00	0.20	0.19	52.60
III +IV	7000.00	4770.00	1505.00	430.00	75.00	20.00	69.25	22.70	6.10	1.05	0.30	0.27	847.40
Differentiation	p = 0.844	p = 0.742	p = 0.242	p = 0.134	p = 0.175	p = 0.844	p = 0.494	p = 0.261	p = 0.061	p = 0.123	p = 0.904	p = 0.028	p = 0.679
Low (G1)	6880.00	4700.00	1490.00	420.00	70.00	20.00	69.05	23.00	6.15	0.90	0.30	0.27	710.60
High (G2+G3)	7030.00	5540.00	1790.00	340.00	80.00	20.00	68.40	23.90	5.10	1.40	0.30	0.20	457.80
Ascites	p = 0.469	p = 1.00	p = 0.046	p = 0.077	p = 0.135	p = 0.852	p = 0.076	p = 0.111	p = 0.118	p = 0.232	p = 0.752	p = 0.697	p < 0.001
-	7540.00	5050.00	2080.00	520.00	100.00	20.00	65.40	26.30	6.30	1.40	0.30	0.25	71.74
+	6850.00	4680.00	1480.00	390.00	70.00	20.00	69.40	22.90	6.10	0.90	0.30	0.26	1118.50
Ascites cytology	p = 0.272	p = 0.076	p = 0.166	p = 0.585	p = 0.475	p = 0.813	p = 0.027	p = 0.048	p = 0.724	p = 0.373	p = 0.727	p = 0.201	p = 0.005

	-	6465.00	4435.00	1530.00	385.00	85.00	20.00	66.20	26.65	5.70	1.25	0.30	0.23	320.10
	+	7000.00	4880	1550.00	430.00	70.00	20.00	70.45	23.00	5.95	1.00	0.30	0.26	1216.59
LNM		p = 0.111	p = 0.049	p = 0.250	p = 0.050	p = 0.286	p = 0.011	p = 0.09	p = 0.036	p = 0.439	p = 0.547	p = 0.083	p = 0.025	p = 0.027
	-	6470.00	4450.00	1730.00	380.00	70.00	20.00	67.80	25.70	5.80	1.00	0.30	0.24	417.80
	+	7330.00	5210.00	1550.00	440.00	80.00	30.00	70.60	20.60	6.10	1.10	0.40	0.30	1094.00
Recurrence		p = 0.380	p = 0.152	p = 0.100	p = 0.371	p = 0.846	p = 0.390	p = 0.028	p = 0.015	p = 0.988	p = 0.540	p = 0.301	p = 0.036	p = 0.001
	-	6775.00	4465.00	1640.00	390.00	80.00	20.00	65.50	26.50	5.65	1.15	0.30	0.22	272.35
	+	7210.00	5050.00	1450.00	430.00	70.00	20.00	70.60	22.00	6.10	1.00	0.30	0.27	1218.00
Dead of disease		p = 0.958	p = 0.838	p =0.239	p = 0.791	p = 0.446	p = 0.536	p = 0.330	p = 0.239	p = 0.728	p = 0.408	p = 0.452	p = 0.152	p = 0.056
	-	6955.00	4685.00	1560.00	390.00	80.00	20.00	68.50	23.85	5.70	1.10	0.30	0.25	552.55
	+	7000.00	4990.00	1465.00	435.00	60.00	20.00	69.05	22.10	6.25	0.90	0.30	0.29	1216.59

Notes: WBC, white blood cells; NE, neutrophils; LY, lymphocytes; MO, monocytes; EO, eosinophil; BA, basophils; MO/LY, ratio of

monocytes to lymphocytes.; LNM, Lymph node metastasis.

3.2. Survival analysis based on preoperative circulating leukocytes

After dividing serous ovarian carcinoma patients into two groups based on the median value, OS and PFS decreased slightly faster in the group with a higher monocyte count (Kaplan–Meier analysis, Figure 3). A higher MO/LY ratio was significantly correlated with shorter PFS (p=0.001) and OS (p=0.048), which was similar to the results for CA125 (p=0.020 and <0.001, respectively).

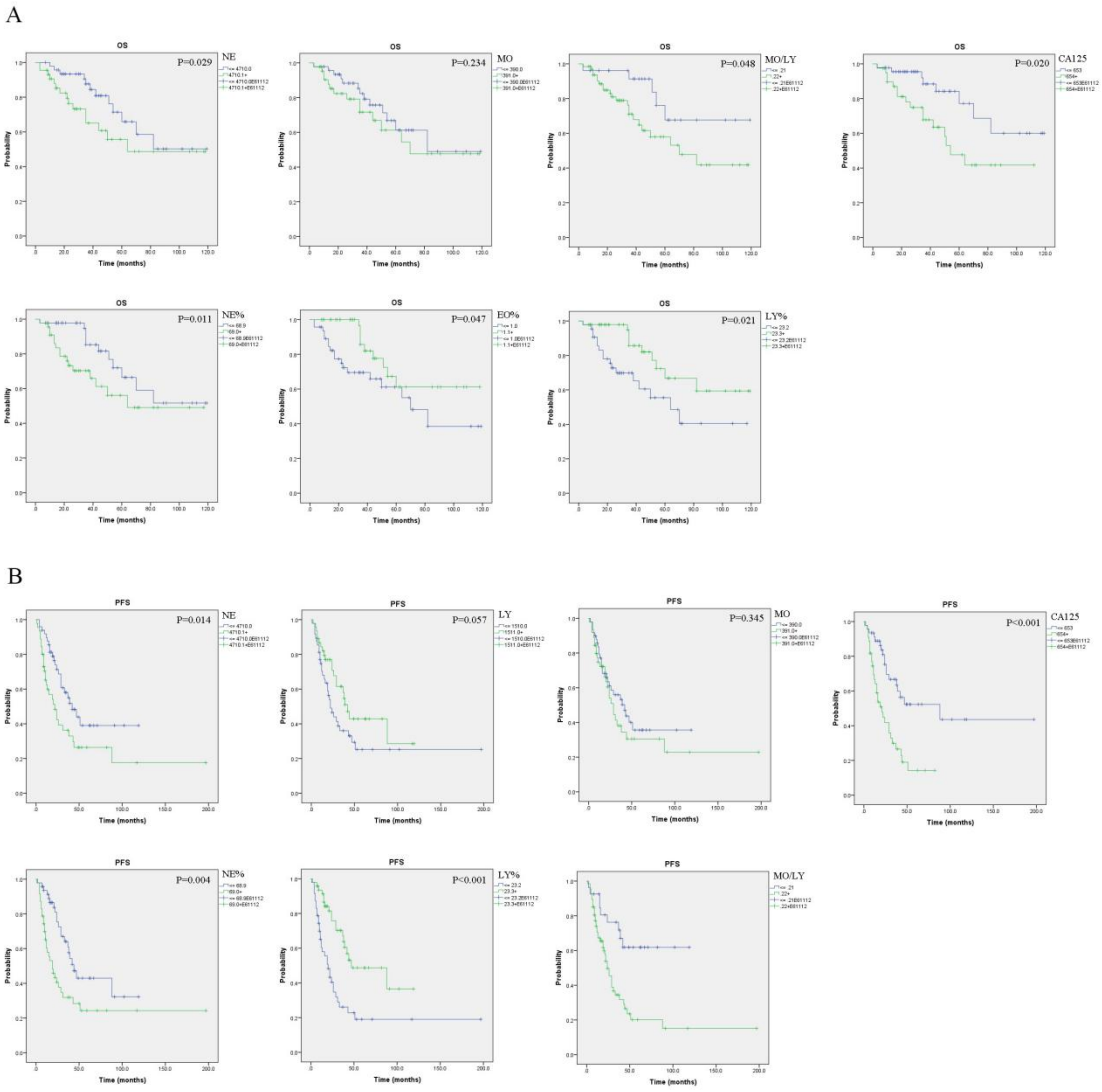


Figure 3. The prognostic value of preoperative blood counts in serous ovarian cancer. The Kaplan–Meier survival curves with the log-rank test were performed and compared between control and serous ovarian cancer samples. A comparison of OS between ovarian cancer and controls; B comparison of PFS between ovarian cancer and controls.

In addition, a higher NE (p=0.029 and 0.014, respectively) and NE% (p=0.011 and 0.004, respectively) significantly predicted shorter OS and PFS times. In contrast, the lower the LY% was (p=0.021 and <0.001, respectively), the worse the prognosis.

When assessing death and recurrence according to the tertiles of the MO/LY ratio cross-classified by the tertiles of the CA125 level, both the death rate and recurrence rate increased across the increasing tertiles of the MO/LY ratio for the first and second tertiles of the CA125 level (Figure 4). All patients within the third tertile of the CA125 level belonged to the first tertile of the MO/LY ratio. Therefore, no cases showed a high MO/LY ratio and high CA125 level at the same time.

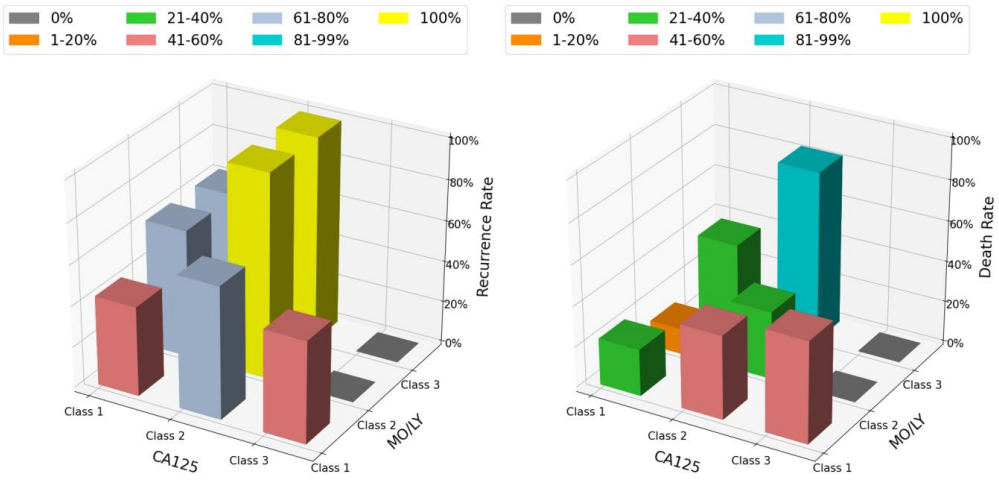


Figure 4. The three-dimensional distribution of CA125, MO/LY and events (death or recurrence). Percentage of patients suffering death (A) or recurrence (B) across tertiles of MO/LY ratio (MO/LY 1–3 = first to third tertiles) and CA125 (CA125 1–3 = first to third tertiles). Graded increases in the risk of death or recurrence are found across increasing tertiles of MO/LY ratio for the first and second tertile of CA125 levels.
Notes: All patients within the third tertile of CA125 level belonged to the first tertile of MO/LY ratio. Therefore, no cases showed a high MO/LY ratio and high CA125 level at the same time.

3.3. Decision tree to predict survival

For the decision tree, CA125 was found to be the root node with the largest information gain by using the built-in method of sklearn (Figure 5). The Gini coefficient reflects the measure of data uncertainty. The smaller the Gini value is, the higher the purity of the potential classes. In each node, the sample number shows the number of samples before being divided, and the value means the number belongs to each class. For example, in the root node, the total number of samples is 42, so the samples are 42. According to whether the CA125 attribute was less than or equal to 3726.05, the samples were split into two groups that contained 35 and 7 samples, respectively.

