Generative models improve radiomics

Generative Models Improve Radiomics Reproducibility and Performance in Low Dose CTs

Dissertation

To obtain the degree of Doctor at 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 Monday 3 July 2023 at 10.00 hrs.

Junhua Chen

陈军华

born on July 5,1994

in Shaoxing, China

Supervisor:

Prof. dr. ir. Andre Dekker

Co-supervisor:

Dr. Iñigo Bermejo

Dr. Leonard Wee

Assessment Committee

Prof. Joachim Wildberger (chair), Maastricht University

Prof. Bram van Ginneken, Radboud University Medical Center

Dr. Wouter van Elmpt, Maastricht University/Maastro Clinic

Prof. Nico van den Berg, University Medical Center Utrecht

The work of this thesis has been funded through the China Scholarship Council (File No. 201906540036)

CONTENT

CHAPTER 1 INTRODUCTION 5

CHAPTER 2 ARE ALL SHORTCUTS IN ENCODER-DECODER NETWORKS BENEFICIAL FOR CT DENOISING? 23

CHAPTER 3 GENERATIVE MODELS IMPROVE RADIOMICS REPRODUCIBILITY IN LOW DOSE CTS: A SIMULATION STUDY 49

CHAPTER 4 LUNG CANCER DIAGNOSIS USING DEEP ATTENTION BASED MULTIPLE INSTANCE LEARNING AND RADIOMICS 91

CHAPTER 5 GENERATIVE MODELS IMPROVE RADIOMICS PERFORMANCE IN DIFFERENT TASKS AND DIFFERENT DATASETS: AN EXPERIMENTAL STUDY 123

CHAPTER 6 IMPROVING REPRODUCIBILITY AND PERFORMANCE OF RADIOMICS IN LOW DOSE CT USING CYCLE GANS 155

CHAPTER 7 GENERAL DISCUSSION AND CONCLUSIONS 203

CHAPTER 8 APPENDICES 225

APPENDIX I SUMMARY 227

APPENDIX II IMPACT PARAGRAPH 235

APPENDIX III ACKNOWLEDGEMENTS 239

APPENDIX IV CURRICULUM VITAE 241

APPENDIX V LIST OF PUBLICATIONS 243

Chapter 1
Introduction

General Introduction

Developing medical imaging methods provide opportunities to assess the characteristics of human tissue noninvasively and continuously. Disease diagnosis and treatment planning based on medical imaging is common in modern medicine.[1][2] However, humans' unassisted discrimination and computation ability and inter- and intra-individual variations reduce the reproducibility, reliability, and accuracy of qualitative disease diagnosis, especially for subtle lesions.[3] Quantitative medical image analysis (QMIA) can overcome the mentioned shortcomings due to computers' strong computation ability, even though humans may still have better interpretation ability. [4] QMIA models established directly from medical images have poor generalization performance due to the high dimensionality of medical imaging and the limited volume of labelled data for model training. Therefore, features which represent the information of images with much lower dimension are of interest. [5][6]

Radiomics [7] is a popular QMIA framework and an active research topic focusing on extracting a high number of features from medical images for multiple clinical applications in oncology, such as tumor phenotype decoding, [8] survival prognostic prediction, [9] diagnostic differentiation of suspected tissue, [10] etc. Radiomic features are quantitative descriptions of the intensity, shape, volume, and texture of the region of interest (ROI). Shape and volume related features are calculated based on the mask of the ROI, while intensity related features are calculated based on descriptive statistics from the intensity histogram, and texture related features are calculated based on the non-uniform spatial disposition of pixel intensities.

The definition of each feature can be found elsewhere [11]. Radiomics has shown the potential for clinical-decision support in a range of cancers (lung cancer, [1] head and neck cancer, [12] rectal cancer [13]) and using most common clinical imaging modalities such as computed tomography (CT), [1][14] magnetic resonance imaging (MRI), [15] and positron emission tomography (PET) [16].

Compared with conventional radiography, CT scanning can provide images of many types of tissue with high resolution in 3 dimensions and even in 4 dimensions for some applications (e.g., 4D CT film clip of a beating heart [17]). Moreover, CT examinations are non-invasive, fast, simple and can be used both in outpatient and emergency patient care. Advantages of CT made it to be one of major examinations in clinical practices and it the third most used medical imaging modality behind plain radiography and ultrasound [18]. More specifically, over 1.29 million CT scans were collected in the Netherlands in 2012 [19] and 80 million in the United States in 2015 [20].

Considering the wide implementation of CT in clinical practice, CT based radiomics has a big potential in clinical practice for decision support. [21] More specifically, CT radiomics can be used for lung cancer screening, [22] tumor treatment outcome prediction, [23] tumor metastasis prediction, [24] etc. Hundreds CT radiomics related studies have been published during last decade. [25][26] In other words, as a hot research topic, CT radiomics has attracted much attention from researchers.

Due to the long-term risk posed by low levels of ionizing radiation exposure, low dose CTs have become more popular (As Low As Reasonably Achievable (ALARA) principle [27]) in clinical practice, especially for

screening and monitoring of populations at risk. In 2013, the US Preventive Services Task Force (USPSTF) recommended annual screening for lung cancer with low-dose CT in adults aged 55 to 80 years who have a 30–pack-year smoking history and currently smoke or have quit within the past 15 years. [28] Low dose CT radiomics and related applications can make screening quicker and more reliable. Therefore, researchers have started to calculate radiomic features based on low dose CTs. [29][30]

Although significant progress has been made during recent years in CT radiomics, there are still important barriers that prevent the widespread implementation of radiomics in clinical settings. For example, the low repeatability and reproducibility of CT radiomic features have been shown in multiple studies. [31][32] Repeatability mainly refers to features that remain unchanged when capturing images multiple times of the same subject under the same conditions. The repeatability of radiomic features is assessed in test-retest analyses [33] and interobserver variability for tumor contouring. [34] Reproducibility of radiomics refers to stability of features when at least one condition is changed. Common metrics for measuring repeatability and repeatability of features are intraclass correlation coefficient (ICC), [35] concordance correlation coefficient (CCC), [36] and Spearman correlation coefficient. [37]

The reproducibility of CT radiomics can be reduced by multiple variations during the whole radiomics extraction and application workflow (as shown in Figure 1-1). [38] The effect of some factors to reduce the reproducibility of radiomic features has been quantitatively studied, such as vendor scanner variability, [39] mask inter-observer delineation variability, [40] features

extraction variability, [41] radiation dose, [42] CT acquisition parameters and reconstruction settings, [43] etc. In low dose CTs, radiation dose is the majority factor that decreases radiomics reproducibility.

Published studies show that improving CT radiomics reproducibility will enhance the performance of CT radiomics in various applications [44] and increase the chances of CT radiomics' widespread implementation in clinical settings, especially for low dose CT images.

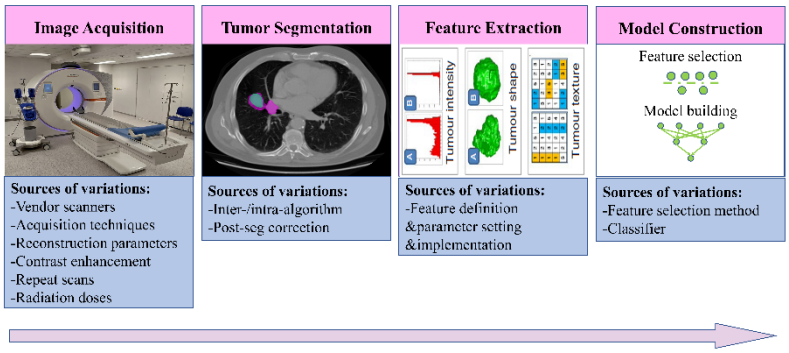


Figure 1-1. Radiomics workflow and variations to reduce radiomics reproducibility. (**Figure reproduced from [38]**)

Different studies have made various efforts to improve CT radiomics reproducibility. For examples, Zwanenburg et al. [45] aimed to reduce effect of implementation bias to radiomics reproducibility. Therefore, they standardized the feature definition, parameter settings and extraction implementation, creating the Image Biomarker Standardization Initiative (IBSI). Jooae et al. [46] used deep learning methods to convert CT images from different reconstruction kernels into one kernel to eliminate reconstruction kernel bias for improving CT radiomics reproducibility.

Mahon et al. [47] used the ComBat algorithm [48] to harmonize multi-center CT radiomics features to eliminate scanner bias. However, very few studies have focused on improving CT radiomics reproducibility and performance in low dose CTs.

As a trade-off of low radiation exposure in low dose CT imaging, higher noise is present in these images. This noise decreases the image texture and the reproducibility of radiomics features. Comparing with radiomics features from high dose CTs, improving radiomics reproducibility and performance in clinical applications from low dose CTs is therefore a timely and potentially impactful research topic, which will be discussed in this thesis.

Improving Radiomics Performance in Low Dose CT

One potential solution worth exploring for improving the reproducibility and performance of radiomics based on low dose CT is denoising the images before extracting radiomic features.

Medical image denoising is a traditional topic in medical image analysis, that has attracted recent attention. Hundreds of related articles have been published in last decade and multiple reviews have summarized these papers. [49][50] Proposed methods can be divided into two categories - traditional denoising methods and deep learning based denoising methods. [49] One major advantage of deep learning based denoisers compared with traditional counterparts is the feature self-representation ability. In other words, domain knowledge such as noise distribution estimation, denoising kernel selection play a key role in traditional denoising methods. However, domain

knowledge as mentioned above is not necessary for deep learning based denoising methods. [51] Multiple studies and reviews have supported the conclusion that deep features-based methods outperform traditional methods in medical image denoising. [49][50]

According to a series of published articles, [52][53] generative models [54] -as a typical deep learning based denoiser- achieved the state-of-the-art (SOTA) in low dose CT denoising performance. Briefly, a generative model is a model that learns the distribution of the data in target domain and can generate new samples that obey the learned distribution based on random input or given data. A more in-depth definition of generative models can be found in some classic literature such as [55][56]. Typical types of generative models are Gaussian mixture models, [57] hidden Markov models, [58] encoder-decoder networks, [59] generative adversarial networks (GAN), [54] diffusion models, [60] etc. Both encoder-decoder networks and GANs belong to deep generative models, and they currently are the main methods for image synthesis. [61] The main difference between the two generative models is that game theory is introduced into GANs for pushing the model network to better learn the distribution of data and output more realistic images. [54]

Generative models have been widely used in medical image analysis for various applications such as segmentation, [62] denoising, [63] image synthesis, [65] registration [66] and more. [64][67] In this thesis, generative models are used as the denoisers to enhance image quality of low dose CT and increase radiomics reproducibility and performance. The rest of the dissertation is structured as follows:

Outline of the thesis

In Chapter 2, we investigate the necessity of shortcuts in encoder decoder networks for CT denoising. Shortcuts are an important part of generative models that connect the encoder and the decoder at different layers. These shortcuts in the networks can help communicate the information of semantic feature maps from the encoder to the decoder. However, in denoising, some semantic transferred by shortcuts from encoder is noise and as such undesirable for the decoder. Therefore, some shortcuts may not be always beneficial for CT denoising. The results of this chapter may provide some guidelines for better generative model design in following chapters.

In Chapter 3, we investigate the effect of SOTA denoising methods to improve radiomics reproducibility from low dose CT. Two generative models – Conditional Generative Adversarial Network (CGAN) [69] and encoder-decoder networks – are adopted as testing denoisers. Most of generative model training – including CGANs and encoder-decoder networks– need paired low dose and high dose CT images. However, collecting this kind of datasets is expensive and time-consuming, if at all possible. Therefore, the selected denoisers are trained based on simulated paired low dose-full dose CT images and then tested by assessing the improvement of radiomics reproducibility. (Chapter 3)

In Chapter 4, we test the performance of improved CT radiomics in a new application. We train a lung cancer classification model to classify at the subject (patient) level from multiple examined nodules, without the need to have specific expert findings reported at the level of each individual nodule. More specifically, this lung cancer diagnosis problem is regarded as a

multiple instance learning (MIL) problem, where lung nodules are regarded as instances in MIL and the diagnosis from the doctor at subject level is regarded as the label of the bag in MIL. CT radiomics are used as biomarkers to extract information from each nodule and deep attention-based MIL [68] is used as the classification algorithm at patient level.

In Chapter 5, we investigated the benefits of generative models for improved CT radiomics performance in real applications. We applied models trained on simulated data (Chapter 3) to denoise low dose CTs and used radiomic features extracted from the denoised CTs for lung cancer diagnosis and tumor pre-treatment survival prediction. The objective in survival analysis is to establish a connection between covariates and the time of an event or the probability of survival at any particular point in time. We used radiomic features as covariates in our study. The improvement on radiomics-based model performance is measured using multiple metrics in different applications.

In Chapter 6, we investigate the application of cycle-consistent adversarial networks (Cycle GAN) to CT denoising. Unlike the previously mentioned generative models, paired data is not mandatory for Cycle GAN training. [70] We trained a Cycle GAN both on paired simulated data and on unpaired real low dose-high dose CT images and applied this to same testing datasets and applications as mentioned in Chapter 3 and Chapter 5. The improvement of CT radiomics reproducibility and performance are compared with CGAN and encoder-decoder network to show if Cycle GAN are an improvement.

Finally, Chapter 7 consists of a general discussion about the potential of generative models for radiomics based on low dose CT, what barriers there

may be for their use in real world applications and potential solutions for these barriers. Moreover, methods other than medical image denoising to improve low dose CT radiomics performance are discussed in this final section.

The outline of the chapters is shown in Figure 1-2.

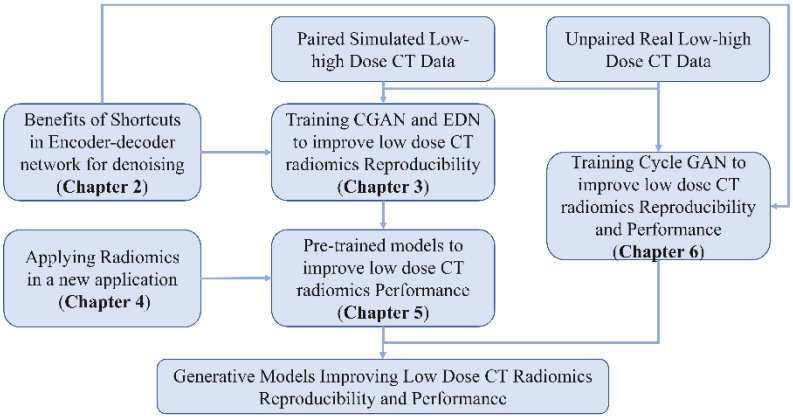


Figure 1-2. Outline of chapters in this thesis

References

[1]Kurland, Brenda F., et al. "Promise and pitfalls of quantitative imaging in oncology clinical trials." *Magnetic resonance imaging* 30.9 (2012): 1301-1312.

[2]Buckler, Andrew J., et al. "A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging." *Radiology* 258.3 (2011): 906-914.

[3]Giger, M. L., et al. "Computer-aided diagnosis: development of automated schemes for quantitative analysis of radiographic images." *Seminars in Ultrasound, CT, and MR*. Vol. 13. No. 2. 1992.

[4]Doi, Kunio. "Computer-aided diagnosis in medical imaging: historical review, current status and future potential." *Computerized medical imaging and graphics* 31.4-5 (2007): 198-211.

[5]Tourassi, Georgia D., et al. "Application of the mutual information criterion for feature selection in computer‐aided diagnosis." *Medical physics* 28.12 (2001): 2394-2402.

[6]Van Ginneken, Bram, Cornelia M. Schaefer-Prokop, and Mathias Prokop. "Computer-aided diagnosis: how to move from the laboratory to the clinic." *Radiology* 261.3 (2011): 719-732.

[7]Aerts, Hugo JWL, et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." *Nature communications* 5.1 (2014): 1-9.

[8]Yuan, Mei, et al. "Comparison of a radiomic biomarker with volumetric analysis for decoding tumour phenotypes of lung adenocarcinoma with different disease-specific survival." *European radiology* 27.11 (2017): 4857-4865.

[9]Cui, S-J., et al. "Role of imaging biomarkers for prognostic prediction in patients with pancreatic ductal adenocarcinoma." *Clinical Radiology* 75.6 (2020): 478-e1.

[10]Haider, Stefan P., et al. "Applications of radiomics in precision diagnosis, prognostication and treatment planning of head and neck squamous cell carcinomas." *Cancers of the head & neck* 5.1 (2020): 1-19.

[11]Van Griethuysen, Joost JM, et al. "Computational radiomics system to decode the radiographic phenotype." *Cancer research* 77.21 (2017): e104-e107.

[12]Bagher‐Ebadian, Hassan, et al. "On the impact of smoothing and noise on robustness of CT and CBCT radiomic features for patients with head and neck cancers." *Medical physics* 44.5 (2017): 1755-1770.

[13]Liu, Zhenyu, et al. "Radiomics analysis for evaluation of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer." *Clinical Cancer Research* 23.23 (2017): 7253-7262.

[14]Bogowicz, Marta, et al. "Stability of radiomic features in CT perfusion maps." *Physics in Medicine & Biology* 61.24 (2016): 8736.

[15]Zhang, Bin, et al. "Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma." *Clinical Cancer Research* 23.15 (2017): 4259-4269.

[16]Tixier, Florent, et al. "Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET." *Journal of Nuclear Medicine* 53.5 (2012): 693-700.

[17]Pan, Tinsu, et al. "4D‐CT imaging of a volume influenced by respiratory motion on multi‐slice CT." *Medical physics* 31.2 (2004): 333-340.

[18]England, N. H. S., and N. H. S. Improvement. "Diagnostic imaging dataset annual statistical release." (2017).

[19]Meulepas, Johanna M., et al. "Trends and patterns of computed tomography scan use among children in The Netherlands: 1990–2012." *European radiology* 27.6 (2017): 2426-2433.

[20]Brenner, David J. "Slowing the increase in the population dose resulting from CT scans." *Radiation research* 174.6b (2010): 809-815.

[21]Ibrahim, Abdalla, et al. "Radiomics analysis for clinical decision support in nuclear medicine." *Seminars in nuclear medicine*. Vol. 49. No. 5. WB Saunders, 2019.

[22]Gillies, Robert J., and Matthew B. Schabath. "Radiomics improves cancer screening and early detection." *Cancer Epidemiology and Prevention Biomarkers* 29.12 (2020): 2556-2567.

[23]Dai, Weixing, et al. "Prognostic and predictive value of radiomics signatures in stage I‐III colon cancer." *Clinical and translational medicine* 10.1 (2020): 288-293.

[24]Wang, Yue, et al. "CT radiomics nomogram for the preoperative prediction of lymph node metastasis in gastric cancer." *European radiology* 30.2 (2020): 976-986.

[25]van Timmeren, Janita E., et al. "Radiomics in medical imaging—"how-to" guide and critical reflection." *Insights into imaging* 11.1 (2020): 1-16.

[26]Song, Jiangdian, et al. "A review of original articles published in the emerging field of radiomics." *European journal of radiology* 127 (2020): 108991.

[27]Musolino, Stephen V., Joseph DeFranco, and Richard Schlueck. "The ALARA principle in the context of a radiological or nuclear emergency." *Health physics* 94.2 (2008): 109-111.

[28]Moyer, Virginia A., and US Preventive Services Task Force*. "Screening for lung cancer: US Preventive Services Task Force recommendation statement." *Annals of internal medicine* 160.5 (2014): 330-338.

[29]Choi, Wookjin, et al. "Radiomics analysis of pulmonary nodules in low‐dose CT for early detection of lung cancer." *Medical physics* 45.4 (2018): 1537-1549.

[30]Mao, Liting, et al. "Quantitative radiomic model for predicting malignancy of small solid pulmonary nodules detected by low-dose CT

screening." *Quantitative imaging in medicine and surgery* 9.2 (2019): 263.

[31]Traverso, Alberto, et al. "Repeatability and reproducibility of radiomic features: a systematic review." *International Journal of Radiation Oncology* Biology* Physics* 102.4 (2018): 1143-1158.

[32]Welch, Mattea L., et al. "Vulnerabilities of radiomic signature development: the need for safeguards." *Radiotherapy and Oncology* 130 (2019): 2-9.

[33]van Timmeren, Janna E., et al. "Test–retest data for radiomics feature stability analysis: generalizable or study-specific?." *Tomography* 2.4 (2016): 361-365.

[34]Huang, Qiao, et al. "Interobserver variability in tumor contouring affects the use of radiomics to predict mutational status." *Journal of Medical Imaging* 5.1 (2017): 011005.

[35]Bartko, John J. "The intraclass correlation coefficient as a measure of reliability." *Psychological reports* 19.1 (1966): 3-11.

[36]Lawrence, I., and Kuei Lin. "A concordance correlation coefficient to evaluate reproducibility." *Biometrics* (1989): 255-268.

[37]Myers, Leann, and Maria J. Sirois. "Spearman correlation coefficients, differences between." *Encyclopedia of statistical sciences* 12 (2004).

[38]Zhao, Binsheng. "Understanding sources of variation to improve the reproducibility of radiomics." *Frontiers in oncology* 11 (2021): 826.

[39]Starmans, Martijn, et al. "A multi-center, multi-vendor study to evaluate the generalizability of a Radiomics model for classifying prostate cancer: high grade vs. low grade." *Diagnostics* 11.2 (2021): 369.

[40]Pavic, Matea, et al. "Influence of inter-observer delineation variability on radiomics stability in different tumor sites." *Acta Oncologica* 57.8 (2018): 1070-1074.

[41]Foy, Joseph J., et al. "Variation in algorithm implementation across radiomics software." *Journal of medical imaging* 5.4 (2018): 044505.

[42]Meyer, Mathias, et al. "Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings." *Radiology* 293.3 (2019): 583-591.

[43]Berenguer, Roberto, et al. "Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters." *Radiology* 288.2 (2018): 407-415.

[44]Park, Ji Eun, et al. "Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives." *Korean journal of radiology* 20.7 (2019): 1124-1137.

[45]Zwanenburg, Alex, et al. "The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping." *Radiology* 295.2 (2020): 328-338.

[46]Choe, Jooae, et al. "Deep learning–based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses." *Radiology* 292.2 (2019): 365-373.

[47]Mahon, R. N., et al. "ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets." *Physics in Medicine & Biology* 65.1 (2020): 015010.

[48]Johnson, W. Evan, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics* 8.1 (2007): 118-127.

[49]Sagheer, Sameera V. Mohd, and Sudhish N. George. "A review on medical image denoising algorithms." *Biomedical signal processing and control* 61 (2020): 102036.

[50]Kollem, Sreedhar, Katta Rama Linga Reddy, and Duggirala Srinivasa Rao. "A review of image denoising and segmentation methods based on medical images." *International Journal of Machine Learning and Computing* 9.3 (2019): 288-295.

[51]LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436-444.

[52]Chen, Hu, et al. "Low-dose CT with a residual encoder-decoder convolutional neural network." *IEEE transactions on medical imaging* 36.12 (2017): 2524-2535.

[53]Yang, Qingsong, et al. "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss." *IEEE transactions on medical imaging* 37.6 (2018): 1348-1357.

[54]Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems* 27 (2014).

[55]Ng, Andrew, and Michael Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." *Advances in neural information processing systems* 14 (2001).

[56]Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science* 349.6245 (2015): 255-260.

[57]Rasmussen, Carl. "The infinite Gaussian mixture model." Advances in neural information processing systems 12 (1999).

[58]Eddy, Sean R. "What is a hidden Markov model?." *Nature biotechnology* 22.10 (2004): 1315-1316.

[59]Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

[60]Rogers, Everett M. "A prospective and retrospective look at the diffusion model." *Journal of health communication* 9.S1 (2004): 13-19.

[61]Bond-Taylor, Sam, et al. "Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models." *arXiv preprint arXiv:2103.04922* (2021).

[62]Dong, Xue, et al. "Automatic multiorgan segmentation in thorax CT images using U‐net‐GAN." *Medical physics* 46.5 (2019): 2157-2168.

[63]Kang, Eunhee, et al. "Cycle‐consistent adversarial denoising network for multiphase coronary CT angiography." *Medical physics* 46.2 (2019): 550-562.

[64]Kazeminia, Salome, et al. "GANs for medical image analysis." *Artificial Intelligence in Medicine* 109 (2020): 101938.

[65]Li, Wen, et al. "Magnetic resonance image (MRI) synthesis from brain computed tomography (CT) images based on deep learning methods for magnetic resonance (MR)-guided radiotherapy." *Quantitative imaging in medicine and surgery* 10.6 (2020): 1223.

[66]Mahapatra, Dwarikanath, et al. "Deformable medical image registration using generative adversarial networks." *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018.

[67]Yi, Xin, Ekta Walia, and Paul Babyn. "Generative adversarial network in medical imaging: A review." *Medical image analysis* 58 (2019): 101552.

[68]Ilse, Maximilian, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning." *International conference on machine learning*. PMLR, 2018.

[69]Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[70]Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.

Chapter 2

Are All Shortcuts in Encoder-Decoder Networks Beneficial for CT Denoising?

Junhua Chen, Chong Zhang, Leonard Wee, Andre Dekker, Inigo Bermejo

Adapted from

Chen, Junhua, et al. "Are all shortcuts in encoder–decoder networks beneficial for CT denoising?." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization (2022): 1-8.

DOI: https://doi.org/10.1080/21681163.2022.2044908

Abstract

Purpose: Denoising of Computed Tomography (CT) scans has attracted the attention of many researchers in the medical image analysis domain during the last decades. Encoder-decoder networks are deep learning neural networks that have become common for image denoising in recent years. Shortcuts (or skip connections) between the encoder and decoder layers are crucial for some image-to-image translation tasks, such as segmentation, style transfer, etc. However, are all shortcuts still necessary for CT denoising?

Method: To answer this question, we set up two encoder-decoder networks representing two popular architectures, then progressively removed shortcuts from the networks from shallow to deep (forwards-removal) and from deep to shallow (backwards-removal). We used two unrelated datasets with different noise levels to test the denoising performance of these networks using two metrics, namely root mean square error (RMSE) and content loss.

Results and Conclusions: The results show that while more than half of the shortcuts are still indispensable for CT scan denoising, removing certain shortcuts leads to performance improvement for denoising. Both shallow and deep shortcuts might be removed thus retaining sparse connections, especially when the noise level is high. Backwards removal seems to have a better performance than forward removal, which means deep shortcuts have priority to be removed. Finally, we propose a hypothesis to explain this phenomenon and validate it in the experiments.

Keywords: Deep Learning; Encoder-decoder Network; Medical Image Denoising; Shortcuts; Comparative Analysis

Introduction

Computed Tomography (CT) denoising is a topic which has been much-investigated in the medical image analysis domain during the last decades. Current denoising methods can be divided into two broad types – filter-based denoising methods and deep learning based denoising methods [1]. Filter methods apply a pre-defined digital image filter to noisy images or try to infer the distribution of noise from data. These methods do not achieve a good performance in a complex noise situation, such as disparate types of noise that manifest as different density distributions introduced by either image acquisition, reconstruction, artefact reduction and image post-processing [2][3]. Deep learning based denoising methods have become popular in recent years due to their potential to autonomously define the optimal parameters in denoising models. In 2016, an international low dose CT denoising contest was held by the American Association of Physicists in Medicine (AAPM), returning this problem to the spotlight in the deep learning era [4]. In this contest, paired full dose and one-quarter dose CT images, including their sinograms, from 10 patients were provided to train a low dose CT denoising network. Kim et al. won the contest by introducing a spatially encoded non-local penalty to the cost function optimized by a neural network [5]. Since then, the associated dataset has been re-used for research on CT denoising contributing to multiple improvements in state-of-art denoising performance [6][7][8][9]. Most of these models have been either generative models, such as generative adversarial networks (GANs) [10], or encoder-decoder networks (EDNs).

As a typical generative model, EDNs and their different variants have been widely used to denoise not only medical images such as CT, MRI,

ultrasound [8][9] [11][12] but also natural images [13][14][15], low light raw images [16], speech [17], desert seismic signals [18], RGB-D images [19], etc. Moreover, EDNs can be used as the generator in GANs which can then used for image denoising [12][20][21].

A typical EDN consists of two parts: an encoder side that extracts features from the input image, based on convolution kernels. The features in the shallow layers become increasingly abstract when progressing towards the deep layers. The decoder side converts deep features to shallow features using deconvolution kernels. The two networks have symmetrical structures that are linked through a bottleneck layer.

Shortcuts, also called ‘ skip connections’ in the literature, connect the encoder and the decoder at different layers and are assumed to be an indispensable part of such networks. The shortcuts in the network can help communicate the information of semantic feature maps (which is lost during the convolution, deconvolution and pooling) from encoder to the decoder [22]. The purpose of introducing shortcuts is to improve the acutance of the output image, make it clearer and retain more “ low level” features or high frequency content, i.e. pixel values that are rapidly changing in space. These shortcuts have been shown to be crucial in certain image-to-image translation tasks, such as segmentation [23][24] and style transfer [25].

However, in noisy images, the information delivered by shortcuts includes not only the content information such as shape boundaries, lines and corners, but also the noise. Part of the delivered features from the encoder to decoder by the shortcuts might thus have a negative impact on denoising.

In this study, we tried to answer the question: are the features from the input image, made available through shortcuts, actually helpful for the explicit task of CT image denoising? Similar question was proposed by other study [26]. Secondly, we consider whether all shortcuts remain equally indispensable in EDNs when performing CT image denoising. (Source code and supplementary materials of this article are available online at https://gitlab.com/UM-CDS/low-dose-ct-denoising/tree/Necessity-of-Shortcuts).

Methods

Data

We re-used (with permission and legal agreement) the images from Low Dose CT Grand Challenge (LDGC) [1] as our first test dataset. This consisted of 10 pairs of clinical abdominal CT examinations, each of which consisted of one full dose CT image (120 kVp and nominally 200 mAs) and one low dose CT image (120 kVp and nominal 50 mAs). We used only the 1 mm slice thicknesses out of which we extracted 5600 single-image frames.

In order to make our conclusions more generalizable, we re-used another dataset for our experiments, namely the NSCLC-Radiomics (LUNG 1) available at The Cancer Imaging Archive (TCIA) [27] under a Creative Commons not-for-commercial use license (CC-BY-NC 3.0). This collection contains clinical patient data and CT images from 422 non-small cell lung cancer (NSCLC) radiotherapy patients and has been cited in several studies [28]. The exposure in these CTs ranged from 42 to 400 mAs. We used 50 CT series (5260 frames in total) that were acquired with 400 mAs as the additional dataset.

Since LUNG1 has no paired full-dose/low-dose images, we created noisy images based on the method proposed by McCollough et al. [4] and Chen et al. [9]. In these methods, the authors simulated the CT scanners' physical behavior: they introduced noise into an acquisition sinogram and reconstructed the CT image from the sinogram to obtain simulated noisy images. Based on this approach, we proposed a closely related method to add noise in the original sinogram. The mathematical description of our method is as follows:

$$z_i = (1 + b_i)e^i + r_i, i = 1, \dots, I, b_i \sim N(\mu, \sigma) \text{(Equation 2-1)}$$

where z_i is the measurement along the i-th ray path, r_i is the read-out error and e^i represents the original line integral of attenuation coefficients along the i-th ray path, and b_i is the black scanner factor, which follows a normal distribution. The intensity of noise added to the image can be controlled using parameter b_i. The main challenge is to pick b_i parameters to match the intensity of noise in low dose CT images compared with their full dose counterparts. We estimated the b_i parameters by comparing RMSE between real low dose CT images and simulated noisy images with different b_i values. In our case, we set μ=0, so only one parameter (σ) needed to be estimated. Additionally, the external noise introduced by thr Radon transform and inverse Radon transform was removed from the generated images.