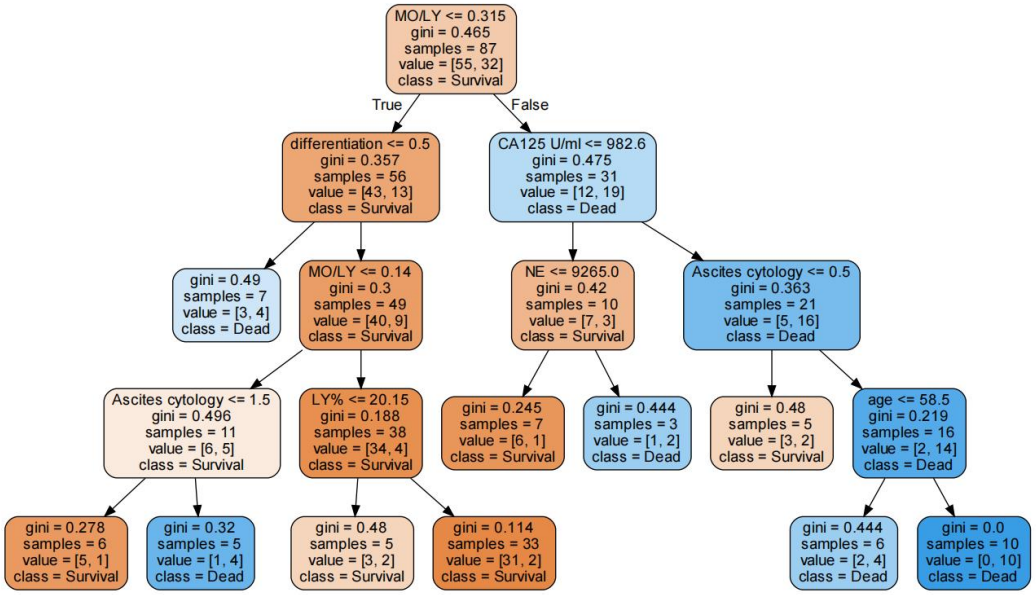


Figure 5. The decision tree visualization for predicting the survival of serous ovarian cancer.

Notes: In the prediction processing, at the root node, the sample is divided into two groups which have the MO/LY value less or equal to 0.315, or not. Then, the divided samples need to be judged by the second layer leaf node. In the second layer leaf nodes, the value CA125 or differentiation are the standards of classification. After that, it will go through into the third layer of leaf nodes until there is no leaf node left. Finally, when the decision reaches the last leaf node, the survival probability is the number of class samples divide total samples in the node. For example, at the leftmost leaf node, the probability of survival is 5/6 and the probability of death is 1/6.

For prediction processing, at the root node, the sample was divided into two groups based on a CA125 value less than or equal to 3726.05. Then, the divided samples were judged by the second layer leaf node. In the second layer leaf nodes, the value WBC or LNM are the standards of classification. After that, the samples will go through into the third layer of leaf nodes until there is no leaf node left. Finally, when the decision reaches the last leaf node, the survival probability is the number of class samples divided by the total samples in the node. For example, at the leftmost leaf node, the probability of survival is 5/6, and the probability of death is 1/6.

The features involved in the decision tree were the MO/LY, differentiation status, CA125 level, NE, ascites cytology, LY% and age. The survival prediction AUC of the decision tree was 0.69 (95% CI: 0.67-0.70).Meanwhile, the survival prediction AUC of the logsitic regression (LR) was 0.55(95% CI: 0.53-0.57), which means that the performance of decision tree is much bether.The performance cooperation between the model trained by the features with (blue line) and without the routine blood test

(RT) (red line) results is shown in Figure 6. From Figure 6, it is obvious that the model trained with RT has better performance than that without RT. The feature importance in the decision tree is shown in Figure 7. The MO/LY, differentiation, CA125, NE, ascites cytology, LY% and age had a high impact, and WBC, MO and LNM had a low impact on the model.

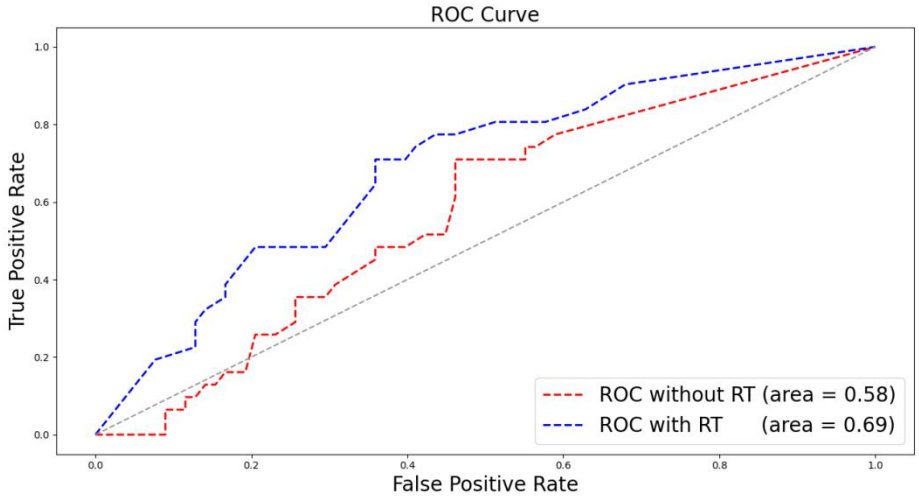


Figure 6. The performance cooperation in ROC curve. The model trained by the features with (blue line) and without blood routine test (RT) (red line)

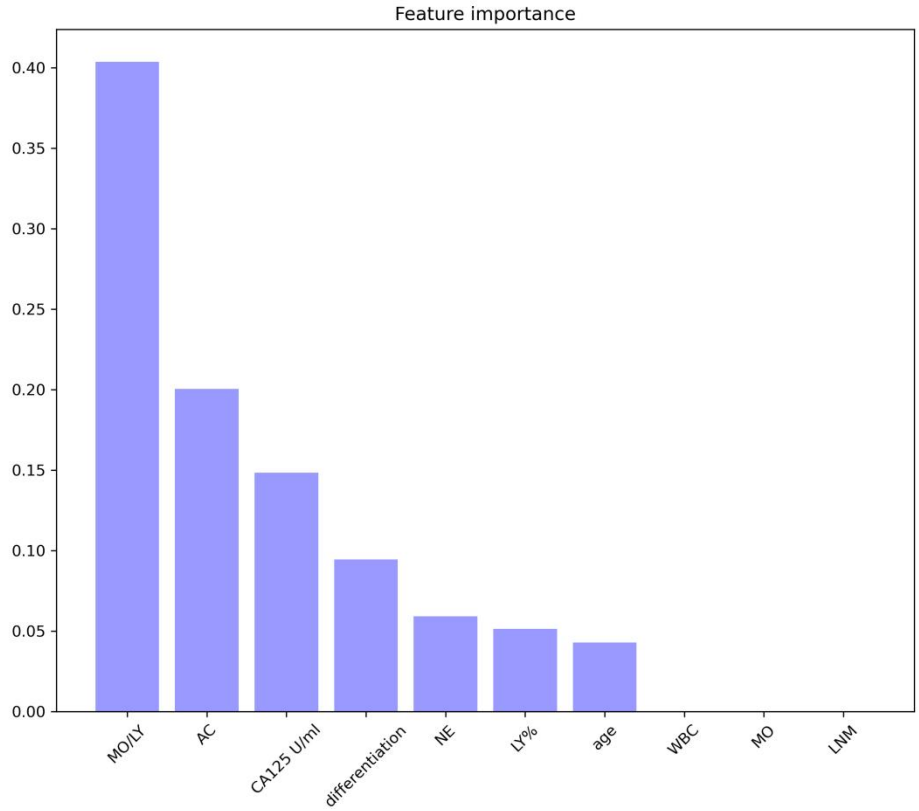


Figure 7. The features importance was shown with the rate of weight distribution of the decision tree.

4. Discussion

Ovarian carcinoma is the deadliest gynecological carcinoma, and epithelial ovarian cancers are the most common type of ovarian carcinoma. Two-thirds of epithelial ovarian cancers are serous carcinoma[3,6]. There are some signs that the inflammation caused by pelvic inflammatory disease may be associated with ovarian cancer[9]. Inflammation, cancer immunity and the immune microenvironment often involve various leukocytes[10-12]. In this study, we explored the role of preoperative circulating leukocytes in serous ovarian carcinoma and investigated their value in predicting survival prognosis. We found that most preoperative subtypes of WBCs, including monocytes, neutrophils, lymphocytes, and eosinophils, were significantly different between the serous ovarian carcinoma group and the control group, both in terms of the count and the percentage. These parameters have also been associated with the clinicopathological features of ovarian serous carcinoma.

Monocytes are the largest leukocytes and account for 2-10% of all leukocytes. These cells can migrate from the blood to tissues and then differentiate into macrophages. Monocytes and macrophages perform 3 major roles in the immune system, namely, phagocytosis, antigen presentation and cytokine production[29,30]. Most macrophages at disease sites are produced via the differentiation of circulating monocytes[31]. Lymphocytes account for 18% to 42% of all circulating leukocytes. Lymphocytes, such as T cells, B cells and natural killer cells, participate in many aspects of the immune response, including cancer immunity[10-12]. Therefore, we also calculated the monocyte-to-lymphocyte (MO/LY) ratio. The results revealed that the preoperative MO/LY was significantly increased in the blood of patients with serous ovarian cancer, similar to the results for monocytes. The higher the ratio is, the worse the prognosis. The possible underlying mechanism may be that monocytes enter the tumor microenvironment and then differentiate into tumor-associated macrophages and promote tumor development[32-34]. Lymphocytes are an important part of the immune response, so when the MO/LY ratio is out of balance, it indicates a poor survival prognosis. It is worth mentioning that the MO/LY seems to show important clinical value, similar to CA125, based on either its predictive value or the results of cross-variable 3D histograms and survival analysis.

In 2012, Vinod Khosla, co-founder of Sun Microsystems, predicted that 80% of clinical work will be replaced by automated machine learning medical diagnostic software in the next 20 years. As an example, in 2020, machine learning technology was used to help diagnose and treat COVID-19[35]. In this study, we applied a machine learning algorithm to predict the survival outcomes of patients with serous ovarian carcinoma and found that the MO/LY, differentiation status, CA125 level, NE, ascites cytology, LY% and age can be analyzed for survival prediction. This is consistent with the results showing that the MO/LY, CA125 level, NE and LY% are significantly associated with OS and with the NCCN guidelines, which indicate that differentiation, ascites cytology

and age are risk factors. In addition, the comparison between the model trained by the features with and without RT shows that the RT has an impact on prediction results. However, when a patient undergoes surgery and/or chemotherapy, the proportion and composition of WBCs in the blood changes significantly, so further research is needed to explore the postoperative situation.

The limitation of this article is mainly the small sample size. In the 3D histogram, there is a lack of data for the MO/LY and CA125, which are both very high. The separate analyses for high-grade and low-grade serous cancers are unable to be performed because of the limitation of the sample size. Therefore, studies on a larger sample size of patients as well as prospective studies are needed. In addition, leukocytes in the blood change according to the state of the body, so the role of postoperative circulating leukocytes still requires much research. Furthermore, the mechanism of action of leukocytes after reaching the tumor tissue site remains unclear.