In order to mimic CT images scanned with 200 mAs and 50mAs from those scanned with 400 mAs, we first measured the noise intensity introduced in images with lower doses by scanning a Gammex 467 CT phantom (Middleton, WI, USA) using a Philips Brilliance Big Bore CT with different

doses (400 mAs, 200mAs, 50 mAs) [29]. We estimated that a σ of 0.0008 and 0.0035 mimicked best the noise in 200 mAs and 50 mAs CT images from 400 mAs image respectively.

In order to assess the necessity of shortcuts with noise of different intensities, we added stronger noise (2.5 times the magnitude of LDGC) to full dose CT (200 mAs), setting σ to 0.0062 to mimic CT images with high noise.

In summary, we used four datasets for our experiments: the original LDGC, LDGC with high noise, LUNG 1 with light noise, and LUNG1 with high noise.

Experiments

To test the necessity of shortcuts, we set up two EDNs with the architecture shown in Figure 2-1 to run our experiments. The first architecture is based on that of the Residual Encoder-Decoder Convolutional Neural Network (RED-CNN) [9], a well-established EDN for CT denoising. It is a 5-layer network, with 3×3 sized convolution and deconvolution kernels that uses padding to keep the image size after convolution or deconvolution layers. Max pooling layers have filters of size 2×2 and a stride of 2. The activation function used is leaky rectified linear unit (LReLU) [30]. The second architecture takes the U-Net as its backbone [22]. The U-Net is a convolutional network architecture that has been used in many image-to-image translation tasks, especially segmentation. The architecture is similar to the RED-CNN (as shown in figure 1 (b)), the main difference being the number of kernels in each convolutional layer.

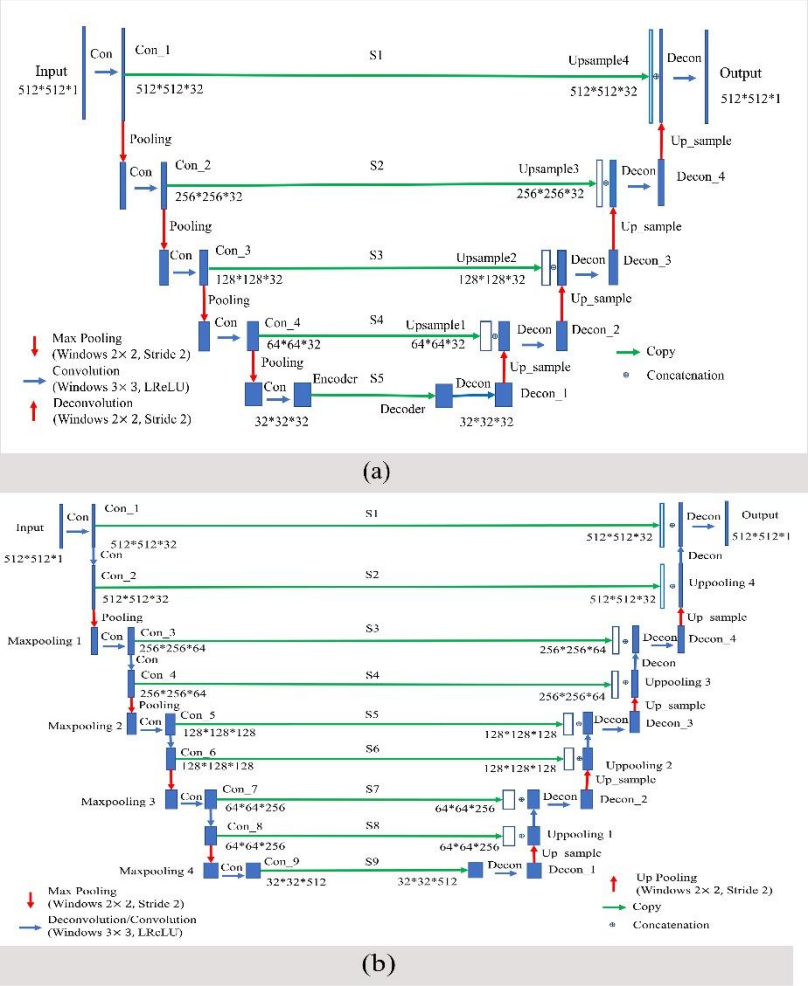


Figure 2-1. Architectures of the encoder-decoder networks used in the experiments.

There are five shortcuts in RED-CNN. All shortcuts except the deepest shortcut (S5) can be removed to test their necessity for medical image denoising. Similarly, in the U-Net all but the deepest shortcut (S9) could be removed. To investigate the impact of shallow and deep shortcuts for denoising, shortcuts were removed from two directions in our experiments.

First, shortcuts were removed from shallow to deep – from S1 to S4 in RED-CNN, from S1 to S8 in U-Net – referred to as forward removing. Second, we removed shortcuts from deep to shallow, referred to as backward removing. We refer to networks with all the shortcuts as fully connected networks and to networks where at least one shortcut has been removed as sparsely connected networks. To assess the impact of shortcuts in the denoising of images with different noise levels, we executed comparative experiments with CT images containing noise of different intensities. We use two metrics, namely the root mean square error (RMSE) and content loss [31] to measure the performance of denoising. The main reason to choose these two metrics is that the former measures the statistical differences between images while the latter measures differences in visual aspect. The mathematic description of RMSE based loss function is as follows:

$$\mathcal{L}_m = 1/N \, \|y - \hat{y}\|_2 \qquad \text{(Equation 2-2)}$$

where y represents the original image, $\hat{y}$ represents the denoised image, N represents the number of pixels in the image and $\|.\|_2$ represents the L2-norm of feature maps. The content loss based metric stems from the VGG-16 architecture [32]. The mathematical description of the content loss-based cost function is as follows:

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{n} \left\| \mathcal{F}_i\left(V\left(O, l_{conv2\_1}\right)\right) - \mathcal{F}_i\left(V\left(D, l_{conv2\_1}\right)\right) \right\|_2$$

(Equation 2-3)

where $\mathcal{F}_i\left(V\left(O, l_{conv2\_1}\right)\right)$ represents the ith feature map of VGG-16 at convolution layer $conv2\_1$ for original image and $\mathcal{F}_i\left(V\left(D, l_{conv2\_1}\right)\right)$

represents the ith feature map for the denoised image, and n represents the numbers of kernels in one convolution layer ($n = 32$ in our case).

In summary, we ran 32 groups of experiments in this study. All our experiments were executed in an Amazon EC2 G3 Graphics Accelerated Instance with a Tesla M60 GPU, 30.5GB of memory and 4 CPUs.

For the LDGC datasets, nine images were used to train the network and one to test it, following the 10-fold cross validation technique. For LUNG 1 based datasets, 40 patients' images were used to train the network and 10 patients' images to test it, using 5-fold cross validation. For each experiment, we conducted 100 tests, each consisting of sampling 4 test slide images from the test set and feeding them to the network.

For training, network weights were initialized following Xavier initialization [33] and the Adam optimizer [34] was used for training with a batch size of 4 and a learning rate of 10^{-5}. Finally, we trained all networks for 45 epochs. We conducted 100 tests, each consisting of sampling 4 test slide images from the test set and feeding them to the network.

In order to interpret the results, we analyzed the output of the different shortcuts in terms of the content and noise transferred by each from the encoder to the decoder. Given that disentangling content from noise is not straightforward from denoised images, we used an alternative route. We considered that the transferred content is best approximated by the output of each shortcut when feeding a full dose CT images (i.e. without noise) as input to the network. We then added noise to the full dose CT image and fed it as input to the model to denoise it. By subtracting the transferred content to the output of each shortcut when denoising a noisy image (which contains

both transferred content and noise), we get the transferred noise. The process is illustrated in Figure 2-2. In order to assess how similar the content (and noise) transferred by each shortcut is to other shortcuts' , we first used a perceptual hash algorithm [35] to summarize the tensor into a fingerprint (an elementwise comparison is not possible due to the output of different shortcuts having different size) and then measured the Hamming distance between them. A higher Hamming distance implies lower similarity and vice versa. We calculated the pairwise similarity of the output of the eight shortcuts (excluding the deepest shortcut) of the U-Net architecture in 400 full dose CT images from LUNG 1.

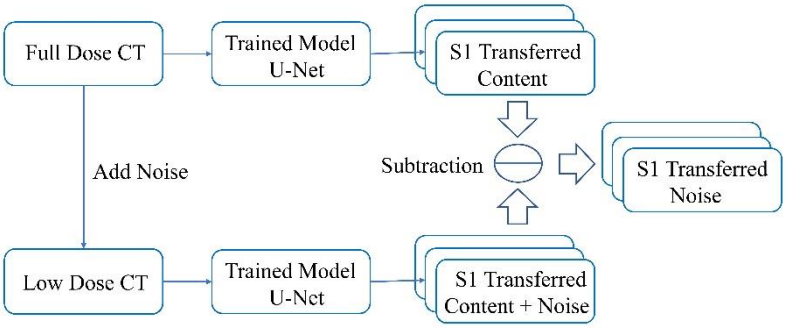


Figure 2-2. Demonstration of extracting transferred noise though shortcuts.

Results

An example result for forward removal of shortcuts from the LDCG dataset by using U-Net is shown in Figure 2-3.(additional examples about backward removing and results from RED-CNN are included in Supplementary Figures 1-3), where the image denoised by the EDNs with all the shortcuts is visually similar to the best performance from sparsely connected networks. Therefore, we analysed the performance based on the two metrics mentioned in Section II. The results summarised in Table 2-1 show that the best average

performance was achieved by sparsely connected networks in 28 groups of experiments, while the fully connected network achieved the best average performance in the remaining 4 groups of experiments. Boxplots with the losses measured across all tests in LUNG 1 based on the two metrics using the two removing strategies are shown in Figure 2-4 (corresponding boxplots for LGDC can be found in Supplementary Figure 4). As shown in the plots, the denoising performance of sparsely connected networks is worse than that of fully connected networks when more than 4 or more shortcuts were removed in U-Net and 2 or more shortcuts were removed in RED-CNN. However, the differences between the best performance of sparsely connected networks and fully connected networks cannot be judged based on visual examination. We tested whether the differences between the performances of the fully connected network and the best performance from sparsely connected networks are statistically significant using the Wilcoxon signed-rank test. The results of the experiments are shown in Table 2-1 (data used in statistical analyses is available at the source code repository). As shown in the table, there are significant differences in 17 out of 32 groups of experiments, 12 out of 16 groups in highly noisy image experiments. The best performing sparsely connected networks outperformed fully connected networks (both by forward removal and backward removal), especially for highly noisy images. In most cases, sparsely connected networks outperform the fully connected network when one shortcut is removed from the RED-CNN architecture and two shortcuts were removed from the U-Net architecture. On the other hand, the backward removing strategy resulted in better performance than forward removing. In our experiments, the U-Net based architecture outperformed the RED-CNN based architecture.

The analysis of the similarity between the output of different shortcuts in the U-Net resulted in a mean Hamming distance of 22.4 for transferred content 26.1 for transferred noise. These results show that delivered content features

have higher similarity than delivered noise features. We provide a discussion for this phenomenon in Section IV. Figure 2-5 shows an example of transferred content information and transferred noise through shortcuts S1 and S2.

Discussion

In this study, we have assessed the impact of shortcuts in two types of EDNs for CT scan denoising. These shortcuts have previously been shown to be beneficial in some fully convolutional networks and some image-to-image translation tasks, but their impact on denoising had not yet been studied in detail. The results of our experiments show that removing certain shortcuts from EDNs either has no detrimental impact or improves the performance of CT denoising. These results support our initial intuition that in noisy images, shortcuts not only transfer content information (semantic features) but also noise from encoder to decoder.

Table 2-1. Comparison of the performance between the full network and the best performance from sparse connected networks

Architecture	Dataset	Metric	Noise	Full network[1]	Forward Moving		Backward Moving	
					Sparse connection[1]	*p*-value[2]	Sparse connection	*p*-value
RED-CNN Based	LDGC	CL[3]	Light	0.1076	0.1059(1)[4]	0.44	0.1026(1)[5]	0.01
	LDGC	CL	High	0.2634	0.2515(1)	<0.01	0.2305(1)	<0.01
	LDGC	RMSE	Light	0.0220	0.0221(1)	0.87	0.0219(1)	0.90
	LDGC	RMSE	High	0.0325	0.0316(1)	<0.01	0.0308(1)	<0.01
	LUNG1	CL	Light	0.1479	0.1451(1)	0.19	0.1447(2)	0.11
	LUNG1	CL	High	0.2139	0.2170(1)	0.29	0.1878(2)	<0.01
	LUNG1	RMSE	Light	0.0287	0.0287(1)	0.80	0.0287(2)	0.08
	LUNG1	RMSE	High	0.0322	0.0324(1)	0.57	0.0308(1)	<0.01
U-Net Based	LDGC	CL	Light	0.0887	0.0880(3)	0.98	0.0882(1)	0.89
	LDGC	CL	High	0.1901	0.1885(1)	0.25	0.1856(2)	0.03
	LDGC	RMSE	Light	0.0202	0.0202(4)	0.96	0.0202(1)	0.10
	LDGC	RMSE	High	0.0281	0.0282(1)	0.75	0.0278(3)	<0.01
	LUNG1	CL	Light	0.1075	0.1005(1)	<0.01	0.0981(3)	<0.01
	LUNG1	CL	High	0.1525	0.1458(1)	<0.01	0.1417(2)	<0.01

LUNG1	RMSE	Light	0.0267	0.0259(2)	<0.01	0.0257(1)	<0.01
LUNG1	RMSE	High	0.0293	0.0287(1)	<0.01	0.0284(3)	<0.01

[1]mean loss values; [2] *p*-value of Wilcoxon signed-rank test; [3]mean content loss; [4]mean the number of moved shortcuts when sparse connection network received the best performance; [5] lower value of loss have a better performance, in this case, 0.1026 better than 0.1059.

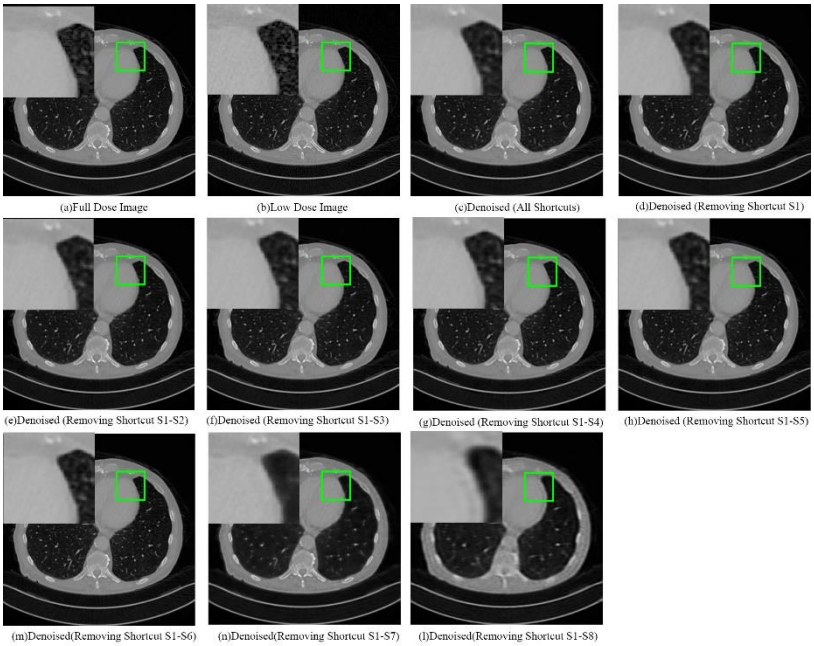


Figure 2-3. Demonstration of the result of denoising an image from LDGC. (a) Original full dose CT image and zoomed of Region of Interest; (b) Paired low dose CT image and zoomed of Region of Interest; (c) Denoised result when all shortcuts are connected in the network and zoomed of Region of Interest; (d-l) Denoised image when shortcut S1-S8 is removed from the network and zoomed of Region of Interest.

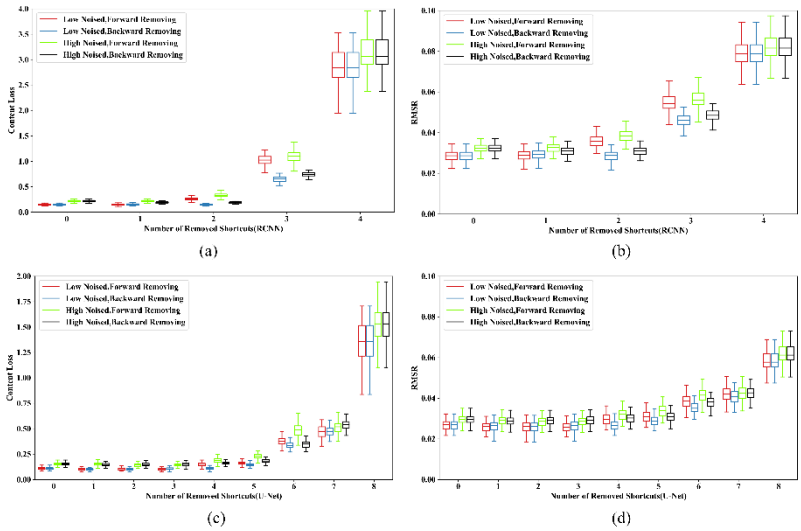


Figure 2-4. Boxplots of (a) content loss in RED-CNN denoised images vs original images along with forward and backward moving in the LUNG1 dataset; (b) RMSE in RED-CNN denoised images vs original images along with forward and backward moving in the LUNG1 dataset; (c) content loss in U-Net denoised images vs original images along with forward and backward moving in the LUNG1 dataset; (d) RMSE in U-Net denoised images vs original images along with forward and backward moving in the LUNG1 dataset.

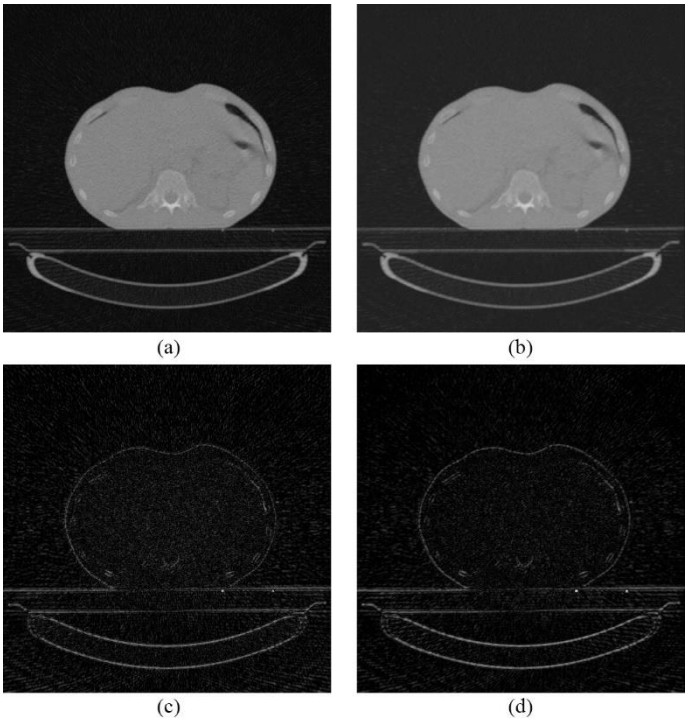


Figure 2-5. A demonstration of transferred information to encoder (based on U-Net) though shortcut S1, S2 and corresponding noise. (a) An example of transferred information though S1; (b) an example of transferred information though S2; (c) corresponding transferred noise for image (a); corresponding transferred noise for image (b).

The results of forward and backward removing show that either shallow or deep shortcuts can be removed while maintaining or improving denoising performance. Backward removal (deep to shallow) outperformed backward removing in our experiments. One potential justification is that the information delivered through deep shortcuts contains less “low level” features compared with shallow shortcuts while most of the content information for reconstruction is carried by the deepest shortcut (which is

never removed). Therefore, the content benefit of deep shortcuts is lower than shallow shortcuts and deep shortcuts should be first.

On the other hand, part of the image content information transferred by shortcuts seems to be duplicated by different shortcuts. The results of our experiments show that the transferred content across different shortcuts have higher similarity than the noise transferred. This phenomenon might be explained by the fact that characteristics of noise might make its pattern harder to learn, and consequently the noise delivered by different shortcuts might be more independent. Therefore, having too many shortcuts might be counter-productive, by reducing the content benefit per shortcut relative to the negative effect of transferred noise.

We have shown that our results are consistent in two otherwise unrelated lung datasets and we believe these results can be broadly generalizable to other CT scans. We used two popular EDN architectures in our experiments and we think it is likely that the results would be consistent with other EDN architectures. We used shallow networks (less than 10 layers) with full shortcuts rather than a deeper network (such as the DenseNet-based encoder-decoder network proposed in [36]) as the experimental network because denoising focuses on “low level” features and does not require a deep network to extract features. The noise introduced in the images used in our experiments to simulate the noise in low dose CT was based on the method described in McCollough et al. [4] and Chen et al. [9]. The validity of our findings depends to a certain extent on the representativeness of this synthetic noise of noise in low dose CTs. For future research designing encoder-decoder networks for medical image denoising, we propose that a sparse connection between encoder and decoder should be considered. Our

results indicate that removal of deep shortcuts should have the higher priority when constructing sparse networks. We have no results supporting a universal solution, so researchers should adjust their networks based on experiments with their own data.

These insights will help the development of denoising models that have a wide range of applications in medical imaging, for example improving the performance of automatic segmentation algorithms [37], or improving the repeatability of radiomics features [38].

Conclusions

In this article, we discuss the necessity and desirability of shortcuts in encoder-decoder networks for denoising of CT scans. In order to answer this question, we adopt two encoder-decoder networks as test architectures, we ran several sets of experiments based on two EDNs using different datasets, different noise intensity levels, different metrics and different moving strategies. The results show that over half of shortcuts are indispensable for denoising. However, a sparse connection will provide a positive effect for denoising and both shallow and deep shortcuts can be removed to achieve a more sparse connection, especially when the noise is high. Backward removal seems have a better performance than forward removal which means deep shortcuts should be kept when making shortcuts sparse. Finally, we believe this conclusion may be suitable not only for medical image denoising but also for other image denoising tasks.

Appendix

Support materials of this Chapter can be found in this link.

References:

[1]Diwakar, Manoj, and Manoj Kumar. "A review on CT image noise and its denoising." *Biomedical Signal Processing and Control* 42 (2018): 73-88.

[2]Goldman, Lee W. "Principles of CT: radiation dose and image quality." *Journal of nuclear medicine technology* 35.4 (2007): 213-225.

[3]Diwakar, Manoj, and Manoj Kumar. "A review on CT image noise and its denoising." *Biomedical Signal Processing and Control* 42 (2018): 73-88.

[4]McCollough, Cynthia H., et al. "Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 Low Dose CT Grand Challenge." *Medical physics* 44.10 (2017): e339-e352.

[5]Kim, Kyungsang, Georges El Fakhri, and Quanzheng Li. "Low-dose CT reconstruction using spatially encoded nonlocal penalty." *Medical physics* 44.10 (2017): e376-e390.

[6]Yang, Qingsong, et al. "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss." *IEEE transactions on medical imaging* 37.6 (2018): 1348-1357.

[7]Kang, Eunhee, et al. "Deep convolutional framelet denosing for low-dose CT via wavelet residual network." *IEEE transactions on medical imaging* 37.6 (2018): 1358-1369.

[8]Shan, Hongming, et al. "3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network." *IEEE transactions on medical imaging* 37.6 (2018): 1522-1534.

[9]Chen, Hu, et al. "Low-dose CT with a residual encoder-decoder convolutional neural network." *IEEE transactions on medical imaging* 36.12 (2017): 2524-2535.

[10]Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

[11]Ravichandran, Vignesh, et al. "Deep Network for Capacitive ECG

Denoising." *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA).* IEEE, 2019.

[12]Ran, Maosong, et al. "Denoising of 3D magnetic resonance images using a residual encoder–decoder Wasserstein generative adversarial network." *Medical image analysis* 55 (2019): 165-180.

[13]Mao, Xiaojiao, Chunhua Shen, and Yu-Bin Yang. "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections." *Advances in neural information processing systems.* 2016.

[14]Yang, Xin, et al. "DEMC: A Deep Dual-Encoder Network for Denoising Monte Carlo Rendering." *arXiv preprint arXiv:1905.03908* (2019).

[15]Chaitanya, Chakravarty R. Alla, et al. "Interactive reconstruction of Monte Carlo image sequences using a recurrent denoising autoencoder." *ACM Transactions on Graphics (TOG)* 36.4 (2017): 98.

[16]Guan, Hao, et al. "NODE: Extreme Low Light Raw Image Denoising using a Noise Decomposition Network." *arXiv preprint arXiv:1909.05249* (2019).

[17]Park, Se Rim, and Jinwon Lee. "A fully convolutional neural network for speech enhancement." *arXiv preprint arXiv:1609.07132* (2016).

[18]Ma, Haitao, et al. "Deep Residual Encoder-Decoder Networks for Desert Seismic Noise Suppression." *IEEE Geoscience and Remote Sensing Letters* (2019).

[19]Sterzentsenko, Vladimiros, et al. "Self-Supervised Deep Depth Denoising." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

[20]Yi, Xin, and Paul Babyn. "Sharpness-aware low-dose CT denoising using conditional generative adversarial network." *Journal of digital imaging* 31.5 (2018): 655-669.

[21]Chen, Shengjie, et al. "Low-resolution palmprint image denoising by generative adversarial networks." *Neurocomputing* 358 (2019): 275-284.

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

[23]Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[24]Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495.

[25]Li, Yijun, et al. "Universal style transfer via feature transforms." *Advances in neural information processing systems*. 2017.

[26]Naseer, Muzammal, et al. "A self-supervised approach for adversarial robustness." Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[27]Aerts, H. J. W. L., Wee, L., Rios Velazquez, E., Leijenaar, R. T. H., Parmar, C., Grossmann, P., … Lambin, P. (2019). *Data From NSCLC-Radiomics* [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI

[28]Aerts, Hugo JWL, et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." *Nature communications* 5 (2014): 4006.

[29]Zhovannik, Ivan, et al. "Learning from scanners: Bias reduction and feature correction in radiomics." *Clinical and translational radiation oncology* 19 (2019): 33-38.

[30]Xu, Bing, et al. "Empirical evaluation of rectified activations in convolutional network." *arXiv preprint arXiv:1505.00853* (2015).

[31]Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." *European conference on computer vision*. Springer, Cham, 2016.

[32]Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[33]Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010.

[34]Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[35]Niu, X. M., & Jiao, Y. H. (2008). An overview of perceptual hashing. *Acta Electronica Sinica*, *36*(7), 1405-1411.

[36]Jégou, Simon, et al. "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017.

[37]Uchikoshi, Kousei, Masaya Hasegawa, and Shigeki Hirobayashi. "Denoising of low dose CT images using mask non-harmonic analysis with edge-preservation segmentation and whitening filter." *Multimodal Biomedical Imaging XIV*. Vol. 10871. International Society for Optics and Photonics, 2019.

[38]Traverso, Alberto, et al. "Repeatability and reproducibility of radiomic features: a systematic review." *International Journal of Radiation Oncology* Biology* Physics* 102.4 (2018): 1143-1158.

Chapter 3
Generative Models Improve Radiomics Reproducibility in Low Dose CTs: A Simulation Study

Junhua Chen, Chong Zhang, Alberto Traverso, Ivan Zhovannik, Andre Dekker, Leonard Wee and Inigo Bermejo

Adapted from

Chen, Junhua, et al. "Generative models improve radiomics reproducibility in low dose CTs: a simulation study." Physics in Medicine & Biology 66.16 (2021): 165002.

DOI: https://doi.org/10.1088/1361-6560/ac16c0

Abstract

Purpose: Radiomics is an active area of research in medical image analysis, however poor reproducibility of radiomics has hampered its application in clinical practice. This issue is especially prominent when radiomic features are calculated from noisy images, such as low dose computed tomography (CT) scans. In this article, we investigate the possibility of improving the reproducibility of radiomic features calculated on noisy CTs by using generative models for denoising.

Method: Our work concerns two types of generative models – encoder-decoder network (EDN) and conditional generative adversarial network (CGAN). We then compared their performance against a more traditional "non-local means" denoising algorithm. We added noise to sinograms of full dose CTs to mimic low dose CTs with two levels of noise: low-noise CT and high-noise CT. Models were trained on high-noise CTs and used to denoise low-noise CTs without re-training. We tested the performance of our model in real data, using a dataset of same-day repeated low dose CTs in order to assess the reproducibility of radiomic features in denoised images.

Results: EDN and the CGAN achieved similar improvements on the concordance correlation coefficients (CCC) of radiomic features for low-noise images from 0.87 [95%CI, (0.833,0.901)] to 0.92[95%CI, (0.909,0.935)] and for high-noise images from 0.68 [95%CI, (0.617,0.745)] to 0.92[95%CI, (0.909,0.936)], respectively. The EDN and the CGAN improved the test-retest reliability of radiomic features (mean CCC increased from 0.89 [95%CI, (0.881,0.914)] to 0.94[95%CI, (0.927,0.951)]) based on real low dose CTs.

Conclusions: These results show that denoising using EDN and CGANs could be used to improve the reproducibility of radiomic features calculated from noisy CTs. Moreover, images at different noise levels can be denoised to improve the reproducibility using the above models without need for re-training, provided the noise intensity is not excessively greater that of the high-noise CTs. To the authors' knowledge, this is the first effort to improve the reproducibility of radiomic features calculated on low dose CT scans by applying generative models.

Keywords: Radiomics; Reproducibility; Computed Tomography; Denoising; Generative Models

Introduction

Radiomics is currently an active area of research in medical image analysis. It involves the automated extraction of (either hand-crafted or deep-learning) quantitative image metrics known as “features”, in the hope of improving the diagnostic, prognostic, or predictive accuracy of clinical models [1].One of the major advantages of hand-crafted radiomics features, as opposed to deep features, is higher potential for interpretability by human operators.

Radiomics has shown potential for clinical-decision support in oncology for a diverse range of cancer sites such as lung [2], head and neck [3], and rectal cancer [4], among others. The most widely used clinical imaging modalities for radiomics are computed tomography (CT) [3][5], magnetic resonance imaging (MRI) [6], and positron emission tomography (PET) [7]. Radiomics has attracted increased attention from researchers following the seminal article by Aerts *et al.* [7] in 2014. In spite of significant progress made during recent years, there remain barriers that hamper widespread adoption of radiomics in clinical settings. One important issue is the generally poor repeatability and reproducibility of radiomic features [8]. Repeatability refers to features that remain the same when extracted multiple times in the same subject. The repeatability of radiomic features may be assessed by test-retest imaging [9] and interobserver studies of tumor contouring [10]. Radiomics reproducibility is the stability of features when at least one processing condition (e.g. equipment, software, acquisition settings) has been changed. The reproducibility of radiomics features is key to external validity and widespread generalizability with respect to differences in image reconstruction [11], radiation dose during CT scanning [12], and other variations that inevitably arise across clinics and scanners.