5. Conclusions

The number and percentage of preoperative leukocytes change significantly in patients with ovarian cancer, as well as the MO/LY, and these changes can be correlated with other clinicopathological characteristics, including survival and recurrence. The clinical value of the MO/LY was similar to that of CA125. In addition, the decision trees generated with machine learning can predict the survival of patients with serous ovarian cancer based on the MO/LY, differentiation status, CA125 level, NE, ascites cytology, LY% and age. However, additional research is still warranted.

Abbreviations

MDSC: myeloid-derived suppressor cells;WBCs: white blood cells;LMR: lymphocyte-to-monocyte ratio;EOC: epithelial ovarian cancer;OS: overall survival;PFS: progression-free survival;AI: artificial intelligence;LNM: lymph node metastasis;DOD: recurrence; death of disease;FIGO: Federation of Gynecology and Obstetrics;3D: three-dimensional;PT: paclitaxel and cisplatin;PAC: cisplatin + adriamycin + cyclophosphamide;PEI: cisplatin + etoposide + ifosfamide;MO/LY: monocytes to lymphocytes;NE: neutrophils;LY: lymphocytes;EO: eosinophils; RT, blood routine test.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A (2021) Cancer Statistics, 2021. CA Cancer J Clin 71 (1):7-33. doi:10.3322/caac.21654
2. (NCCN®) NCCN (2021) Ovarian Cancer, Including Fallopian Tube Cancer and Primary Peritoneal Cancer. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Version 1.2021 — February 26, 2021
3. Lheureux S, Gourley C, Vergote I, Oza AM (2019) Epithelial ovarian cancer. Lancet 393 (10177):1240-1253. doi:10.1016/S0140-6736(18)32552-2
4. Noone AM HN, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds) (Updated September 10, 2018) SEER Cancer Statistics Review (CSR) 1975-2015, based on November 2017 SEER data submission, posted to the SEER web site, April 2018. Bethesda, MD: National Cancer Institute
5. Yi X, Walia E, Babyn P Generative adversarial network in medical imaging: A review. (1361-8423 (Electronic))
6. Network NCC (2019) Ovarian Cancer. Including Fallopian Tube Cancer and Primary Peritoneal Cancer. NCCN Guidelines Version 3.2019 NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Version 3.2019 — November 26, 2019
7. Bakacak M, Serin S, Ercan O, Kostu B, Bostanci MS, Bakacak Z, Kiran H, Kiran G (2016) Utility of preoperative neutrophil-to-lymphocyte and platelet-to-lymphocyte ratios to distinguish malignant from benign ovarian masses. J Turk Ger Gynecol Assoc 17 (1):21-25. doi:10.5152/jtgga.2015.0152
8. Yildirim MA, Seckin KD, Togrul C, Baser E, Karsli MF, Gungor T, Gulerman HC (2014) Roles of neutrophil/lymphocyte and platelet/lymphocyte ratios in the early diagnosis of malignant ovarian masses. Asian Pac J Cancer Prev 15 (16):6881-6885. doi:10.7314/apjcp.2014.15.16.6881
9. Ingerslev K, Hogdall E, Schnack TH, Skovrider-Ruminski W, Hogdall C, Blaakaer J (2017) The potential role of infectious agents and pelvic inflammatory disease in ovarian carcinogenesis. Infect Agent Cancer 12:25. doi:10.1186/s13027-017-0134-9
10. Wagner M, Koyasu S (2019) Cancer Immunoediting by Innate Lymphoid Cells. Trends Immunol 40 (5):415-430. doi:10.1016/j.it.2019.03.004
11. Mantovani A (2010) The growing diversity and spectrum of action of myeloid-derived suppressor cells. Eur J Immunol 40 (12):3317-3320. doi:10.1002/eji.201041170
12. Ha TY (2009) The role of regulatory T cells in cancer. Immune Netw 9 (6):209-235. doi:10.4110/in.2009.9.6.209
13. Meng-Hsiun Tsai S-SY, Yung-Kuan Chan, Chun-Chu Jen (2015) Blood smear image based malaria parasite and infected-erythrocyte detection and segmentation. J Med Syst
39 (10):118. doi:10.1007/s10916-015-0280-9
14. Enas Abdulhay MAM, Dheyaa Ahmed Ibrahim, N Arunkumar, V Venkatraman (2018) Computer aided solution for automatic segmenting and measurements of blood

leucocytes using static microscope images. J Med Syst 42 (4):58. doi:10.1007/s10916-018-0912-y
15. Hong Liu HC, Enmin Song (2019) Bone marrow cells detection a technique for microscopic image analysis. J Med Syst 43 (4):82. doi:10.1007/s10916-019-1185-9
16. Eo WK, Kim KH, Park EJ, Kim HY, Kim HB, Koh SB, Namkung J (2018) Diagnostic accuracy of inflammatory markers for distinguishing malignant and benign ovarian masses. J Cancer 9 (7):1165-1172. doi:10.7150/jca.23606
17. Zhang H, Yang Z, Zhang W, Niu Y, Li X, Qin L, Su Q (2017) White blood cell subtypes and risk of type 2 diabetes. J Diabetes Complications 31 (1):31-37. doi:10.1016/j.jdiacomp.2016.10.029
18. Carlos Cardoso-Vigueros TvB, Beate Rückert, Arturo Rinaldi, Ge Tan, Anita Dreher, Urszula Radzikowska, Guenter Menz, Peter Schmid-Grendelmeier, Cezmi A Akdis, Milena Sokolowska (2022) Leukocyte redistribution as immunological biomarker of corticosteroid resistance in severe asthma. Clin Exp Allergy Online ahead of print. doi:10.1111/CEA.14128
19. Burgess B, Levine B, Taylor RN, Kelly MG (2020) Preoperative Circulating Lymphocyte and Monocyte Counts Correlate with Patient Outcomes in Type I and Type II Endometrial Cancer. Reprod Sci 27 (1):194-203. doi:10.1007/s43032-019-00009-4
20. Li L, Tian J, Zhang L, Liu L, Sheng C, Huang Y, Zheng H, Song F, Chen K (2021) Utility of Preoperative Inflammatory Markers to Distinguish Epithelial Ovarian Cancer from Benign Ovarian Masses. J Cancer 12 (9):2687-2693. doi:10.7150/jca.51642
21. Eo W, Kim HB, Lee YJ, Suh DS, Kim KH, Kim H (2016) Preoperative Lymphocyte-Monocyte Ratio Is a Predictor of Suboptimal Cytoreduction in Stage III-IV Epithelial Ovarian Cancer. J Cancer 7 (13):1772-1779. doi:10.7150/jca.15724
22. Li Z, Hong N, Robertson M, Wang C, Jiang G (2017) Preoperative red cell distribution width and neutrophil-to-lymphocyte ratio predict survival in patients with epithelial ovarian cancer. Sci Rep 7:43001. doi:10.1038/srep43001
23. Rajkomar A, Dean J, Kohane I (2019) Machine Learning in Medicine. Reply. N Engl J Med 380 (26):2589-2590. doi:10.1056/NEJMc1906060
24. Hu J, Niu H, Carrasco J, Lennox B, Arvin F (2020) Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning. IEEE Transactions on Vehicular Technology 69 (12):14413-14423. doi:10.1109/tvt.2020.3034800
25. Berek JS, Kehoe ST, Kumar L, Friedlander M (2018) Cancer of the ovary, fallopian tube, and peritoneum. International Journal of Gynecology & Obstetrics 143:59-78. doi:10.1002/ijgo.12614
26. Network NCC (2022) Ovarian Cancer, Including Fallopian Tube Cancer and Primary Peritoneal Cancer, Version 1.2022. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) January 18, 2022

27. Sankari ES, Manimegalai D Predicting membrane protein types using various decision tree classifiers based on various modes of general PseAAC for imbalanced datasets. (1095-8541 (Electronic))
28. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WPJJoair (2002) SMOTE: synthetic minority over-sampling technique. 16:321-357
29. B A Nichols DFB, M G Farquhar (1971) Differentiation of Monocytes. Origin, Nature, and Fate of Their Azurophil Granules The Journal of cell biology 50 (2):498-515
30. Filip K Swirski MN, Martin Etzrodt, Moritz Wildgruber, Virna Cortez-Retamozo, Peter Panizzi, Jose-Luiz Figueiredo, Rainer H Kohler, Aleksey Chudnovskiy, Peter Waterman, Elena Aikawa, Thorsten R Mempel, Peter Libby, Ralph Weissleder, Mikael J Pittet (2009) Identification of Splenic Reservoir Monocytes and Their Deployment to Inflammatory Sites. Science 325 (5940):612-616. doi:10.1126/science.1175202
31. Pittet MJ, Nahrendorf M, Swirski FK (2014) The journey from stem cell to macrophage. Ann N Y Acad Sci 1319:1-18. doi:10.1111/nyas.12393
32. Qianxia Tan HL, Jie Xu, Yanqun Mo, Furong Dai (2021) Integrated analysis of tumor-associated macrophage infiltration and prognosis in ovarian cancer. Aging (Albany NY) 11;13(undefined)
33. Nowak M, Klink M (2020) The Role of Tumor-Associated Macrophages in the Progression and Chemoresistance of Ovarian Cancer. Cells 9 (5). doi:10.3390/cells9051299
34. Feng Y, Xiao M, Zhang Z, Cui R, Jiang X, Wang S, Bai H, Liu C, Zhang Z (2020) Potential interaction between lysophosphatidic acid and tumor-associated macrophages in ovarian carcinoma. J Inflamm (Lond) 17:23. doi:10.1186/s12950-020-00254-4
35. Vaishya R, Javaid M, Khan IH, Haleem A (2020) Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr 14 (4):337-339. doi:10.1016/j.dsx.2020.04.012

Chapter 7: An applicable machine learning model based on preoperative examinations predicts histology, stage and grade for endometrial cancer

Adapted from Ying Feng, Zhixiang Wang*, Meizhu Xiao, Jinfeng Li, Yuan Su, Bert Delvoux, Zhen Zhang, Andre Dekker, Sofia Xanthoulea, Zhiqiang Zhang, Alberto Traverso, Andrea Romano, Zhenyu Zhang, Chongdong Liu, Huiqiao Gao, Shuzhen Wang and Linxue Qian. An Applicable Machine Learning Model Based on Preoperative Examinations Predicts Histology, Stage, and Grade for Endometrial Cancer. Front Oncol 2022, 12, 904597. https://doi.org/10.3389/fonc.2022.904597.*
** indicates equal contributions*

Abstract

Purpose

To build a machine learning model to predict histology (type I and type II), stage, and grade preoperatively for endometrial carcinoma to quickly give a diagnosis and assist in improving the accuracy of the diagnosis, which can help patients receive timely, appropriate and effective treatment.

Materials and methods

This study used a retrospective database of preoperative examinations (tumor markers, imaging, diagnostic curettage etc.) in patients with endometrial carcinoma. Three algorithms (Random Forest, Logistic Regression, and Deep neural network) were used to build models. The AUC and accuracy were calculated. Furthermore, the performance of machine learning models, doctors' prediction, and doctors with the assistance of models were compared.

Results

A total of 329 patients were included in this study with 16 features (age, BMI, stage, grade, histology etc.). A Random Forest algorithm had the highest AUC and Accuracy. For histology prediction, AUC and Accuracy was 0.69 (95% CI=0.67-0.70) and 0.81 (95%CI=0.79-0.82). For stage they were 0.66 (95% CI=0.64-0.69) and 0.63 (95% CI=0.61-0.65) and for differentiation grade 0.64 (95% CI=0.63-0.65) and 0.43 (95% CI=0.41-0.44). The average accuracy of doctors for histology, stage and grade was 0.86 (with AI) and 0.79 (without AI), 0.64 and 0.53, 0.5 and 0.45, respectively. The accuracy of doctors' prediction with AI was higher than that of Random Forest alone and doctors' prediction without AI.