An important source of non-reproducibility or limited generalizability in radiomics that demands more attention is image noise. Due to the long term risk posed by low levels of ionizing radiation exposure, low dose CTs have become more popular (e.g. the ALARA principle [17]) especially for screening and monitoring of populations at risk. Therefore, researchers have interest to use radiomic features based on low dose CTs. As an inevitable trade-off for low radiation exposure, higher noise is present in these images. This noise decreases the image texture [13]. As reported in [11], changes in radiation dose reduce the reproducibility of radiomic features, and features from low dose CT images tend to have lower reproducibility [12]. Image noise has been shown to adversely impact the reproducibility of radiomic features if signal to noise ratio (SNR) falls below 50 dB [5], but the results also show that some radiomics features are robust to low-pass filtering. Thus, improving the reproducibility of radiomics from low dose CTs is a timely and potentially impactful clinical research topic.

To the best of the authors' knowledge, there are presently no published detailed analysis on how to improve radiomics features robustness in low dose CT. A potential solution worth exploring is pre-extraction denoising of images [12]. Jiang, *et al.* [14] described a new semi-supervised generative adversarial network (GAN) that reconstructed higher-resolution CT images from their low-resolution counterparts with state-of-the-art performance. However, they did not investigate the possibility of using similar generative models to improve radiomics reproducibility in low dose CT.

In the topic of medical image denoising, many effective alternative procedures have been proposed to improve low dose image quality and recover texture information, including building a more sophisticated

imaging platform [15], denoising in the CT sinogram domain [16], and denoising in the reconstructed CT image domain. The most convenient and popular denoising methods for low dose CT images operate in the CT image domain because hardware details and sinograms of CTs are hard to access for most researchers [23]. Generally speaking, since higher image quality should lead to higher reproducibility of radiomics, denoising seems to be a useful pre-processing step to consider.

Many articles describing image denoising techniques have been published so far, as shown in reviews [18][19]. They can be divided into two main categories: traditional denoising methods and deep learning-based denoising. The latter views denoising as a type of restricted image-to-image conversion task. Traditional denoising methods [20] have known limitations such as loss of detail in images. With the advent of deep learning, a series of publications have shown that generative models outperform traditional methods in low dose CT denoising [21][22-24]. The most widely-tested generative models are autoencoders (AE) [25][26], encoder-decoder networks (EDN) [21][22], fully convolutional neural networks (FCN) [24][27] and various GANs [23][28].

An AE is an unsupervised neural network that learns how to efficiently compress (encode) an image by learning how to reconstruct the image from a tightly compressed (encoded) representation. EDNs are thus a convolutional version of AEs that allow connections across encoder and decoder layers. FCNs replace the fully connected layer in traditional convolutional neural networks with a deconvolution layer to perform the image-to-image translation.

GANs were originally proposed by Goodfellow *et al.* [28], and have been applied to diverse image-to-image translation tasks [29][30]. One of the major disadvantages of original GANs was the difficulty to control the output. However, Mirza *et al.* [31] proposed conditional GANs (CGANs), which introduced conditional restrictions into GANs to make the output more controllable and training more stable. Yang *et al.* [23] then achieved state-of-art performance in low dose CT denoising using a CGAN with Wasserstein distance penalty and perceptual loss. The general limitation of deep learning based methods remains high computational resource requirements during training and the need for large datasets.

In this article, we focus on using generative models to improve the image quality of low dose CT images and assess the impact it has on the reproducibility of radiomic features. Source code of the whole project with detailed instructions, pre-trained models and supplementary materials are available (https://gitlab.com/UM-CDS/low-dose-ct-denoising/-/branches). We hope our codes and related documents could assist future researchers to interpret and re-use our work.

Methods and Materials

Institutional Review Board approval was not applicable for this study, since the primary source of data was an open access collection on The Cancer Imaging Archive (National Institutes of Health) [32] and all patients' private information had been moved from CT scans. This dataset has been used for this study in accordance with the Creative Commons Attribution-NonCommercial 3.0 Unported (CC BY-NC) conditions.

Model Development

Though FCNs have shown good performance in image-to-image translation tasks, we excluded them from our experiments because they are unfeasibly slow, since they generate new images pixel-by-pixel. As convolutional versions of AEs, EDNs are expected to have better performance for image-to-image translation tasks [22]. Therefore, we excluded AEs from our experiments in favour of EDNs. CGANs were included in the experiments due to proven performance for low dose CT denoising work [23].

The architecture of our EDN is shown in Figure 3-1. It is a 5-layer network consisting of 32 (3×3)-sized convolution and deconvolution kernels with padding to keep the image size constant after each convolution or deconvolution. Max-pooling layers are used with 2×2 size filters and a stride of 2. We used cross entropy as the loss function and leaky rectified linear units (LReLUs) as activation functions. An original 512×512×1 CT image was fixed as the constant dimension of input and output images when training and testing.

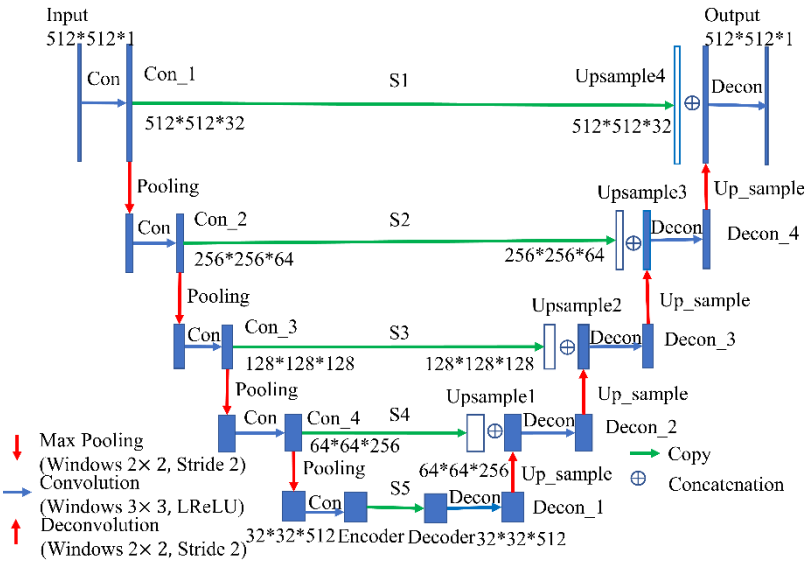


Figure 3-1. The architecture of the encoder-decoder network

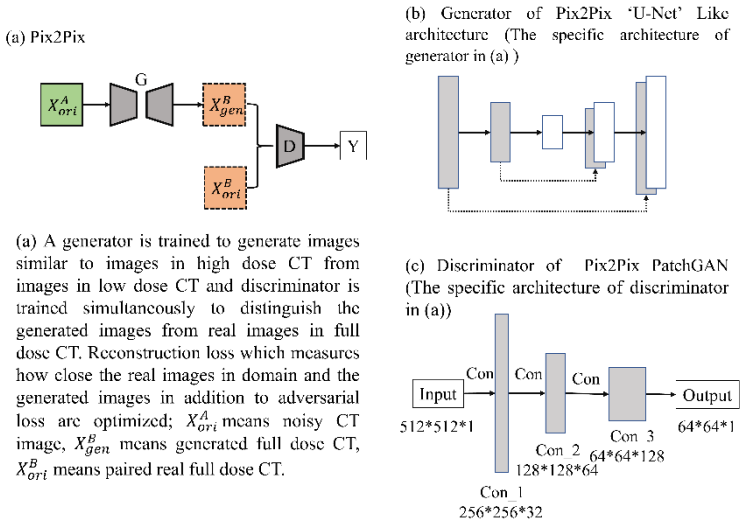


Figure 3-2. The architecture of the CGAN (Pix2Pix)

For our CGAN, we used the same architectures and parameter settings as proposed elsewhere [31][33]. The architecture of the CGAN is illustrated in Figure 3-2. We adjusted the network's input and output dimensions from the original to 512×512×1. Finally, we adjusted the networks to output DICOM files directly for archival and radiomics feature calculation. The loss function in the CGAN was also set to cross entropy.

In order to compare the performance of generative models with that of traditional denoising methods, we included a type of low-pass filter, non-local means algorithm [34], as a good representative for traditional denoising methods. There are a couple of reasons to choose non-local means as our comparison denoising method. First, non-local means had the better performance amongst other traditional denoising methods [35]. Second, non-local means executed faster than other algorithms such as Block-Matching and 3D filtering (BM3D) [35][36] but gave similarly denoising outcome. For our non-local means algorithm, we set the size of the 'search windows' to 5 and the filtering parameter 'h' to 7.

Data Acquisition

We used the high quality NSCLC-Radiomics collection [37] (hereafter called LUNG 1), which contains CT scans of 422 non-small cell lung cancer (NSCLC) patients, as our experimental dataset. These CT scans included annotations drawn by specialist radiation oncologists that delineate a region of interest (ROI), the gross tumor volume. ROIs were necessary to be able to compute radiomic features. The CT images for which the dose level ('parameter exposure' in DICOM metadata) was missing (n=200) were excluded from further analyses. We considered CT images scanned at 400 milliampere-seconds (mAs) and above as full dose CT (n=157, the index of

LUNG 1 patients included in the experiments can be found in Supplementary Table 1, supplementary materials are available: https://gitlab.com/UM-CDS/low-dose-ct-denoising/-/branches). These data were be used for training (n=40, 4260 frames) and testing (n=117, 13423 frames). Conversely, we designated CT images scanned at 50 mAs as low dose CT, taking the same definition as a prior Low Dose CT Grand Challenge [22][38].

As mentioned, training of EDNs and CGANs require paired images, in our case, pairs of matching low dose and full dose CT scans. However, LUNG 1 contains no paired images, thus we simulated the noisy degradation present in low dose CT images by introducing noise using the method proposed in literature [22][38]. In these, the authors had mimicked CT scanners' behavior by adding noise with a normal distribution into a sinogram (by Radon transform) and reconstructed the CT image from the modified sinogram to obtain simulated noisy images. We used a similar method to add noise in the original sinogram as follows:

$$z_i = (1 + b_i)e^i + r_i, i = 1, \dots, I, b_i \sim N(\mu, \sigma) \quad \text{Equation 3-1}$$

where z_i is the measurement along the i-th ray path; r_i is the read-out error; e^i represents the original line integral of attenuation coefficients along the i-th ray path; and, b_i is the black scanner factor, which follows a normal distribution. The intensity of noise added to the image can be controlled through the parameter b_i.

To simulate low dose CT images (scanned with 50mAs) from full dose CT images (scanned with 400 mAs), we first measured the noise intensity introduced in images with lower doses by scanning a Gammex 467 CT

phantom (Middleton, WI, USA) using a Philips Brilliance Big Bore CT at two dose levels (400 mAs and 50 mAs) [29]. The signal-to-noise ratio (SNR) of the real phantom dataset was 19.7 dB (95%CI [17.8, 21.6]). We thus estimated that a σ value of 0.0035 best estimated the noise in 50 mAs CT images when generated from 400 mAs images. The SNR in the simulated low-noise images was 18.3 (95%CI, [16.9, 20.1]) dB, close to the real value. To assess the reproducibility of radiomic features with noise of different intensities, we added stronger noise (25 times noise power) by setting σ to 0.0068 to mimic CT images with stronger noise (referred to as simulated high-noise images hereafter). The SNR in the simulated high-noise images had thus reduced to 6.0 (95%CI, [5.9,6.1]) dB. Additionally, extraneous noise introduced by the Radon transform and inverse Radon transform was filtered from the simulated images. A comparison of noise in simulated images and in real phantom scans is shown in Figure 3-3, the intensity of noise in real phantom is 17.1 dB and average noise power spectra density within whole image is 45.8 W/Hz. The intensity of noise in simulated low-noise images is 19.4 dB and average noise power spectra density within whole image is 3.6 W/Hz, intensity of noise in simulated high-noise images is 6.1 dB and average noise power spectra density within whole image is 6.0 W/Hz.

We used 40 subjects from LUNG 1 (4260 images in total) for training in all three denoising models and then 117 subjects from LUNG 1 for testing. Training was only based on paired low dose (high-noise) and full dose images for all models. Denoising for the low-noise images was performed using the trained models without any additional retraining.

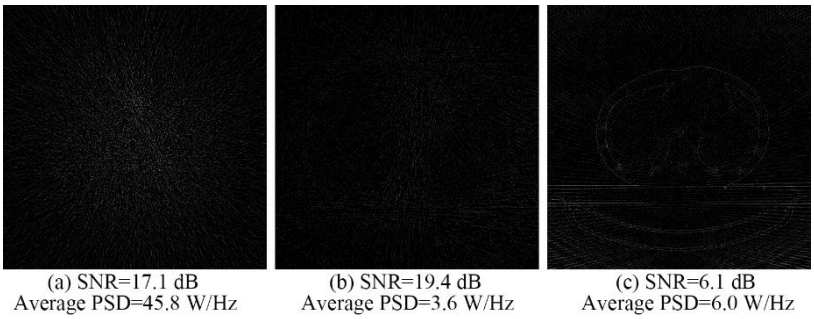


Figure 3-3. A comparison of noise in simulated images and in a real phantom image. (a) real phantom image; (b) simulated low-noise image; (c) simulated high-noise image.

To test the performance of our models in denoising real low dose CTs, we used two additional datasets. First, a collection of phantoms CTs were scanned at different exposure levels, as in [29]. Second, we used the Reference Image Database to Evaluate Therapy Response (RIDER) collection, a collection of same-day repeat CT scans collected to assess the variability of tumor measurements [40]. This dataset comprised paired CT scans for 32 NSCLC patients with corresponding gross tumor volume annotations. These CT images had been scanned at low doses (7 to 13 mAs), making it suitable tests for our denoising experiment.

Calculation of Radiomic Features

Radiomics features from images in DICOM format were extracted using the open source O-RAW extension [41] to PyRadiomics [42]. Radiomic features can be divided into three classes – shape features, intensity histogram (first-order) features and textural (Haralick) features. In our experiments, the ROI masks for calculating radiomic features were not affected by denoising, therefore shape features were excluded from further analysis. Finally, 90

radiomic features (listed in Supplementary Table 2, supplementary materials are available at https://gitlab.com/UM−CDS/low−dose−ct−denoising https://gitlab.com/UM-CDS/low-dose-ct-denoising) were included for our analyses. We followed the recommendations of the Image Biomarker Standardization Initiative (IBSI). The IBSI checklist can be found in Supplementary Table 3.

Experiments

Experiments were executed in a virtual Amazon Elastic Compute instance (G3 Graphics Accelerated Instance with Tesla M60 GPU, 30.5GB of memory and 4 CPUs).

We executed three kinds of comparisons of radiomics feature reproducibility, comparing: (i) denoising of CT images using different types of generative models (EDN, CGAN) and one traditional denoising algorithm (non-local means), (ii) performance for different numbers of training epochs (25, 50, 75 and 100) and (iii) performance for different noise intensities (low and high noise images. We compared feature reproducibility by calculating the correlation of each radiomic feature between each full dose CT image with its a corresponding noise-added then post-denoised CT.

To assess the impact of denoising on real low dose CT scans using the models trained above on LUNG 1, we ran two additional experiments. First, we used the trained models to denoise real low dose CT scans of phantoms scanned at 50 mAs and then compared the difference between low and full dose CT to the difference between denoised and full dose CT. Second, we assessed the impact of denoising on the test-retest reproducibility of radiomic features using the aforementioned RIDER dataset. Finally, we

compared the test-retest reliability of radiomic features in original versus denoised CT scans by calculating the correlation between the CT scan pairs for each radiomic feature.

Statistical Analysis

Correlation was defined as the concordance correlation coefficient (CCC) [43]. We measured the difference between the original full dose CT images and denoised images using Root Mean Square Error (RMSE) and content loss [31]. Content loss was calculated based on a pretrained VGG-16 [32]. The definition of RMSE is shown in

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2} \qquad \text{Equation 3-2}$$

where y_i and $\hat{y}_i$ represent the image value in position i for original full dose CT and denoised CT, respectively. Image intensities y and $\hat{y}$ were normalized to 0-1 before calculating RMSE. M represents the total number of pixels in an image and it is 262144 (512 x 512) in our case.