Conclusion

A random forest model can predict histology, stage, and grade of endometrial cancer preoperatively and can help doctors in obtaining a better diagnosis and predictive results.

Keywords: Machine learning, endometrial carcinoma, diagnosis, prediction, random forest, preoperatively

Introduction

Endometrial carcinoma (EC) represents the 6th most common malignant tumor worldwide. [1] In the past 2020, the number of new cases of endometrial cancer was 417,367, and the number of new deaths was 97,370. [1] This may be due to increased obesity, aging, and physical inactivity. [2, 3] Endometrial carcinoma occurs most commonly in postmenopausal women. [4] The first symptom is often abnormal vaginal bleeding. Transvaginal ultrasound is an effective examination to evaluate the presence of endometrial carcinoma, besides pelvic and physical examination. [2, 5] A histopathology diagnosis is commonly assessed by dilation and curettage (D&C) or endometrial biopsy before surgery. However, the preoperative endometrial biopsy and final diagnosis are not completely consistent with only a moderate agreement rate on grade, especially for grade 2 tumors. [2] In addition, other serological and imaging tests are routine tests for the diagnosis of endometrial carcinoma. [2, 3].

With the development of computer science, clinical decision support systems (CDSSs) are being developed. A CDSS is defined as a system that enhances clinical information and medical knowledge to help doctors and nurses with clinical decisions for better health care[6]. CDSS is a major subject of medical artificial intelligence (AI). CDSS can be used pre-diagnosis (prepare diagnoses), during diagnosis (review and filter diagnoses), and post-diagnosis (predict future events).

However, there are no studies that use an AI model to predict histology, stage, and grade for endometrial carcinoma based on the preoperative examinations. Such an AI model can be a part of an endometrial cancer CDSS to improve the efficiency of doctors, reduce the rate of misdiagnosis, and improve the quality of health care.

Machine learning (ML), a type of AI [7], is widely used in medical fields, such as anatomy, medical diagnoses, and brain-machine interfaces [8]. In 2022 Otani et al proposed a ML based classifier to predict the EC risk from the multiparametric magnetic resonance images (MRI)[9]. And, in 2021, Nakajo et al proved that a 18F-FDG PET-based radiomic analysis using a machine learning approach may be useful for predicting tumor progression and prognosis in patients with endometrial cancers[10].

In this study, we used ML to build three models to predict histology (type I and type II), stage, and grade for endometrial carcinoma to quickly give a diagnosis and assist in improving the accuracy of the diagnosis, which can help patients receive timely, appropriate and effective treatment.

Methods

Study subject

This study used a retrospective database of preoperative examinations in patients with endometrial carcinoma who were first treated in the Department of Obstetrics and Gynecology at Beijing Chaoyang Hospital, Capital Medical University, from January 2000 to April 2014. Inclusion criteria were as follows: (1) undergoing surgical treatment at Beijing Chaoyang Hospital, (2) confirmation of endometrial carcinoma by postoperative pathology, (3) without neoadjuvant chemotherapy and hormone therapy, (4) all treatments have been completed, (5) complete clinical-pathological data. The case exclusion criteria were: (1) presence of primary malignant tumors of other organs, (2) metastatic cancer caused by malignant tumors of other organs, (3) not the first-time surgical treatment at Beijing Chaoyang Hospital, (4) with neoadjuvant chemotherapy and hormone therapy, (5) incomplete clinical-pathological data. The obtained data included age, BMI, childbirth history, preoperative serum tumor markers, imaging results, histopathology diagnosis after D&C, hypertension, diabetes, menopause, symptoms, postoperative histology, stage based on the 2014 International Federation of Gynecology and Obstetrics (FIGO) staging system[11], and grade. Ethics approval for this research was given by the Beijing Chaoyang Hospital, Capital Medical University.

Data and Machine Learning Algorithms

A total of 16 features mentioned above were used for the development of the classification models.

For data preprocessing, first, we transformed semi-structured and unstructured features such as preoperative serum tumor markers, imaging results into structured features. Then, we normalized the continuous variables such as age and BMI into 0 to 1.

In this study, we trained and compared 3 classifiers, including logic regression(LR)[12], random forest (RF)[13], and a deep neural network(DNN)[14]. The DNN is based on the extension of the perceptron: a neural network with many hidden layers. Random forest is an ensemble algorithm (Ensemble Learning), which belongs to the Bagging algorithms. By combining multiple weak classifiers, the final result is voted or averaged, so that the result of the overall model has higher accuracy and generalization performance.

The DNN model was composed of 2 fully connection layers which have a Rectified Linear Unit (ReLU) activation function to increase the nonlinearity of the neural network model and dropout layers with the rate of 0.5 to avoid over-fitting and 1 fully connection layer without activation function. The cross-entropy loss was used to guide the training process by using a stochastic gradient descent (SGD) optimizer with a 0.0002 learning rate. The random forest included one hundred decision trees.

The classification models were trained and tested with the selected features to predict the histology, stage and grade of endometrial cancer. For model training, we trained and validated the model 100 times (RF, LR) and 10 times (DNN) repeating random sampling verification. We split the dataset into training and testing datasets with a ratio of 7:3 in each validation. Then we used the Synthetic Minority Oversampling Technique (SMOTE) method in the training set for over-sampling, which adds artificially simulated new samples to the data set to decrease the influence of imbalanced data.

To evaluate the preformace of the classification models, we calculated the Area Under the Curve(AUC) and the accuracy.
In addition, we also investigated whether the AI algorithms can play a role in the diagnosis accuracy and speed of the doctor's diagnosis. We generated 4 test sets for doctors with 40 patients, half of the patients with an AI prediction class and its possibility, and the other half of the patients without any assistance. Then we sent the test sets to obstetric oncologists to measure the AUC, accuracy, and the time consumption for the predicting the disease category with and without AI assistance. The function of accuracy shown below.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP: True Positive , TN: True Negative, FP:False Positive, FN: False Negative
Data pre-processing and machine learning models were implemented within Python 3.8, and scikit-learn 0.24 and PyTorch 1.10 packages.
Comparison of Different Models
The comparison of accuracy between models was performed by using the two-way ANOVA test in GraphPad Prism.

Results

Clinical information of cases

A total of 344 endometrium cancer cases were reviewed and collected. Of these, 14 cases were excluded because of 70% or more of missing clinical data. As there was only one undifferentiated case, this category could not be tested because the test sample would be 0. Therefore, 329 cases were enrolled into the train and test\. The mean age was 56 (range 28-83) years old (Table 1). The mean BMI was 26.87±4.43. Among these cases, 86.3% of the patients were type I EC. Most (75.7%) of the cases were FIGO stage I and 31 cases were grade (G) 1, 114 cases were G2, 38 cases were G3, and 17 cases were unknown.

Table 1. Clinicopathological data of patients with endometrial cancer

Features	Frequency (%) N=329
Age, mean (range)	56 (28-83)
BMI, mean±SD	26.87±4.43
Hypertension	
+	144 (43.8)
-	184 (55.9)
Unknown	1 (0.3)
Diabetes	
+	71 (21.6)
-	256 (77.8)
Unknown	2 (0.6)
Gestation	
+	312 (94.8)
-	17 (5.2)
Parturition	
+	301 (91.8)
-	28 (8.5)
Menopause	
+	192 (58.3)
-	13 (4.0)
Unknown	124 (37.7)
Histology	
type I	284 (86.3)
type II	45 (13.7)
FIGO Stage (2009)	

I	249 (75.7)
II	28 (8.5)
III	42 (12.8)
IV	10 (3.0)
Differentiation	
G1	31 (37.7)
G2	114 (45.6)
G3	38 (11.6)
Unknown	17 (5.2)

Notes: G, grade; SD, standard deviation; FIGO, the international federation of obstetrics and gynecology.

Comparison of the Models for the Prediction
Histology

The AUC and the Accuracy score of the LR were 0.69 (95% CI=0.67-0.70) and 0.74 (95%CI=0.72-0.75). The AUC and the Accuracy score of RF were 0.69 (95% CI=0.67,0.70) and 0.81 (95%CI=0.79-0.82). And the AUC and the Accuracy score of DNN were 0.60 (95% CI=0.54-0.65) and 0.83 (95% CI=0.75-0.90). The LR and RF algorithms have a similar score which was significantly better ($p<0.05$) than DNN.

Stage

The AUC and the Accuracy score of the logistic regression were 0.56 (95% CI=0.54-0.59) and 0.42 (95% CI=0.41-0.44). The AUC and the Accuracy score of the random forest were 0.66 (95% CI=0.64-0.69) and 0.63 (95% CI=0.61-0.65). And the AUC and the Accuracy score of DNN was 0.48 (95% CI=0.46-0.51) and 0.78 (95% CI=0.71,0.84). The RF was significantly better than LR and DNN.

Grade

The AUC and the Accuracy score of the LR were 0.61 (95% CI=0.60-0.62) and 0.36 (95% CI=0.35-0.38). The AUC and the Accuracy score of RF was 0.64 (95% CI=0.63-0.65) and 0.43 (95% CI=0.41-0.44). And the AUC and the Accuracy score of DNN were 0.47 (95% CI=0.45-0.50) and 0.43 (95% CI=0.40-0.45). The LR and RF algorithms have a similar score significantly better than DNN.

Performance comparison between ML model, doctors' prediction and doctors with the assistance of AI

The result of the doctor's prediction is shown in Table 3. The average accuracy for histology was 86% (with AI) and 79% (without AI), respectively. The average accuracy for the stage was 64% and 53%, respectively. The average accuracy for differentiation was 50% and 45%, respectively. The time consumption for each patient to make a decision was 29.25 s (with AI) and 28.75 s (without AI), respectively. For type and stage diagnosis, the AI model can improve 6% and 10% of a doctor's accuracy. But the accuracy decreases 7% for the differentiation diagnosis. The average time consumption with AI was 10 seconds longer than that without AI, though the AI model only cost 3 ms to predict one patient..

Table 3. Comparison of the doctors' prediction with and without AI assistance

Project	Without AI (accuracy %)	With AI (accuracy %)
Histology	79	86
Stage	53	64
Differentiation	45	50

Notes: AI, artificial intelligence.

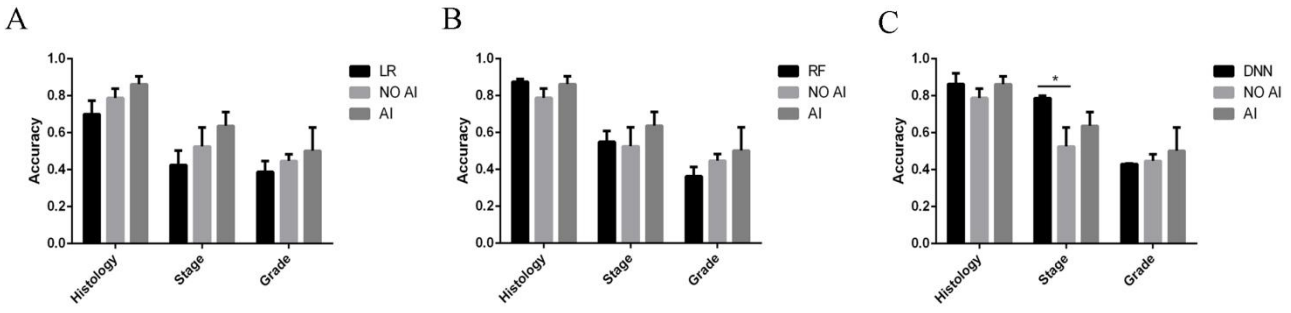


Figure 2. the accuracy comparison between whether doctors with and without AI assistance and AI in predict stage, and grade.A,,B,C shows the different AI assistance model.

The comparison of a doctors' prediction with and without AI assistance is shown in Fig.2. Compared to LR (Fig.2A), the accuracy of doctors' prediction with AI is higher than that of LR and doctors' prediction without AI among histology, stage, and grade. The comparison with RF (Fig.2B) also showed similar results. However, the accuracy of the DNN's prediction of the stage was significantly higher than that of doctors' prediction with and without AI assist (Fig.2C). But the accuracy of the combination of doctor and AI was relatively better as a whole.

Discussion

Endometrial cancer is a relatively common gynecological tumor. The development and application of AI in the medical field has gradually generated significance and value. This study built AI models to predict histology, stage, and grade of EC. Besides the prediction of AI models, we also compared the AI models, doctors' predictions, and doctors' predictions assisted by the AI model.

From the point of AUC alone, LR and RF models perform better in the prediction of histology and grade. RF is better in the prediction of the stage (Figure 1). If only accuracy is considered, DNN and RF models work well in the prediction of histology and grade (Table 2). In the real world, not every patients can complete all examinations. In this way, the patients with missing values were also included in the dataset. However, compared with RF and DNN, the LR is sensitive to missing values, which means the missing values will significantly influence the performance of LR[15]. On the other hand, though DNN with hidden layers has more capability to learn from nonlinear and complex relationships. But tt has higher requirements for the sample size of training data than LR and RF[16].

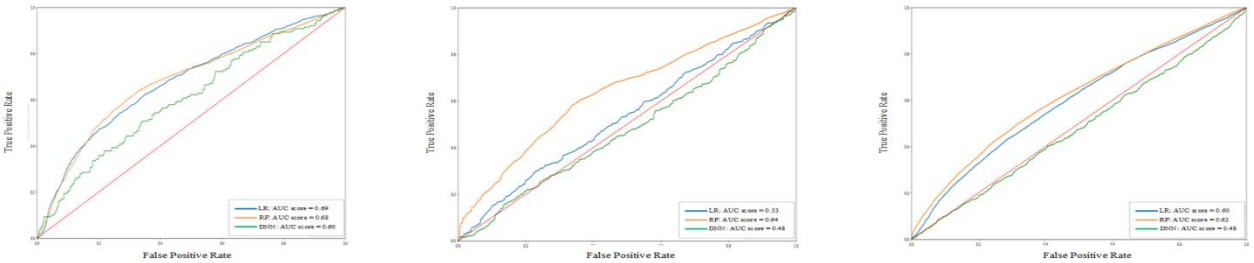


Figure 1. The ROC curve of the histology stage and grade between different models. (a,b,c) shows the ROC curve and AUC score of three different models for histology, stage and grade prediction, respectively

Model	Histology		Stage		Grade	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
LR	0.69(0.67-0.70)	0.74(0.72-0.75)	0.56(0.54-0.59)	0.42(0.41-0.44)	0.61(0.60-0.62)	0.36(0.35-0.38)
RF	0.69(0.67-0.70)	0.81(0.79-0.82)	0.66(0.64-0.69)	0.63(0.61-0.65)	0.64(0.64-0.65)	0.43(0.41-0.44)
DNN	0.60(0.54-0.65)	0.83(0.75-0.90)	0.48(0.46-0.51)	0.78(0.71-0.84)	0.47(0.45-0.50)	0.43(0.40-0.45)

Notes: LR,Logic Regression; RF,Random Forest ;DNN, Deep Neural Network ; AUC, Area Under the Curve ;

Taking into account of above reasons, the RF model was relatively better than other models, so RF was used to assist doctors.

The doctor's clinical experience combined with the assistance of AI increases the accuracy of histology, stage, and grade (Table 3). The main reason is that doctors analyze the highly relevant features of the disease (such as BMI, D&C, imaging, etc.) based on their clinical experience and draw conclusions, while the algorithm learns the influence weights of different features according to the distribution of training data, and more accurate judgments can be obtained for some patients who are not obvious in the preoperative features. And overall, the accuracy of doctors with AI assistance is relatively the best choice among the histology, stage, and grade whether compared to AI alone or doctor alone (Figure 2). Therefore, the judgment of the doctor with the RF assistance is the best choice.

The accuracy of grade and stage is not that high, and the AUC is also relatively low. The reasons can be: 1. The pathological results of preoperative curettage are not completely accurate, and there are false negatives [3]; 2. The staging of endometrial cancer is the clinicopathological stage, the determination of staging requires a combination of preoperative conditions, staged surgery and postoperative pathology, as well as grade, but the aim of this study is the preoperative diagnosis, so only preoperative features are given to AI models and doctors, and the intraoperative and postoperative characteristics were not included. Despite this, the AUC of RF is greater than 0.6 among histology, stage, and grade, so it has predictive value, especially given that it is only based on preoperative features.

Furthermore, in the past years, there is general agreement that AI may assist physicians to make better clinical decisions. This technology can provide additional information to help doctors make proper diagnoses.[17] In the classification of grade, the outcome of AI alone and doctor alone is not very good, but doctors' prediction including the AI results improved the accuracy. In the classification of histology, both doctors and AI had high accuracy, but the accuracy of doctors combined with AI was improved. The same is true for staging. The accuracy of staging is not high, but doctors combined with AI improved the accuracy. Compared with without AI assist , the time consumption for doctors with AI assist is only 10 seconds longer, only 0.5 second per patient 1.7% longer than before, which can be seen as almost no additional time cost. The extension of time consumption is not because of the speed predicted by AI, but because doctors need to analyze the information from AI.

Therefore, the AI model we built can effectively assist doctors in preoperative diagnosis and prediction of histology, stage, and grade.

There are several limitations to this study. Some multi-category classifications, such as staging and differentiation have small sample sizes, resulting in poor overall performance. This was a single-center (country) study and an independent validation set from another country can make the results more convincing. Prospective, multi-

center, large sample size research will help improve the performance of this AI model. In addition, the features of the database are mainly derived from text information, and the dimension of information should be improved. In the future, more dimensional information can be directly extracted from the images and examinations, so that intuitive information can be extracted.

Conclusions

This study demonstrated that a random forest model can predict histology, stage, and grade of endometrial cancer preoperatively and help doctors in obtaining a better diagnosis and predictive results with minimal additional time, which can help patients receive timely, appropriate and effective treatment.

Reference

1. Hyuna Sung JF, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, Freddie Bray: Global Cancer Statistics 2020 GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA CANCER J CLIN* 2021, 71:209-249.
2. Martin Koskas FA, Mansoor Raza Mirza, Carien L Creutzberg: Cancer of the corpus uteri: 2021 update. *Int J Gynaecol Obstet* 2021, 155 45-60.
3.) NCCNN: Uterine Neoplasms *NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®)* 2021, Version 1.2022 — November 4, 2021.
4. Kong A JN, Kitchener HC, Lawrie TA: Adjuvant radiotherapy for stage I endometrial cancer. *The Cochrane Database of Systematic Reviews* 2012, 4:CD003916.
5. B Karlsson SG, M Wikland, P Ylöstalo, K Torvid, K Marsal, L Valentin: Transvaginal ultrasonography of the endometrium in women with postmenopausal bleeding--a Nordic multicenter study. *Am J Obstet Gynecol* 1995, 172:1488-1494.
6. Osheroff JA, Teich JM, Levick D, Saldana L, Velasco FT, Sittig DF, Rogers KM, Jenders RA: *Improving outcomes with clinical decision support: an implementer's guide.* Himss Publishing; 2012.
7. Rajkomar A, Dean J, Kohane I: Machine Learning in Medicine. Reply. *N Engl J Med* 2019, 380:2589-2590.
8. Hu J, Niu H, Carrasco J, Lennox B, Arvin F: Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning. *IEEE Transactions on Vehicular Technology* 2020, 69:14413-14423.
9. Otani S, Himoto Y, Nishio M, Fujimoto K, Moribata Y, Yakami M, Kurata Y, Hamanishi J, Ueda A, Minamiguchi SJMri: Radiomic machine learning for pretreatment assessment of prognostic risk factors for endometrial cancer and its effects on radiologists' decisions of deep myometrial invasion. 2022, 85:161-167.
10. Nakajo M, Jinguji M, Tani A, Kikuno H, Hirahara D, Togami S, Kobayashi H, Yoshiura TJMI, Biology: Application of a Machine Learning Approach for the Analysis of Clinical and Radiomic Features of Pretreatment [18 F]-FDG PET/CT to Predict Prognosis of Patients with Endometrial Cancer. 2021:1-10.
11. Berek JS, Kehoe ST, Kumar L, Friedlander M: Cancer of the ovary, fallopian tube, and peritoneum. *International Journal of Gynecology & Obstetrics* 2018, 143:59-78.
12. Wright RE: Logistic regression. 1995.
13. Breiman LJMl: Random forests. 2001, 45:5-32.
14. LeCun Y, Bengio Y, Hinton GJn: Deep learning. 2015, 521:436-444.
15. Amini P, Maroufizadeh S, Hamidi O, Samani RO, Sepidarkish MJE, Biostatistics, Health P: Factors associated with macrosomia among singleton live-birth: A

comparison between logistic regression, random forest and artificial neural network methods. 2016, 13.
16. Yoo W, Ference BA, Cote ML, Schwartz A: A Comparison of Logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Environmental Interactions. *Int J Appl Sci Technol* 2012, 2:268.
17. Haleem A, Javaid M, Khan IH: Current status and applications of Artificial Intelligence (AI) in medical field: An overview. *Current Medicine Research and Practice* 2019, 9:231-237.

Chapter 8: Discussion

In this thesis, I investigated two applications of artificial intelligence in oncology. First, Artificial intelligence (AI) applications for data augmentation (Chapters 2,3,4). Second, applications of AI algorithms in clinical decision-making (Chapter 5,6,7).

Although AI plays a significant role as a powerful tool in the medical field, the requirement for data quantity and quality is one of the biggest challenges for clinical AI applications. State-of-art data augmentation methods, such as generative adversarial network (GAN) represent a potential solution to the above-mentioned problems.

However, the variety of formats and contents of medical data is so heterogeneous that using a single unified approach to generating diverse types of data is not an optimal solution. Identifying and selecting the appropriate data augmentation method remains challenging and lacks consensus among the computer science community.

Therefore, in Chapter 2 I reviewed the applications of GANs in radiotherapy proposed in the last five years and categorize them according to the ultimate task (such as generation of synthetic CT and RT-derived data). GAN models can automatically learn the anatomical features from different modalities of images to generate synthetic images. However, due to its design, a GAN still has several drawbacks. First, it is prone to pattern collapse during the training process, resulting in a lack of diversity of generated samples. Second, it is difficult to control the details due to the large size of the generated image.