The flowchart summarizing our study methodology is shown in Figure 3-4.

Results

Training took 8 hours for the EDN model and 20 hours for the CGAN model. The loss function (cross entropy loss) curves of EDN and CGAN as a function of training epochs is shown in Figure 3-5, showing similar convergence. The trained models were used to generate denoised CT scans of the test set. An example of an original, low-noise and post-denoising CT scan is shown in Figure 3-6. A corresponding figure for high-noise level images is given in Supplementary Figure 1.

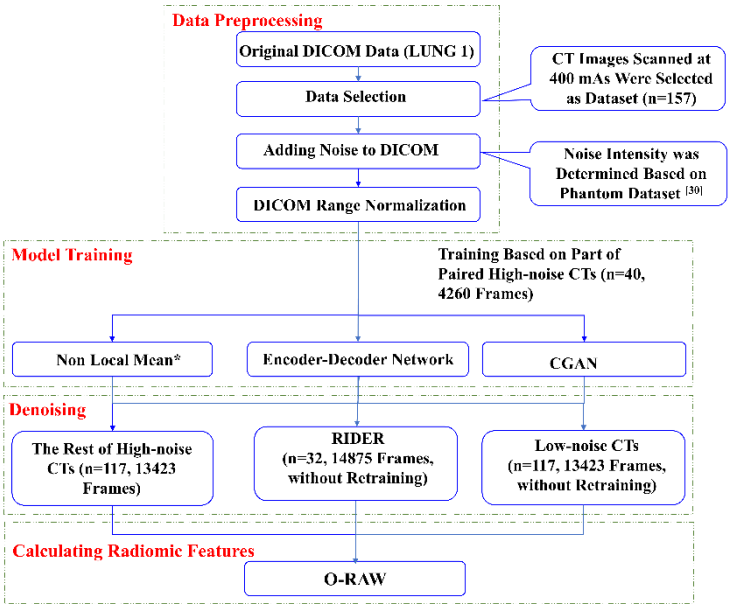


Figure 3-4. Flowchart of methods. *Non-local means algorithm applied only on The Rest of High-noise CTs and Low-noise CTs datasets.

Following the classification in [8], the reproducibility of a feature was deemed good, medium or poor when CCC≤0.85, 0.65≤CCC<0.85 and CCC<0.65, respectively. Table 3-1 shows the RMSE, content loss and ratio of poor-, medium-, and highly-reproducibility radiomic features. We summarized the reproducibility of radiomic features in low-noise images and their denoised counterparts using a heatmap in Figure 3-7 (a corresponding figure for high-noise images is shown in Supplementary Figure 2, the CCCs for each feature in different models can be found in Supplementary Tables 4-5).

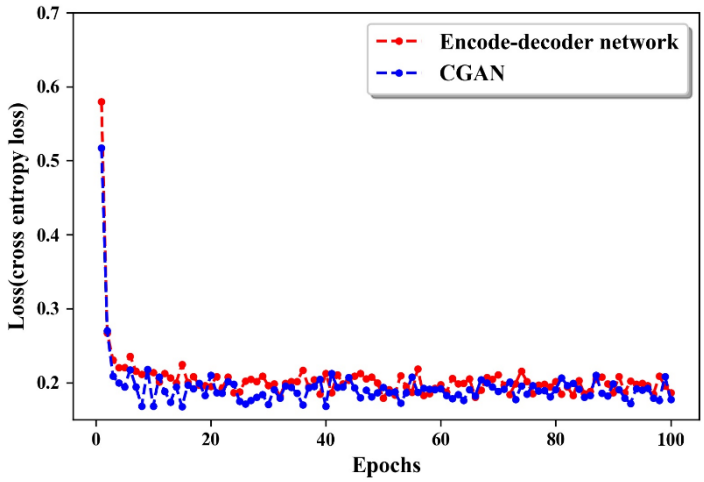


Figure 3-5. Loss function (cross entropy loss) curves of encoder-decoder network and CGAN along with different training epochs

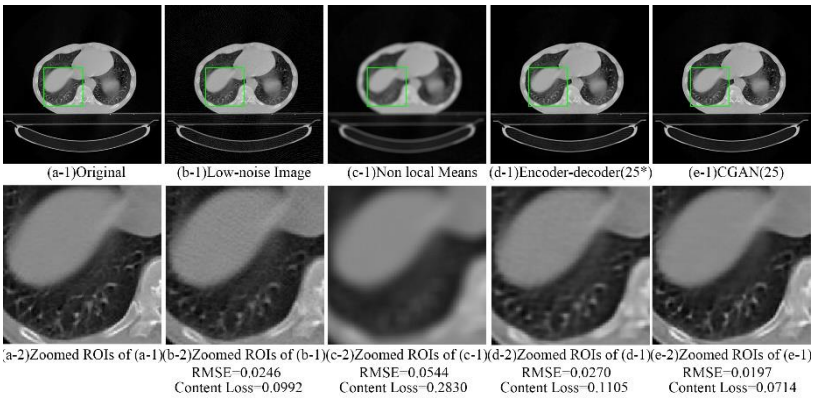


Figure 3-6. Example of low dose CT denoising. (a-1) The original full dose CT image; (b-1) Low-noise image; (c-1) Image denoised using non-local means; (d-1) Image denoised by encoder-decoder network (Training at 25 epochs); (e-1) Image denoised by CGAN; (a-2) to (e-2) Zoomed ROIs for

(a-1) to (e-1). We regard the higher noise in (d-2) by comparing with (b-2) as a coincidence.

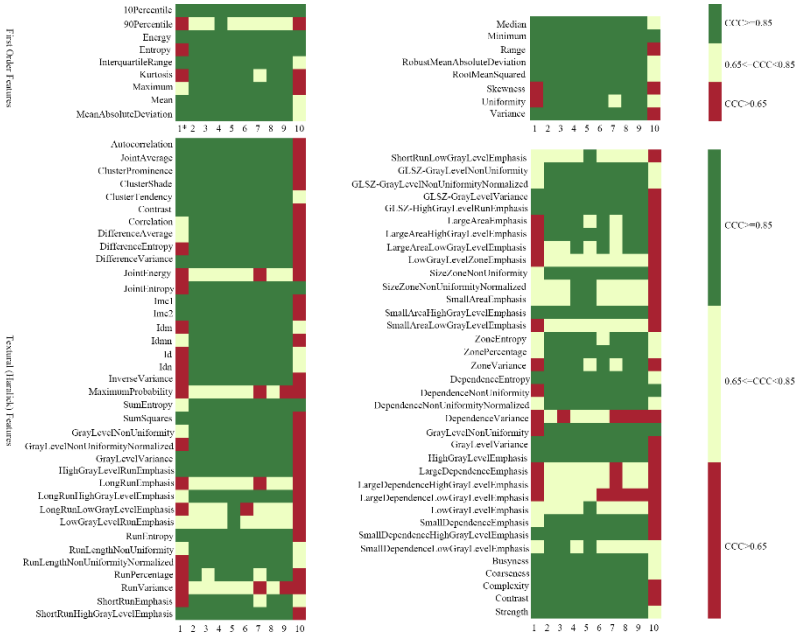


Figure 3-7. Heatmap of radiomic features' reproducibility based on high-noise/denoised images. *1 represents CCC of radiomic features calculated based on high-noise images; 2-5 represent CCC of radiomic features calculated based on denoised images by using CGAN when network trained at 25, 50, 75,100 epochs; 6-9 represent CCC of radiomic features calculated based on denoised images by using encoder-decoder network when network trained for 25, 50, 75,100 epochs respectively; 10 represent CCC of radiomic features calculated based on denoised images by using non-local means algorithm.

Table 3-1. Summary of RMSE, content loss and distribution of CCCs of radiomic features

Distribution / Models	RMSE	Content loss	CCCs<0.65	0.65≤CCCs<0.85	CCCs≥0.85
Low-noise Images					
Without denoising	0.0225	0.0706	0/17(0%)* 9/73(12%)** 9/90(10%)***	3/17(18%) 17/73(24%) 20/90(22%)	14/17(82%) 47/73(64%) 61/90(68%)
Non-Local Means	0.0993	0.3280	5/17(29%) 48/73(66%) 53/90(59%)	8/17(47%) 21/73(29%) 29/90(32%)	4/17(24%) 4/73(5%) 8/90(9%)
Encoder-decoder	0.0173	0.0427	0/17(0%) 0/73(0%) 0/90(0%)	1/17(6%) 16/73(22%) 17/90(19%)	16/17(94%) 57/73(78%) 73/90(81%)
CGAN	0.0143	0.0290	0/17(0%) 3/73(4%) 3/90(3%)	0/17(0%) 15/73(21%) 15/90(17%)	17/17(100%) 55/73(75%) 72/90(80%)
High-noise Images					
Without denoising	0.0237	0.0781	5/17(29%)	1/17(6%)	11/17(65%)

			27/73(37%)	20/73(27%)	26/73(36%)
			32/90(36%)	21/90(23%)	37/90(41%)
Non-Local Means	0.1095	0.3941	6/17(35%)	7/17(41%)	4/17(24%)
			52/73(71%)	17/73(23%)	4/73(6%)
			58/90(64%)	24/90(27%)	8/90(9%)
Encoder-decoder	0.0175	0.0443	0/17(0%)	1/17(6%)	16/17(94%)
			4/73(5%)	13/73(18%)	56/73(77%)
			4/90(4%)	14/90(16%)	72/90(80%)
CGAN	0.0146	0.0305	0/17(0%)	1/17(6%)	16/17(94%)
			0/73(0%)	13/73(18%)	60/73(82%)
			0/90(0%)	14/90(16%)	76/90(84%)

* represents the summary of fist order features; ** represents the summary of textural features; *** represents the summary of all features.

Effect of Different Models

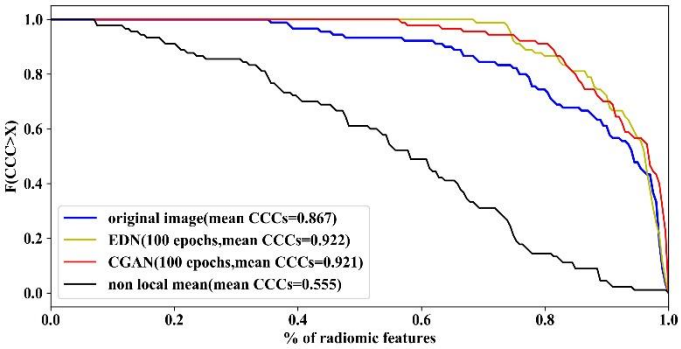
As shown in Table 1, the baseline RMSE and content loss of high-noise and low-noise images (prior to denoising) were 0.0237(RMSE)/0.0781(content loss) and 0.0225/0.0706, respectively. The RMSE and content loss decreased to 0.0175/0.0443 for the high-noise images and 0.0173/0.0427 for low-noise images, respectively, by using EDN denoising. In comparison, RMSE and content loss were decreased even further to 0.0146/0.0305 and 0.0143/0.0290 using CGAN denoising.

As shown in Figure 3-8, the baseline mean CCC of radiomics in high-noise and low-noise images were 0.681 [95%CI, (0.617,0.745)] and 0.867 [95%CI, (0.833,0.901)], respectively. By comparison, the mean CCC for denoised images using the EDN and the CGAN (both trained for 100 epochs) were significantly improved to about 0.92 [95%CI, (0.909,0.935)] for high-noise as well as low-noise images.

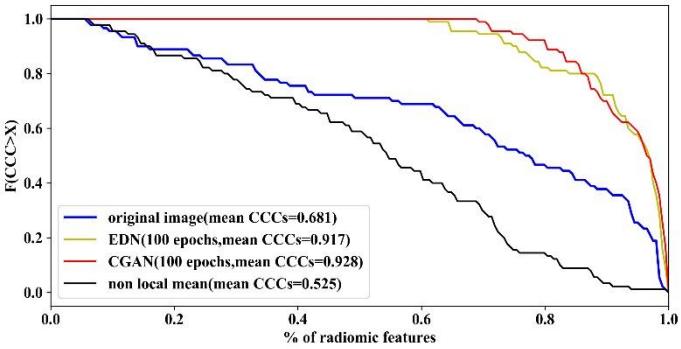
In regards to a traditional denoising method, the RMSE and content loss increased to 0.1095/0.3941 and 0.0993/0.3280 when using the non-local means algorithm. Likewise, the mean CCC of radiomics in images denoised using non-local means were decreased - 0.525 [95%CI, (0.474, 0.576)] and 0.555 [95%CI, (0.507, 0.604)] - for high-noise and low-noise images, respectively.

A cumulative distribution function of CCCs for different models when trained for 100 epochs is shown in Figure 3-8. The EDN and the CGAN both improved the overall reproducibility of radiomic features significantly, especially in the high-noise images. Non-local means algorithm was able to remove noise from images as we can see in Figure 3-6 (c-1), however it also

led to detail loss, as shown in Figure 3-6 (c-2). This is a well-known compromise of denoising referred to as “smoothing” in literature [5]. In other words, smoothing due to a specific low-pass noise filter in the traditional method caused a deterioration of radiomics reproducibility when measured by its CCC (Wilcoxon signed-rank test, p-value <0.01).



(a)



(b)

Figure 3-8. Cumulative distribution function of radiomic features’ CCCs on images denoised using different models: (a) low-noise images; (b) high-noise images. The plots show the proportion of radiomic features with a CCC higher than x, across all possible values of the CCC (horizontal axis).

For example, a point at (0.8, 0.7) implies that 70% of the radiomic features have a CCC higher than 0.8.

Effect of Different Numbers of Training Epochs

An example of original, noisy and denoised CT scan after different training epochs by EDN and CGAN is shown in Figure 3-9. A cumulative distribution function of the CCCs of radiomic features on images denoised with the CGAN trained for different numbers of epochs is shown in Figure 3-10.

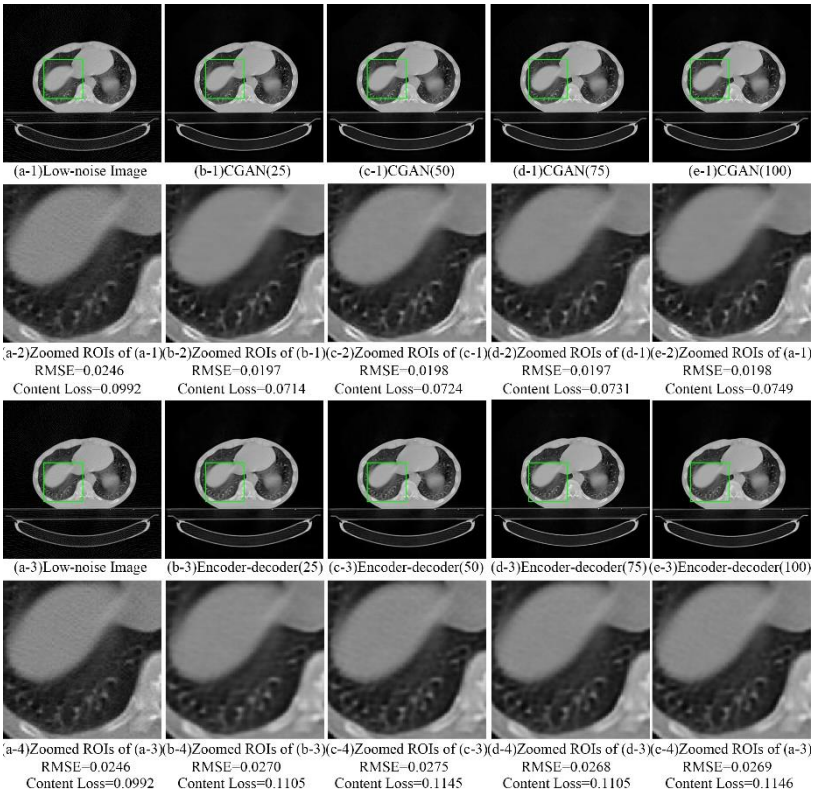
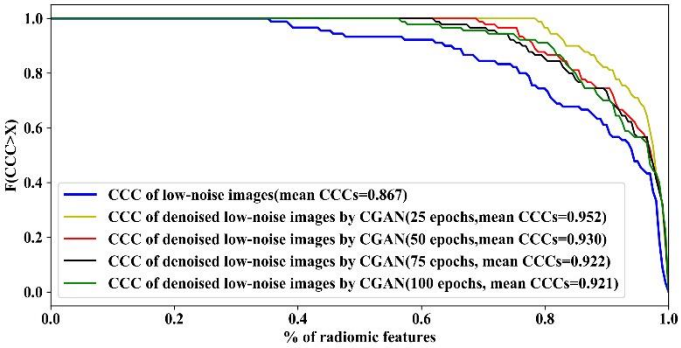
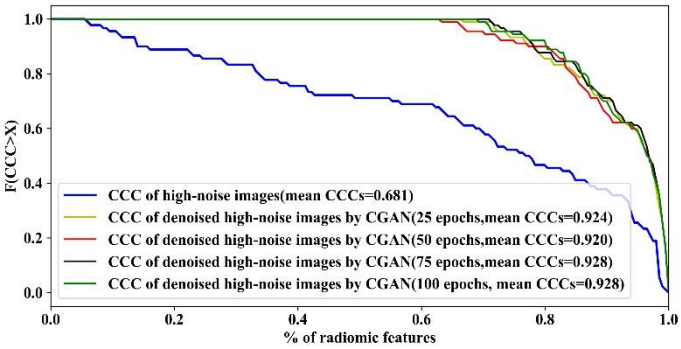


Figure 3-9. An example of an original, noisy and denoised CT scan at different training epochs by using the encoder-decoder network and the CGAN. As we can see figures (b-4) to (e-4), the RMSE and content loss of denoised images is higher in this particular case than the original low-noise images.