Consequently, in order to address these issues, I designed a dual-discriminator GAN model to generate synthetic ground glass nodules (GGN) in Chapter 3. This structure increases the model complexity and sample space by utilizing a dual discriminator, thus reducing the occurrence of pattern collapse in the training process. Moreover, it enables the model to generate data while retaining local details and global images at the same time. It is worth mentioning that in the method of evaluating the quality of generated images, I designed a Visual Turing Test (VTT) method to evaluate the quality of generated images. This method invited radiologists to participate in the evaluation of image quality and measure the authenticity of generated images. Nonetheless, the proposed method depends on hand-picking specific areas of interest (ROI), and therefore, it can only be employed when annotated data is accessible.

This defects inspired the research in Chapter 4, where I designed a self-attention GAN. The experimental results illustrated that the self-attention mechanism can automatically select the ROI, so that the generated data samples retain the original structure of organs and create clearer details. Finally, I evaluated the accuracy gain of

the synthetic data on the AI classification model, demonstrating that it can address the shortage of insufficient data for AI and improve the diagnosis accuracy of a classification model.

The above research in Chapters 2,3,4 proves the possibility of AI as the data augmentation method to expand datasets.

At the same time, I realized that a reliable and robust AI model is not only necessary to obtain sufficient image data, but also to make efficient and accurate use of image information. To this end, I utilized the method of radiomics in Chapter 5 to convert the images into imaging features and set up a logical regression classification model, which proved that the quantified image can be used for effective prognosis prediction of Radiation Pneumonitis (RP), providing a theoretical basis for further research. Clinically speaking, the prediction model can accurately predict the RP risk for patients, which is useful for tailoring treatments and improving prognosis.

Nevertheless, imaging examinations are not applicable for all diseases. Thus, it is also important to use microscopic biomarkers, for example acquired from blood, prognosis and prediction. Hence, I attempted to use machine learning to automatically analyze biomarkers (e.g., preoperative circulating leukocytes) to predict the prognosis of ovarian cancer in Chapter 6. The experimental results demonstrated that the use of biomarkers can accurately predict the prognosis of cancer, which offers a way to avoid excessive imaging examination and treatment.

Ultimately, the purpose of an AI is to assist clinicians in the diagnosis and prediction of prognosis in the real world. So, I designed an experiment in Chapter 7 to explore the added value of using AI as a clinical decision support system. The experimental results showed that the AI diagnostic model can improve the accuracy of a clinician diagnosis as an assistant, which provides theoretical feasibility for the application of AI in the clinical field.

Overall, in Chapters 5,6,7, I explore the role of AI in clinical decision-making, showing that it can combine different types of data to assist physicians in clinical decision-making, while also reducing potential risks and helping patients receive timely, appropriate, and effective treatment.

Challenges

The quality of generated data

Even though GANs are gaining more and more popularity in the medical field, most of the models still present a high level of complexity compared to traditional DL

algorithms such as convolutional neural networks, which will bring more instability when generating data. This will significantly affect the quality of the generated image.

For example, there is no consensus on the most appropriate metric to be used to stop the training at the optimal point (global minimum of the loss function). It is susceptible to various artifacts and noises, resulting in a poor image quality. Especially when processing medical images, the risks of introducing novel, undesired artifacts and blurred images are not negligible. Preprocessing of input images and the postprocessing of generated images to decrease the artifacts and noises are worth exploring.

A low-quality image can also lead to credibility issues. The credibility of synthetic images generated by GANs has an important influence on their clinical application. Sorin V et.al (1) mentioned that the synthetic image generated by GANs can easily deceive radiologists, and determining real or fake images is a challenging task. And a deep learning model trained with a large amount of wrong synthetic data may lead to completely wrong results.

Moreover, high-quality data generation requires a large amount of dataset, which is a huge barrier for GANs applications(2).

The high requirement for hardware

Due to the GAN architecture being composed of two DL models, it has a dual hardware requirement during the training process. Especially in the task of 3D medical image generation, medical images such as CT and MRI always occupy a large memory, which sharply slows down the calculation speed, resulting in several weeks of training. This is unacceptable for most researchers. On the other hand, the higher requirement on hardware, makes deploying GAN models difficult in most hospitals. Model compression needs to be considered in future model design and deployment.

Evaluation metrics for synthetic data

GANs can learn and imitate the distribution of real data for image generation. As an unsupervised task, the generation task has no clear evaluation metrics such as accuracy, the Area Under the Curve (AUC), and mean Mean Squared Error (MSE), which play a huge role in guiding model training and evaluation. Usually, researchers evaluate the generated image quality according to the pixel distribution such as SSIM, PSNR, MSE et.al. The content information of the image is not available by the above metrics, which is particularly important for medical images. So, the Turing test by experts is always required for clinical image generation evaluation tasks to distinguish

whether the synthetic image is clinically meaningful or not. But the time required for experts to manually revise the data is the largest obstacle to performing Turing tests. Though in Chapters 3 and 4, I tried to summarize image evaluation metrics, the image content evaluation method (for example whether the tumor shape is reasonable) remains an important challenge.

Multi-modalities and Genomic data

Multi-modality data refers to a variety of medical data used for the in-depth study of disease mechanisms and phenotypes, which takes into account multiple forms and sources of data, such as MRI, CT, electroencephalograms, optical microscopy images, (3).

More and more studies have found that multi-modality data can provide complementary information for clinical decision-making. For example, multi-modality data can be used to extract medical data from different perspectives by computer vision methods, and Natural Language Processing algorithms to provide highly granular information(4). Furthermore, gene technologies can also play an important role in improving diagnosis and treatment plan in AI healthcare. Gene technologies can help doctors trace the relationship between genes and diseases. Gene analysis and AI technologies can be used to analyze the genome of the patients to improve healthcare decisions and care quality(5). However, the collection of genetic data is difficult due to its high cost, legal and privacy issues. In my previous work, I collected patient data (include biomarkers and medical images) from different countries and hospitals. In future work, I will incorporate gene data into the research, establish the mapping between genes and other modalities data through AI algorithm, to find out the expression of genes in different modalities such as the effect of gene expression on CT images.

The long-tail distribution

The long-tail distribution of data is a common phenomenon in machine learning tasks, that is, the minority of categories have a higher frequency while most categories have a lower frequency. In medical AI, the long-tail distribution of data can lead to insufficient and inaccurate training. This is because smaller attributes may be overlooked, and the model is not sensitive to complex patterns and possible anomalies.

In addition, a model based on this distribution may overfit the data because many variables of the same size are ignored. Finally, the long-tail distribution of data may lead to model biases, as many common variables will be emphasized and re-positioned while fewer variables may be overlooked. To address this problem, I designed some GAN models as data augmentation methods that can generate synthetic data, which

can help to solve part of the long-tail distribution problem. However, optimizing and iterating the algorithm from the perspective of decision-making is also worth exploring, such as few-shot learning and zero-shot learning.

Future perspectives

Model structure and loss function design

A GAN is composed of a generator and a discriminator, which are trained and compete alternately to make the generator create more realistic images. However, there is no direct loss function to train the generator, which causes the generator to not always achieve the expected effect. The model structure and loss function should be designed carefully to solve the non-convergence and pattern collapse. Optimizing the loss function is a significant challenge in deep learning, particularly in unsupervised tasks like GANs. This process can help reduce errors and improve the network's convergence speed during training. (6). In chapters 3 and 4, I tackled this challenge by optimizing the GAN model's structure and loss function. This modification allowed the model to focus more on the specific details of the image, rather than treating the entire image as a single entity.

Second, in traditional generation tasks, the image is always directly created from random noise, which makes it hard for the model to control the detail of generated image. The diffusion model proposed by Ho et. al(7) uses the model to reverse the process of gradually turning the image into noise. It became the hot point in 2022, because of its strong and stable generation capability. The key point of the diffusion model is that it tries to restore step-by-step the image from noise. This inspiration can also be used in GANs applications to make them more stable and powerful. That will possibly bring GAN applications closer to real-world clinical use.

Finally, most research on image generation is based on single timepoint 1D, 2D and 3D images, presently. The development of diseases or tumors over time in medical imaging is also an interesting question, and in future work using AI to predict the disease development image may be an interesting application.

Model compression

Except for more stable and high-quality generation, model compression is one of the most important directions. The smaller the model, the lower the hardware requirement, will make it usable to more hospitals. Meanwhile, the smaller model can sharply decrease the speed of forward calculation. Time consumption performance has a significant impact on clinical tasks, especially for auto-plan generation and medical

image super-resolution tasks. For example, some researchers achieve real-time applications by design of a smaller model like in Chapter 2, which is one of the model compression methods. Model pruning, quantization, and distillation can also be used to compress the model in future research.

Multi-center and federated learning

Multi-center medical data is composed of medical data collected and analyzed from different hospitals and research cases with the same standard indicators, allowing for the comparison of results between multiple centers and/or multiple research studies across different regions/locations (8). They facilitate conducting large-scale studies as more samples and greater data can be used from different distribution, allowing researchers to draw conclusions and make sound inferences more quickly, avoid over-fitting and make the model more close to real-world applications. They can also facilitate comparison between different regions while exploring different practices in different hospitals. Multi-center data can provide more information for AI to analyze and predict more accurately for accurate diagnoses, further improving the application of AI in the healthcare. However, multi-center data requires a lot of time and hardware costs for image transmission and storage; while clinical data is protected by privacy and law, many challenges and restrictions are faced when it comes to the collaboration of multiple centers. Federated learning is proposed(9), which is a distributed machine learning approach that combines models from multiple machines to create more accurate solutions. Its advantages include accelerated computation speed, reduced load on computer databases, protection of user privacy, and better support for big data processing tasks.

Few-shot/Zero-shot learning

The "Few-shot/zero-shot learning" algorithm can be used to build medical AI models in the case of insufficient sample size or imbalanced data, thus improving the generalization ability of the models (10). The basic idea of few-shot/zero-shot learning is to use some auxiliary data (usually non-labeled samples) for model training in the case of very few input samples, thus obtaining more accurate prediction results.

This technology can be used for diagnosing rare diseases, as the sample size for rare diseases is extremely limited. Using a few-shot/zero-shot learning algorithm can greatly improve the accuracy of diagnosing rare cancers, even when there's only a small amount of data available.(11). In addition, it can also be used in intelligent inspection technologies, such as the detection of heart disease, to more accurately identify the patterns of heart disease (12).

In my research, although I collected hundreds of samples to train the model, I found that for the prediction of some specific categories of data (Chapter 6), such as

pathology, it is difficult to collect samples of rare categories, the sample size is insufficient, and the commonly used machine/deep learning methods cannot effectively establish the prediction model. Facing this situation, few-shot learning would be the appropriate method. In my future research, I will try to develop an algorithm based on the few/zero shot learning method to make a predictive model in rare cases.

Conclusions

In this thesis, several AI applications for medical data augmentation applications were proposed and extensively verified though various experiments. The presented research showed the value of AI-based generation tasks, highlighted the benefits of deep learning models trained with synthetic data, and demonstrated the potential of AI-generated synthetic data for medical research and AI training. Meanwhile, the potential of AI diagnosis to make accurate decisions from different types of clinical data from macroscopic to microscopic has also been demonstrated. Finally, AI can be an auxiliary tool to support clinicians.