(a)



(b)

Figure 3-10. Cumulative distribution function of CCCs for image denoised by CGAN trained for different numbers of epochs. (a) Cumulative distribution function of CCCs based on denoised low-noise by using CGAN trained for different numbers of epochs; (b) Cumulative

distribution function of CCCs based on denoised high-noise by using CGAN trained for different numbers of epochs.

This indicated the best results in low-noise images were achieved when the CGAN was trained for 25 epochs. However, there was no significant difference in high-noise images. This same observation is holds for the EDN (Supplementary Figure 3), where the results seemed poorest when trained for 50 epochs. We speculate that this dip was an artefact of training and model convergence, rather than any meaningful finding. Table 3-2 shows the RMSE, content loss and proportion of poor, medium, and good reproducibility radiomic features denoised with CGAN, as a function of training epochs. (A corresponding table for the EDN is in Supplementary Table 6). The RMSE and content loss in the whole dataset for images denoised using CGANs trained for different numbers of epochs are also shown in Table 3-2.

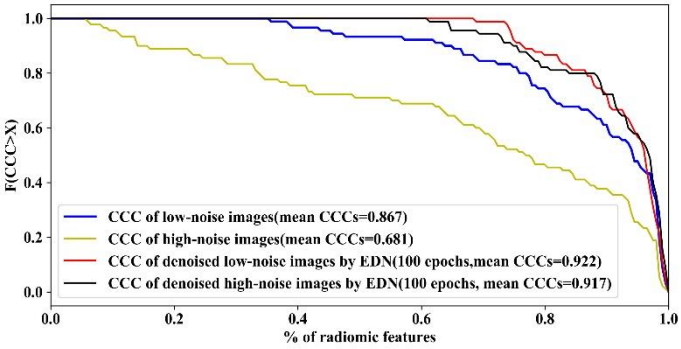
Table 3-2. RMSE, content loss and ratio of poor, medium, and good reproducibility radiomic features for images denoised by the CGAN trained for different numbers of epochs

Training length / Noisy images	25 Epochs	50 Epochs	75 Epochs	100 Epochs
Low-noise Images				
RMSE	0.0148	0.0144	0.0142	0.0143
Content loss	0.0295	0.0290	0.0276	0.0290
CCCs ≥ 0.85	90%	81%	78%	80%
0.65≤CCCs<0.85	10%	19%	20%	17%

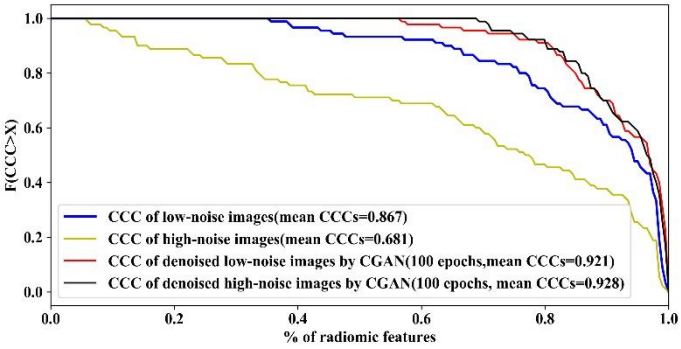
CCCs<0.65	0%	0%	2%	3%
High-noise Images				
RMSE	0.0150	0.0148	0.0144	0.0146
Content loss	0.0309	0.0312	0.0291	0.0305
CCCs > 0.85	80%	79%	83%	84%
0.65≤CCCs<0.85	20%	20%	17%	16%
CCCs<0.65	0%	1%	0%	0%

Effect of Different Noise Intensities

As mentioned in Methods Section, the models were not retrained for denoising low-noise images. Figure 3-11 shows the cumulative distribution functions of CCCs of radiomic features extracted from images to which different levels of noise intensity had been applied. We compared the CCC distributions of radiomic features calculated on images denoised from high-noise images with those of images denoised from low-noise images using the Wilcoxon signed-rank test. The p-value for the CGAN and the EDN were 0.671 and 0.109, respectively, implying no significant differences. That is, our results show CGANs and EDNs trained to denoise high-noise images can also be applied to denoise images with variable levels of noise with comparable performance. A single well-trained model can be used to improve radiomics reproducibility in images with different levels of noise, and especially when the imaging dose is lower than 50 mAs.



(a)



(b)

Figure 3-11. Cumulative distribution functions of CCCs for image noised with different intensities. (a) Cumulative distribution functions of CCCs for image noised with different intensities (Encoder-decoder network); (b) Cumulative distribution functions of CCCs for image noised with different intensities (CGAN).

Effect in real low dose CT scans

The RMSEs of denoised versus full dose CT scans of phantoms using the EDN and CGAN were 0.0182 and 0.0140 respectively, which was better than 0.0231 in the original low dose CTs. The content loss in CT scans

denoised using EDN and CGAN was also improved - 0.0433 and 0.0289, respectively - compared to 0.0702 in the original low dose CTs.

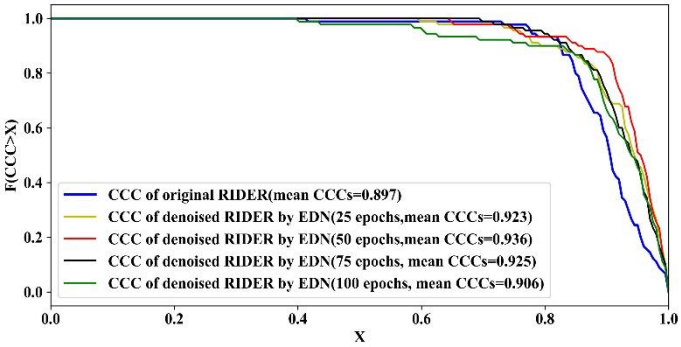
The results in Table 3-3 show that denoising using the EDN and the CGAN improved the mean CCCs of radiomic features in the RIDER dataset from 0.89 [95%CI, (0.881, 0.914)] to around 0.94 [95%CI, (0.927,0.951)], and the percentage of features with a CCC higher than 0.85 increased from 80% to around 90%. The cumulative distribution of CCCs for radiomic features in the RIDER test set is given in Figure 3-12. An example of an original, denoised RIDER image using EDN and CGAN is shown in Figure 3-13. This scan proved especially troublesome during previous experimentation with cycle GANs (not in the scope of this paper), so it was excluded from the analysis.

We may conclude that these generative models can improve the test-retest reliability of radiomic features calculated from real low dose CT scans, such as the ones in the RIDER dataset.

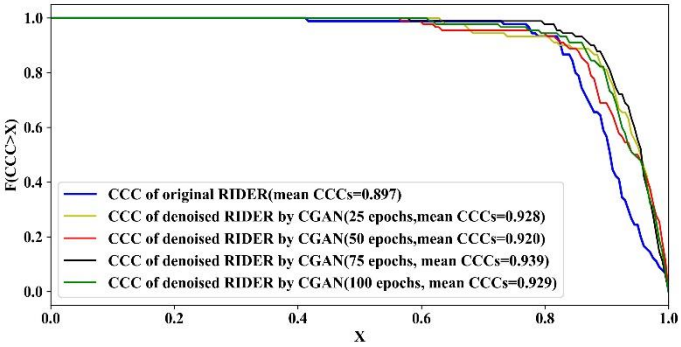
Table 3-3. Effect of denoising on test-retest reliability of radiomic features

Epochs / CCCs>0.85	25 Epochs	50 Epochs	75 Epochs	100 Epochs	Original RIDER
Encoder-decoder	78/90(0.92*) (0.91,0.94)**	82/90(0.94) (0.92,0.95)	78/90(0.93) (0.91,0.94)	78/90(0.91) (0.88,0.93)	72/90(0.90) (0.88,0.91)
CGAN	81/90(0.93) (0.91,0.95)	81/90(0.92) (0.90,0.93)	85/90(0.94) (0.93,0.95)	83/90(0.93) (0.91,0.94)	

*Mean CCCs of radiomic features, ** Mean 95% confidence intervals of CCCs.



(a)



(b)

Figure 3-12. Cumulative distribution functions of CCCs for original and denoised CT scans in the RIDER dataset (a) using a CGAN trained for different numbers of epochs; (b) using an encoder-decoder network trained for different numbers of epochs

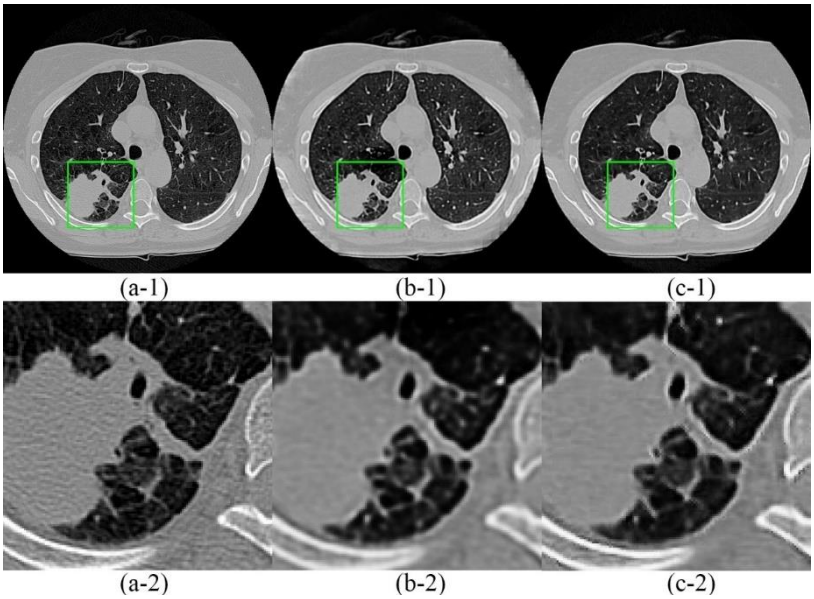


Figure 3-13. Example of denoised image from the RIDER dataset. (a-1) Original image; (b-1) Image denoised by EDN (100 epochs); (c-1) Image denoised by CGAN (100 epochs); (a-2) to (c-2) Zoomed ROIs for (a-1) to (c-1).

Discussion

Our objective was to test two different deep learning generative models, EDN and CGAN, to improve the SNR in CT images and explore its effect after denoising on increasing radiomic features reproducibility. The overall results of our experiments show that an equally good performance, in terms of reducing RMSE and content loss, as well as increasing the average CCC of radiomic features, was obtained by the CGAN and the EDN. However, the poor performance of the non-local means algorithm likely stems from the reduced ability of the traditional algorithm to keep fine image details

during denoising, relative to CGANs and EDNs, which is the point made in Figure 3-6.

We chose the CCC as our metric for reproducibility rather than the intraclass correlation coefficient (ICC) [46] because it is well suited to paired before-and-after values, and it relies on fewer assumptions than the ICC [47]. As a sanity check, we also calculated ICCs for a subset of these experiments and found that they were equal to their respective CCCs up to the second decimal place. Therefore, we considered the additional reporting of ICCs to be of limited added value.

The non-local means algorithm showed poorer relative performance against EDN and CGAN in both aspects of noise removal and content loss, as can be clearly seen by CCC of radiomics feature subgroups in Table 3-1 and also by visual inspection in Figure 3-6. The non-local means method appears to have moderate performance in maintaining first order features' reproducibility. We speculate, however, that it is the content loss (i.e. aforementioned "smoothing" phenomenon, resulting in loss of fine details from the image) that is associated with significantly worse reproducibility among the subset of textural features, when using the non-local means algorithm. The lack of reproducibility among textural features due to the non-local means algorithm is also clearly apparent when comparing column 10 in the heatmap (Figure 3-7) with the other columns 2-9.

Among the two generative models, CGAN showed slightly better performance than EDN in terms of removing random noise and retaining image details, but in terms of radiomics reproducibility both had similar outcomes. One of the possible reasons for this might be that radiomics

features are no longer sensitive to small differences in noise or detail left behind after adequate denoising [5].

The improvement in feature reproducibility after denoising of low dose CT scans (<=50 mAs) has been demonstrated above. However, it is still worth testing if our generative models can perform equally well in a wider range of scanners and imaging conditions. The results from the low-noise images suggest that the denoising models generalize to images with different noise intensity, but greater variation in the scanning and reconstruction setting are needed to establish how generalizable this is. If so, our models might significantly reduce application barriers for clinicians and radiomics researcher.

The main limitation of this study is that training data were not actually real paired low/full dose CT images taken of the same human subjects. There is the obvious difficulty of justifying and collecting such paired images in a practical clinical setting. Overcoming these practical constraints with simulated noise, we were able to show in our experiments that denoising did have a beneficial impact on real low dose (RIDER) CT scans in terms of radiomics reproducibility. Further, we did not show a direct benefit of reproducibility for any clinical application of radiomics, such as improved performance of a prediction model. This was beyond the scope of this article, but we argue that better reproducibility of radiomic features in itself is likely to improve external validity, in general, for any potential application of radiomics.

We expended only limited time on fine-tuning and exhaustive testing of hyperparameters of the generative models. We used the same

hyperparameters and training strategies for CGAN used in the original setting pix2pix [33]. The results shown in our experiments might not be the best possible results achievable with these models, especially for the CGAN, but nonetheless we demonstrated the concept of improving radiomics feature reproducibility.

The loss functions used to train the generative models might not have been the most optimal for radiomics feature reproducibility either, however we did try to improve feature reproducibility independently by minimising RMSE and content loss. Choice of loss functions can significantly affect the convergence of the network, but we did not fully investigate alternative loss functions for convergence. Other than training curves, no additional measures were implemented here to guarantee the convergence of the networks.

We transformed DICOM images to PNG images to use them as the input to the networks. The transformation will result in minimal information loss due to numerical rounding errors, but we do not believe this to have had a major detrimental effect. All of our training data originated from the same single clinic, though the RIDER test images were obtained from a different hospital. The robustness and generalizability of our models needs to be extensively tested using data from multiple centers.

Lastly, it may be possible to select other quantitative metrics to evaluate the goodness of the generative model-based approaches, however only RMSE, content loss and CCC were used for this article. This may admit the possibility of apparently worse image quality after denoising when using an EDN, as shown in Figure 3-9.