Reference

1. Sorin V, Barash Y, Konen E, Klang EJAr. Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) - A Systematic Review. 2020.
2. Frid-Adar M, Diamant I, Klang E, Amitai MM, Goldberger J, Greenspan HJN. GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification. 2018;321:321-31.
3. Amal S, Safarnejad L, Omiye JA, Ghanzouri I, Cabot JH, Ross EG. Use of Multi-Modal Data and Machine Learning to Improve Cardiovascular Disease Care. Front Cardiovasc Med. 2022;9:840262.
4. López-Úbeda P, Martín-Noguerol T, Juluru K, Luna A. Natural Language Processing in Radiology: Update on Clinical Applications. Journal of the American College of Radiology : JACR. 2022;19(11):1271-85.
5. Yeoh KG, Tan P. Mapping the genomic diaspora of gastric cancer. Nat Rev Cancer. 2022;22(2):71-84.
6. Zhang Y, Miao S, Mansi T, Liao RJMia. Unsupervised X-ray image segmentation with task driven generative adversarial networks. 2020;62:101664.
7. Ho J, Jain A, Abbeel PJAiNIPS. Denoising diffusion probabilistic models. 2020;33:6840-51.
8. Chu J, Chen J, Chen X, Dong W, Shi J, Huang Z. Knowledge-aware multi-center clinical dataset adaptation: Problem, method, and application. J Biomed Inform. 2021;115:103710.
9. McMahan B, Moore E, Ramage D, Hampson S, Arcas BAy. Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Aarti S, Jerry Z, editors. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics; Proceedings of Machine Learning Research: PMLR; 2017. p. 1273--82.
10. Ye HJ, Ming L, Zhan DC, Chao WL. Few-Shot Learning with a Strong Teacher. IEEE Trans Pattern Anal Mach Intell. 2022;Pp.
11. Xu Z, Niu K, Tang S, Song T, Rong Y, Guo W, et al. Bone tumor necrosis rate detection in few-shot X-rays based on deep learning. Comput Med Imaging Graph. 2022;102:102141.
12. Pałczyński K, Śmigiel S, Ledziński D, Bujnowski S. Study of the Few-Shot Learning for ECG Classification Based on the PTB-XL Dataset. Sensors (Basel). 2022;22(3).

Appendix

Summary

The application of Artificial Intelligence (AI) to solve medical tasks has seen remarkable progress in the recent years, yet a number of challenges persist concerning data augmentation and decision-making. In terms of data augmentation, the availability of large, high-quality unimodal and multi-modal clinical datasets is essential for deep learning but is often a significant obstacle. While traditional data augmentation methods such as flipping, rotating, zooming, etc. can increase the size of the dataset, they are insufficient to address data imbalance issues.

Generative Adversarial Networks (GANs) have shown great promise as a generative model for clinical data augmentation, capable of learning and imitating the distribution and features (e.g., shape, boundary, strength, texture, etc.) of each image to generate synthetic, but realistic data that does not exist in the original dataset. This can be used to generate simulated clinical data for training deep learning models. With regards to decision-making, the diversity of formats of medical data presents considerable challenges from macro image data to micro biomarkers. Different data sources have their own specific type and structure, which results in significant discrepancies in the format and content of the data. However, faced with different data and diseases, it is still necessary to design different models to deal with specific complex problems. In addition, the contribution of an AI-based diagnostic or prognostic model to clinical decision making still needs to be evaluated.

In this thesis, advances are made in AI applications in oncology for data augmentation and decision-making.

For the data augmentation, I mainly focus on the GAN model to generate clinical data as a data augmentation method. To further explore this, I tried to design an SRGAN (Super-Resolution GAN) for Ground Glass Nodule (GGN) generation, with two discriminators to capture global and local details during training processing. Qualitative visual Turing test (VTT) with clinicians and quantitative radiomics experiments were conducted to assess the similarity between GAN-generated and real lung lesions.

The results of these experiments indicate that GAN-generated data could be used to train and test junior doctors, particularly in hospitals without large datasets, established Picture Archiving and Communication Systems (PACS) or for privacy-preserving synthetic open datasets for research purposes. I also proposed a self-attention cycleGAN model to generate synthetic pneumonia from Radiation Pneumonitis (RP) and COVID-19 radiography datasets as a data augmentation method to solve data imbalance problems and create synthetic data with clearer details. A

classification model was trained with generated data, and it had an excellent performance on several classification metrics indicating that the method can significantly improve the accuracy of the classification model compared with traditional data augmentation methods. Therefore, the potential of the proposed method as a general data augmentation tool to assist in overcoming the sample imbalance problem in medical image datasets, is demonstrated.

In terms of decision-making, AI diagnosis models with different modalities of medical data source were developed. Bootstrap samples and a prospective validation set were used to validate a model which demonstrated that radiomic features extracted from the CT and the radiotherapy dose matrices could assist doctors in predicting of Radiation Pneumonitis (RP). Furthermore, a comprehensive nomogram was built to support clinical decision-making and personalized treatment. A decision tree model was then developed to identify the correlation between preoperative circulating leukocytes (such as MO/LY, differentiation status, CA125 level, NE, ascites cytology, LY%, and age) and ovarian cancer survival.

Finally, machine learning models were built and evaluated to provide a rapid diagnosis prediction and assist clinicians in providing effective treatment advice for endometrial carcinoma, with comparisons between clinicians with and without the assistance of models demonstrating the contribution of AI-based models.

Overall, this thesis has confirmed the hypothesis that AI-based techniques can yield high-caliber medical data suitable for instruction of machine learning and deep learning models, with the ultimate goal of improving their classification performance.

Furthermore, AI decision models can leverage different sources of medical data for diagnoses and provide support to clinicians in the field of oncology.

Samenvatting

De toepassing van Kunstmatige Intelligentie (KI) om medische taken op te lossen heeft opmerkelijke vooruitgang geboekt in de afgelopen jaren, maar er blijven nog steeds uitdagingen bestaan op het gebied van gegevensuitbreiding en besluitvorming. Met betrekking tot gegevensuitbreiding is de beschikbaarheid van grote, hoogwaardige unimodale en multimodale klinische datasets essentieel voor deep learning, maar is vaak een belangrijke hindernis. Traditionele methoden voor gegevensuitbreiding, zoals omdraaien, roteren, zoomen, enz., kunnen de omvang van de dataset vergroten, maar zijn ontoereikend om problemen met gegevensonevenwichtigheden aan te pakken.

Generative Adversarial Networks (GAN's) hebben veelbelovende resultaten getoond als generatief model voor gegevensuitbreiding in de klinische praktijk, waardoor synthetische maar realistische gegevens kunnen worden gegenereerd die niet in de oorspronkelijke dataset voorkomen. Dit kan worden gebruikt om gesimuleerde klinische gegevens te genereren voor de training van deep learning-modellen. Met betrekking tot besluitvorming presenteren de verschillende formaten van medische gegevens aanzienlijke uitdagingen van macrobeeldgegevens tot microbiomarkers.

In deze thesis worden vorderingen gemaakt in AI-toepassingen in de oncologie voor gegevensuitbreiding en besluitvorming.

Voor de gegevensuitbreiding richt ik me voornamelijk op het GAN-model om klinische gegevens te genereren als een methode voor gegevensuitbreiding. Om dit verder te verkennen, heb ik geprobeerd een SRGAN (Super-Resolution GAN) te ontwerpen voor Ground Glass Nodule (GGN) generatie, met twee discriminatoren om globale en lokale details vast te leggen tijdens het trainingsproces. Kwalitatieve visuele Turing-test (VTT) met clinici en kwantitatieve radiomics-experimenten werden uitgevoerd om de gelijkenis tussen GAN-generatie en echte longlaesies te beoordelen.

De resultaten van deze experimenten geven aan dat GAN-genereren gegevens gebruikt kunnen worden om junior artsen op te leiden en te testen, vooral in ziekenhuizen zonder grote datasets, gevestigde Picture Archiving and Communication Systems (PACS) of voor privacy-beschermende synthetische open datasets voor onderzoeksdoeleinden. Ik stelde ook een zelf-aandacht CycleGAN-model voor om synthetische longontsteking te genereren van stralingspneumonitis (RP) en COVID-19 radiografie datasets als een methode voor gegevensaugmentatie om gegevensonevenwichtigheden op te lossen en synthetische gegevens met duidelijkere details te creëren. Er werd een classificatiemodel getraind met gegenereerde gegevens, en het had uitstekende prestaties op verschillende classificatiemetrics, wat aangeeft

dat de methode de nauwkeurigheid van het classificatiemodel aanzienlijk kan verbeteren in vergelijking met traditionele methoden voor gegevensaugmentatie. Daarom wordt de potentie van de voorgestelde methode als algemeen hulpmiddel voor gegevensaugmentatie aangetoond om te helpen bij het overwinnen van het probleem van steekproefonevenwichtigheid in medische beeldgegevens.

Op het gebied van besluitvorming werden AI-diagnosemodellen ontwikkeld met verschillende modaliteiten van medische gegevensbron. Bootstrap-samples en een prospectieve validatieset werden gebruikt om een model te valideren dat aantoonde dat radiomische kenmerken, geëxtraheerd uit de CT- en radiotherapie-doseermatrices, artsen konden helpen bij het voorspellen van stralingspneumonitis (RP). Bovendien werd een uitgebreide nomogram gebouwd om klinische besluitvorming en gepersonaliseerde behandeling te ondersteunen. Vervolgens werd er een beslissingsboommodel ontwikkeld om de correlatie tussen preoperatieve circulerende leukocyten (zoals MO/LY, differentiatie-status, CA125-niveau, NE, ascites cytologie, LY%, en leeftijd) en ovariumkankersurvival te identificeren.

Ten slotte werden machine learning-modellen gebouwd en geëvalueerd om een snelle diagnosevoorspelling te bieden en clinici te ondersteunen bij het geven van effectief behandelingadvies voor baarmoederkanker, waarbij vergelijkingen tussen clinici met en zonder de hulp van modellen de bijdrage van op AI gebaseerde modellen aantonen.

Over het algemeen heeft deze scriptie de hypothese bevestigd dat op AI gebaseerde technieken medische gegevens van hoog kaliber kunnen opleveren die geschikt zijn voor het instrueren van machine learning- en deep learning-modellen, met als uiteindelijk doel hun classificatieprestaties te verbeteren. Bovendien kunnen AI-beslissingsmodellen verschillende bronnen van medische gegevens benutten voor diagnoses en ondersteuning bieden aan clinici op het gebied van oncologie.

Research Impact

Artificial intelligence (AI) has shown remarkable power in the medical field in applications such as diagnosis[1], tumor detection[2], organ segmentation[3]. However, the large data requirement and complex diagnosis processing hinders the application of AI in the clinic.

In this thesis, I studied AI applications for clinical data augmentation and how they play a role in diagnosis. These studies will have clinical, technological, societal and scientific impacts.

Clinical impact

This thesis examines the application of artificial intelligence (AI) in oncology at the level of data and decisions. Firstly, data augmentation using AI can help to improve the effectiveness of diagnostic or classification models for diseases that are not common and alleviate the problem of insufficient data to some extent (Chapter 2,3,4). Secondly, for clinical decision making, different sources of information such as medical images, clinical features and biomarkers can be valuable in the field of medical AI (Chapter 5,6). Finally, AI has great promise as an auxiliary tool to clinical diagnostics (Chapter 7). In summary, I proved that AI can generate high quality data and can support diagnostics in oncology.

Technological impact

There are several lessons that can help during Artificial intelligence (AI) model design and deployment. At the data level, Generative Adversarial Networks (GANs) are prone to gradient collapse during training due to a flaw in their fundamentals, where all generated samples are concentrated in the same class. In this case, the design of discriminators and loss functions can be effective in reducing this situation (Chapter 3). Secondly, if the quality of the images generated by the GAN is unsatisfactory, focusing on optimizing the generator such as adding an attention module or modifying the generator loss function can improve the image quality (Chapter 4). At the decision-making level, the importance and data types of the different modalities differ, and the data pre-processing part should be paid attention to before using multimodal data, unifying the different data types and magnitudes to avoid the excessive impact of a single data (Chapter 5,6,7).

Societal impact

In this thesis, I demonstrate the potential of AI for clinical applications in oncology. Data augmentation methods can lower the threshold for deployment of AI models, expanding the range of diseases to which they can be applied, reducing upfront

preparation time, and speeding up the development process. AI can significantly improve the efficiency of doctors and reduce their workload, thereby improving the efficiency of the healthcare system. Finally, this thesis presents experiments on the clinical application of AI, demonstrating that AI as a tool can significantly improve the diagnostic accuracy of clinicians and safeguard the lives of patients.

Scientific impact

First, all studies are open access and are published in scientific and professional journals with high impact factors (e.g., International journal of radiation oncology, IEEE Access, European Radiology Experimental, Precision Cancer Medicine, Journal of ovarian research) that have more influence and transmissibility in the scientific community. Second, through this thesis, I built a strong connection between the Chinese hospitals and Maastricht University, which will promote the cooperation of scientific researchers between the two countries in the future and contribute to international academic development and cooperation. Third, all the code and projects have followed the tenet of open science and open source. This may help promote interdisciplinary research, obtain technical support, reduce the time spent on academic research in the same field, and improve the productivity of researchers.

References

1. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, Allison T, Arnaout O, Abbosh C, Dunn IF, Mak RH, Tamimi RM, Tempany CM, Swanton C, Hoffmann U, Schwartz LH, Gillies RJ, Huang RY, Aerts H (2019) Artificial intelligence in cancer imaging: Clinical challenges and applications. CA: a cancer journal for clinicians 69 (2):127-157. doi:10.3322/caac.21552

2. Lin DJ, Johnson PM, Knoll F, Lui YW (2021) Artificial Intelligence for MR Image Reconstruction: An Overview for Clinicians. Journal of magnetic resonance imaging : JMRI 53 (4):1015-1028. doi:10.1002/jmri.27078

3. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G (2020) Artificial Intelligence in Anesthesiology: Current Techniques, Clinical Applications, and Limitations. Anesthesiology 132 (2):379-394. doi:10.1097/aln.0000000000002960

List of Publications

Published research on international journals (1 (shared) first author, * (shared) corresponding author)

1. Zhen Zhang[1], **Zhixiang Wang**, Meng Yan, Jiaqi Yu, Andre Dekker, Lujun Zhao, Leonard Wee. Radiomics and Dosiomics Signature from Whole Lung Predicts Radiation Pneumonitis: A Model Development Study with Prospective External Validation and Decision-Curve Analysis. Int J Radiat Oncol Biol Phys 2022, S0360-3016(22)03189-3. https://doi.org/10.1016/j.ijrobp.2022.08.047.

2. Zhen Zhang[1], **Zhixiang Wang**[1], Tianchen Luo, Meng Yan, Andre Dekker, Dirk De Ruysscher, Alberto Traverso, Leonard Wee, Lujun Zhao Computed tomography and radiation dose images-based deep-learning model for predicting radiation pneumonitis in lung cancer patients after radiation therapy[J]. Radiotherapy and Oncology, 2023: 109581. doi: 10.1016/j.radonc.2023.109581

3. **Zhixiang Wang**[1], Zhen Zhang[1], Ying Feng, Lizza E. L. Hendriks, Razvan L. Miclea, Hester Gietema, Janna Schoenmaekers, Andre Dekker, Leonard Wee and Alberto Traverso. Generation of Synthetic Ground Glass Nodules Using Generative Adversarial Networks (GANs). Eur Radiol Exp 2022, 6 (1),59.https://doi.org/10.1186/s41747-022-00311-y.

4. Ying Feng[1], **Zhixiang Wang**[1], Meizhu Xiao, Jinfeng Li, Yuan Su, Bert Delvoux, Zhen Zhang, Andre Dekker, Sofia Xanthoulea, Zhiqiang Zhang, Alberto Traverso, Andrea Romano, Zhenyu Zhang, Chongdong Liu, Huiqiao Gao, Shuzhen Wang and Linxue Qian. An Applicable Machine Learning Model Based on Preoperative Examinations Predicts Histology, Stage, and Grade for Endometrial Cancer. Front Oncol 2022, 12, 904597. https://doi.org/10.3389/fonc.2022.904597.

5. Junzhuo Liu[1], **Zhixiang Wang**[1], Ye Zhang, Alberto Traverso, Andre Dekker, Zhen Zhang * and Qiaosong Chen *. CycleGAN Clinical Image Augmentation Based on Mask Self-Attention Mechanism. IEEE Access 2022, 10, 105942–105953. https://doi.org/10.1109/ACCESS.2022.3211670.

6. Ying Feng[1], **Zhixiang Wang**[1], Ran Cui, Meizhu Xiao, Huiqiao Gao, Huimin Bai, Bert Delvoux, Zhen Zhang, Andre Dekker, Andrea Romano, Shuzhen Wang, Alberto Traverso, Chongdong Liu and Zhenyu Zhang. Clinical analysis and artificial intelligence survival prediction of serous ovarian cancer based on preoperative

circulating leukocytes. J Ovarian Res 15, 64 (2022). https://doi.org/10.1186/s13048-022-00994-2

7. Wen Q[1], **Wang Z**[1], Traverso A, Liu Y, Xu R, Feng Y, Qian L. A radiomics nomogram for the ultrasound-based evaluation of central cervical lymph node metastasis in papillary thyroid carcinoma. Front Endocrinol (Lausanne). 2022 Nov 30;13:1064434. doi: 10.3389/fendo.2022.1064434. PMID: 36531493; PMCID: PMC9748155.

8. Feng Y[1], **Wang Z**[1], Xiao M, Li J, Su Y, Delvoux B, Zhang Z, Dekker A, Xanthoulea S, Zhang Z, Traverso A, Romano A, Zhang Z, Liu C, Gao H, Wang S, Qian L. An Applicable Machine Learning Model Based on Preoperative Examinations Predicts Histology, Stage, and Grade for Endometrial Cancer. Front Oncol. 2022 May 30;12:904597. doi: 10.3389/fonc.2022.904597. PMID: 35712473; PMCID: PMC9196302.

9. Pengfei Sun, Ying Feng, Chen Chen, Andre Dekker, Linxue Qian, **Zhixiang Wang*** & Jun Guo * An AI model of sonographer's evaluation+ S-Detect + elastography + clinical information improves the preoperative identification of benign and malignant breast masses. Front Oncol. 2022 Nov 11;12:1022441. doi: 10.3389/fonc.2022.1022441. PMID: 36439410; PMCID: PMC9692079.

10. **Zhixiang Wang**[1], Glauco Lorenzut[1],Zhen Zhang , Andre Dekker,Alberto Traverso *Applications of generative adversarial networks (GANs) in radiotherapy: narrative review[J]. 2022. Precision Cancer Medicine.https://dx.doi.org/10.21037/pcm-22-28

Submitted research:

1. Ye Zhang[1], **Zhixiang Wang** [1], Zhen Zhang , Junzhuo Liu , Ying Feng , Leonard Wee , Andre Dekker , Qiaosong Chen *& Alberto Traverso * 'GAN Based One-Dimension Medical Data Augmentation'(Soft Computing , under revision)

2. **Zhixiang Wang**[1],Ying Feng[1], Zhen Zhang, Yupeng Deng, Meizhu Xiao, Zhiqiang Zhang, Andre Dekker, Shuzhen Wang, Linxue Qian, Alberto Traverso*, Chongdong Liu*, Zhenyu Zhang 'Auto-segmentation of the ureter and uterine artery in video sequences of laparoscopic hysterectomy' (American Journal of Obstetrics & Gynecology , under review)

3. **Zhixiang Wang** , Huiqiao Gao , Jinfeng Li , Hengzi Sun , Yidi Ma , Xuefang Zhang , Zhen Zhang , Andre Dekker , Alberto Traverso , Zhenyu Zhang , Linxue Qian * , Meizhu Xiao *& Ying Feng * 'A Multi-Task Learning-based applicable AI model precisely predicts stage, histology, grade and LNM for cervical cancer' (Obstetrics & Gynecology , under review)

Conference:
Oral:
1. 'Three Branch net Lane Detection on Complex Road Conditions' IEEE SMC 2019
2. 'Generative adversarial networks based ground glass opacities (GGOs) synthetic' ECMP 2022
3. 'Radiomics and dosiomics signature from whole lung predicts radiation pneumonitis: a model development study with prospective external validation and decision-curve analysis' ESTRO 2022

Poster:
4. 'Dual discriminator Super-Resolution Generative Adversarial Network-based synthetic GGO nodule image augmentation' ESTRO 2021
5. 'Synthetic generation of pulmonary nodules using super resolution generative adversarial model' ESTRO 2022
6.'Generation of synthetic radiation pneumonitis images using deep learning' ESTRO 2023
7.'Deep transfer features for radiation pneumonitis prediction' NACP 2023

Acknowledgments

I would like to take this opportunity to express my sincere gratitude to the many people who have supported me throughout my PhD journey. First and foremost, I am immensely grateful to my three supervisors whose kindness, care, and dedication filled me with happiness and joy during my doctorate studies. I consider myself fortunate to have spent my doctoral years in our laboratory, learning from the best in the field.

I owe a great debt of gratitude to Andre Dekker for giving me the chance to join Maastricht University and for changing the trajectory of my life. His support, encouragement, and provision of ample space for me to pursue my research interests freely were instrumental in my success.

I am also deeply grateful to my daily supervisors, Alberto Traverso and Leonard Wee. Alberto has always been prompt in providing me with necessary assistance whenever I required it. His constant guidance and care have been invaluable during moments of confusion. Similarly, Leonard's rigorous scientific approach and extensive technical knowledge have been a source of inspiration to me. I am grateful for their supervision and support throughout my research journey.

I would like to extend my appreciation to my colleagues and friends. Their constant support and encouragement have been crucial in helping me achieve my goals. It has been a privilege to work and learn from them and grow together in our research. They have taught me valuable qualities that will remain with me throughout my life.

I owe a special debt of gratitude to my parents Tingyin Wang and Hongyan Li for their unwavering support and encouragement, which has sustained me throughout my studies.

Finally, I am grateful to the experts Prof. Frank Verhaegen, Dr. Cecile Wolfs, Prof. Marius Staring, and Dr. Zhenwei Shi of my thesis assessment committee for taking the time to read and improve my thesis. Their feedback and guidance have been immensely helpful in enhancing the quality of my work.

Thank you all for your invaluable contributions and support.

Curriculum Vitae

Zhixiang was born on the 14th of September 1993 in Liaoning, China. In 2012, he studied applied physics at the Chongqing University of Posts and Telecommunications (CQUPT). In 2017, he started his master's study of computer technology at CQUPT. He focuses on computer science and artificial intelligence development, and accumulated experience in the work with technology companies.

After obtaining his master's degree in 2020, he started a Ph.D. program funded by the China Scholarship Council at Maastricht University in The Netherlands at the Faculty of Health, Medicine and Life Sciences, under the supervision of Prof. Andre Dekker.

His main research interests are artificial intelligence applications in the medical field.

Here, he cooperated with doctors, which enhanced his medical knowledge and experience, allowing him to complete the Ph.D. program in less than three years. In Beijing Friendship hospital, he was offered a position as a researcher at Capital Medical University