Future work might improve the reproducibility of radiomics features even further. For example, the mask used to calculate radiomic features was drawn by a clinician on full dose CT images. However, the masks might have been different if they had been drawn based on noisy images, such as low dose CT scans. Therefore, a noise-insensitive tumor segmentation algorithm could potentially improve low dose CT radiomic feature calculation. Moreover, we inserted Gaussian noise into the sinogram for our study, but the noise distribution in real low dose CT images might not exactly be Gaussian. This could lead to an overestimation of the performance of our models. Therefore, further studies using real data are needed.

Conclusions

In this article, we compared two different deep-learning generative models for image denoising – an EDN and a CGAN – and evaluated its utility for improving radiomics feature reproducibility (assessed by the CCC metric) in noisy images such as low dose CT scans. We compared their performance to a well-established non-deep learning based denoising method – the non-local means algorithm. We added noise at two intensities to real full dose CT images to simulate different kinds of low-dose CT images. All models were trained using high-noise images, then high and low-noise images were denoised using these models for validation, without any retraining. The results show that a non-local means algorithm for denoising may not be suitable for improving reproducibility of radiomic features. EDNs and CGANs do indeed improve the reproducibility of radiomics features in post-denoised CT, and both generative methods were about equivalent in terms of noise removal and detail retention. In addition, the results from low-noise images were not significantly different to those of high-noise images. These results imply that images with varying levels of noise can be denoised using our trained models to potentially improve the reproducibility of radiomic features. To the authors' best knowledge, this article is the first to show that improvement in the reproducibility of radiomics features is feasible based on denoising low-dose CT images.

Appendix

Support materials of this Chapter can be found in this link.

References:

[1]Lambin, Philippe, et al. "Radiomics: the bridge between medical imaging and personalized medicine." *Nature reviews Clinical oncology* 14.12 (2017): 749.

[2]Desseroit, Marie-Charlotte, et al. "Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non–small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort." *Journal of Nuclear Medicine* 58.3 (2017): 406-411.

[3]Bogowicz, Marta, et al. "Stability of radiomic features in CT perfusion maps." *Physics in Medicine & Biology* 61.24 (2016): 8736.

[4]Tixier, Florent, et al. "Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET." *Journal of Nuclear Medicine* 53.5 (2012): 693-700.

[5]Bagher-Ebadian, Hassan, et al. "On the impact of smoothing and noise on robustness of CT and CBCT radiomic features for patients with head and neck cancers." *Medical physics* 44.5 (2017): 1755-1770.

[6]Zhang, Bin, et al. "Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma." *Clinical Cancer Research* 23.15 (2017): 4259-4269.

[7]Aerts, Hugo JWL, et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." *Nature communications* 5.1 (2014): 1-9.

[8]Traverso, Alberto, et al. "Repeatability and reproducibility of radiomic features: a systematic review." *International Journal of Radiation*

Oncology Biology* Physics* 102.4 (2018): 1143-1158.

[9]van Timmeren, Janna E., et al. "Test–retest data for radiomic feature stability analysis: generalizable or study-specific?." *Tomography* 2.4 (2016): 361.

[10]Huang, Qiao, et al. "Interobserver variability in tumor contouring affects the use of radiomics to predict mutational status." *Journal of Medical Imaging* 5.1 (2017): 011005.

[11]Berenguer, Roberto, et al. "Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters." *Radiology* 288.2 (2018): 407-415.

[12]Meyer, Mathias, et al. "Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings." *Radiology* 293.3 (2019): 583-591.

[13]Zuo, Wangmeng, et al. "Gradient histogram estimation and preservation for texture enhanced image denoising." *IEEE transactions on image processing* 23.6 (2014): 2459-2472.

[14]Jiang, Xin, et al. "A novel super-resolution CT image reconstruction via semi-supervised generative adversarial network." *Neural Computing and Applications* 32.18 (2020): 14563-14578.

[15]Badawi, Ramsey D., et al. "First human imaging studies with the EXPLORER total-body PET scanner." *Journal of Nuclear Medicine* 60.3 (2019): 299-303.

[16]Wang, Jing, et al. "Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose X-ray computed tomography." *IEEE transactions on medical imaging* 25.10 (2006): 1272-1283.

[17]Musolino, Stephen V., Joseph DeFranco, and Richard Schlueck. "The ALARA principle in the context of a radiological or nuclear emergency." *Health physics* 94.2 (2008): 109-111.

[18]Sharma, Abhishek, and Vijayshri Chaurasia. "A review on magnetic resonance images denoising techniques." *Machine Intelligence and Signal Analysis*. Springer, Singapore, 2019. 707-715.

[19]Kollem, Sreedhar, Katta Rama Linga Reddy, and Duggirala Srinivasa Rao. "A Review of Image Denoising and Segmentation Methods Based on Medical Images." *International Journal of Machine Learning and Computing* 9.3 (2019).

[20]Mredhula, L., and M. A. Dorairangasamy. "An extensive review of significant researches on medical image denoising techniques." *International Journal of Computer Applications* 64.14 (2013).

[21]Shan, Hongming, et al. "3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network." *IEEE transactions on medical imaging* 37.6 (2018): 1522-1534.

[22]Chen, Hu, et al. "Low-dose CT with a residual encoder-decoder convolutional neural network." *IEEE transactions on medical imaging* 36.12 (2017): 2524-2535.

[23]Yang, Qingsong, et al. "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss." *IEEE transactions on medical imaging* 37.6 (2018): 1348-1357.

[24]Kang, Eunhee, et al. "Deep convolutional framelet denosing for low-dose CT via wavelet residual network." *IEEE transactions on medical*

imaging 37.6 (2018): 1358-1369.

[25]Fan, Fenglei, et al. "Quadratic Autoencoder (Q-AE) for Low-dose CT Denoising." *IEEE Transactions on Medical Imaging* (2019).

[26]Xu, Wenju, Shawn Keshmiri, and Guanghui Wang. "Adversarially approximated autoencoder for image generation and manipulation." *IEEE Transactions on Multimedia* 21.9 (2019): 2387-2396.

[27]Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[28]Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

[29]Yi, Xin, Ekta Walia, and Paul Babyn. "Generative adversarial network in medical imaging: A review." *Medical image analysis* (2019): 101552.

[30]Zhang, Zizhao, Yuanpu Xie, and Lin Yang. "Photographic text-to-image synthesis with a hierarchically-nested adversarial network." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[31]Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." *arXiv preprint arXiv:1411.1784* (2014).

[32]Clark, Kenneth, et al. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository." *Journal of digital imaging* 26.6 (2013): 1045-1057.

[33]Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[34]Buades, Antoni, Bartomeu Coll, and J-M. Morel. "A non-local algorithm for image denoising." *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 2. IEEE, 2005.

[35]Dabov, Kostadin, et al. "Image denoising by sparse 3-D transform-domain collaborative filtering." *IEEE Transactions on image processing* 16.8 (2007): 2080-2095.

[36]Liu, Yan-Li, et al. "A robust and fast non-local means algorithm for image denoising." *Journal of computer science and technology* 23.2 (2008): 270-279.

[37]Aerts, H. J. W. L., Wee, L., Rios Velazquez, E., Leijenaar, R. T. H., Parmar, C., Grossmann, P., … Lambin, P. (2019). *Data From NSCLC-Radiomics* [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI

[38]McCollough, Cynthia H., et al. "Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 Low Dose CT Grand Challenge." *Medical physics* 44.10 (2017): e339-e352.

[39]Zhovannik, Ivan, et al. "Learning from scanners: Bias reduction and feature correction in radiomics." *Clinical and translational radiation oncology* 19 (2019): 33-38.

[40]Zhao, Binsheng, et al. "Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non–small cell lung cancer." *Radiology* 252.1 (2009): 263-272.

[41]Shi, Zhenwei, et al. "Ontology-guided radiomics analysis workflow (O-RAW)." *Medical Physics* 46.12 (2019): 5677-5684.

[42]Van Griethuysen, Joost JM, et al. "Computational radiomics system to

decode the radiographic phenotype." *Cancer research* 77.21 (2017): e104-e107.

[43]Lawrence, I., and Kuei Lin. "A concordance correlation coefficient to evaluate reproducibility." *Biometrics* (1989): 255-268.

[44]Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." *European conference on computer vision*. Springer, Cham, 2016.

[45]Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[46]McGraw, Kenneth O., and Seok P. Wong. "Forming inferences about some intraclass correlation coefficients." *Psychological methods* 1.1 (1996): 30.

[47]Chen, Chia-Cheng, and Huiman X. Barnhart. "Comparison of ICC and CCC for assessing agreement for data without and with replications." *Computational statistics & data analysis 53.2* (2008): 554-564.

Chapter 4
Lung Cancer Diagnosis Using Deep Attention Based Multiple Instance Learning and Radiomics

Junhua Chen, Haiyan Zeng, Chong Zhang, Zhenwei Shi, Andre Dekker, Leonard Wee, Inigo Bermejo

Adapted from

Junhua Chen, et al. Lung cancer diagnosis using deep attention-based multiple instance learning and radiomics. Medical Physics. 2022; 49(5): 3134– 3143.

DOI: https://doi.org/10.1002/mp.15539

Abstract

Background: Early diagnosis of lung cancer is a key intervention for the treatment of lung cancer in which computer aided diagnosis (CAD) can play a crucial role. Most published CAD methods perform lung cancer diagnosis by classifying each lung nodule in isolation. However, this does not reflect clinical practice, where clinicians diagnose a patient based on a set of images of nodules, instead of looking at one nodule at a time. Besides, the low interpretability of the output provided by these methods presents an important barrier for their adoption.

Method: In this article, we treat lung cancer diagnosis as a multiple instance learning (MIL) problem, which better reflects the diagnosis process in the clinical setting and provides higher interpretability of the output. We selected radiomics as the source of input features and deep attention-based MIL as the classification algorithm. The attention mechanism provides higher interpretability by estimating the importance of each instance in the set for the final diagnosis. In order to improve the model's performance in a small imbalanced dataset, we propose a new bag simulation method for MIL.

Results and Conclusion: The results show that our method can achieve a mean accuracy of 0.807 with a standard error of the mean (SEM) of 0.069, a recall of 0.870 (SEM 0.061), a positive predictive value of 0.928 (SEM 0.078), a negative predictive value of 0.591 (SEM 0.155) and an area under the curve (AUC) of 0.842 (SEM 0.074), outperforming other MIL methods. Additional experiments show that the proposed oversampling strategy significantly improves the model's performance. In addition, our experiments show that our method provides a good indication of the

importance of each nodule in determining the diagnosis, which combined with the well-defined radiomic features, make the results more interpretable and acceptable for doctors and patients.

Keyword: Lung Cancer diagnosis; Multiple Instance Learning; Attention Mechanism; Radiomics

Introduction

According to the statistics from the World Health Organization (WHO), lung cancer is the most frequently diagnosed malignant carcinoma and the leading cause of cancer death worldwide, accounting for an estimated 2.09 million deaths in 2018 [1][2]. Early diagnosis and treatment can reduce a lung cancer patient's mortality significantly. A plausible method for early lung cancer diagnosis is the routine use of low dose computed tomography (CT) scans [3]. To date, radiologists typically need to visually inspect CT scans slice by slice, which is costly and time-consuming as well as susceptible to human error [4][5]. Computer aided diagnosis (CAD) for rapid early lung nodules classification based on low-dose CT imaging has therefore attracted much attention from researchers during the last decades [6][7].

The development of CAD for lung nodules classification has reached new peaks in last decade mainly due to breakthroughs in deep learning neural networks [8] and its application to a wide range of medical image analysis tasks. Several deep learning-based lung nodule classification methods have been proposed in recent years, with steadily improving state-of-art performance. Shen et al. [9] developed a multi-scale convolutional neural network (CNN) to extract features (referred to as 'deep features' [10] in the literature) then applied a supervised random forest classifier to the deep features, reporting an accuracy of 86%. Xie et al. [11] combined handcrafted features with deep features to classify each nodule as either benign or malignant, achieving an AUC of 0.96. Alakwaa et al. [12] combined the LUNA16 [13] dataset with a subset of the National Lung Screening Trial (NLST) [14], then used a pre-trained U-Net to segment potential nodules

from a CT scan automatically. The segmented nodules were passed to a 3D CNN to detect early-stage lung cancer, achieving an AUC of 0.83 in a randomly-split test cohort from the abovementioned data. Ardila et al. [15] developed an end-to-end set of 3D CNN modules to compute the overall risk of lung malignancy based on autodetection of nodules, using the full-size publicly available NLST dataset. In a retrospective reader study, their model outperformed six experienced radiologists with absolute reductions of 11% and 5% in false positives and false negatives, respectively.

The need for transparency, interpretability and explainability in such computer-aided diagnostic recommendations will grow to become increasingly prominent in the immediate future. A crucial piece of law, the General Data Protection Regulation (GDPR), governs the rights of European Union (EU) citizens as human data subjects and addresses processing by automated means for decision-making anywhere in the world if it concerns an EU individual. Specifically, the GDPR enshrines the right of an individual to receive “meaningful information about the logic involved” in an automated decision concerning them, and on that basis to either legally challenge the decision, or exercise conscientious objection to the use of an automated means for deriving the decision [16].

While definition of “meaningful” is open for debate, it is clearly helpful to be able to point at specific regions of interest (ROIs) that were strongly triggering for the diagnostic recommendation, along with related features of lung cancer and non-lung cancer cases. In this way, a human radiologist can review the information in depth, and either confirm or over-rule the recommendations of an automated system. Irrespective of a right to an explanation, a computerized diagnostic support system with high

transparency and high interpretability would be immensely valuable in clinical practice.

For automated diagnosis of lung cancer, a deep learning-based system can be applied in two levels: at nodule level, to identify potential malignant nodule(s) for further biopsy and performing diagnosis at patient level. Generally speaking, nodule classification methods need a label for each nodule to be able to train a model [9][11]. However, labelling each nodule is more time-consuming and expensive than having a label for each patient, which is usually already available in hospital records. In this study we focus on deep learning methods for lung cancer diagnosis that can make use of the existing data to develop a lung cancer CAD system that classifies patients based on multiple suspected nodules in the entire CT series without the need to assign a label to each nodule (i.e. each instance), and at the same time provide high visibility of the triggering features of its recommendation. Multiple instance learning (MIL) with attention mechanism [17][19] fits this need well. In MIL, the nodules are grouped into 'bags of instances' (assuming multiple nodules in one CT examination of the chest area). The task is hence to determine the diagnosis for the subject as a whole. Only the subject-level diagnoses (i.e. the bag labels) are needed, but not individual labels of every nodule found in the subject [20]. This approach is thus more amenable to real-world data mining in lung cancer, since the subject level diagnosis is much more widely available than annotations on each nodule.

Research on MIL problems has progressed along instance-level versus embedding-level solutions [21], with the latter seeming to perform better at subject-level classification [22]. Widely-used embedding approaches include MI-SVM [23], mi-Graph [24], miVLAD [25] and MI-Net [25], but

the shortcoming of these is the lack of transparency of triggering instance(s). An attention-based deep MIL [21][33] has been recently introduced, that allows a deep learning model to estimate the contribution of each instance to the predicted subject label, using the well-established attention mechanism [26].

The objective of this work was to develop a lung cancer classification model at the subject (patient) level from multiple examined nodules, without the need to have specific expert findings reported at the level of each individual nodule. An MIL method with an additional deep attention mechanism was used to help draw an expert clinician's eye towards the individual nodules that were strongly triggering for the model's diagnostic recommendation. We propose that this will be important by way of offering better interpretability and the possibility of human expert verification of the internal logic of the algorithm. A selection of commonly-used hand-crafted radiomics features was used as a source of image features, and we also compared a number of alternative MIL methods. We have re-used an existing open access data collection for training and cross-validation. Source code will open access for public at https://gitlab.com/UM-CDS/combine-mil-and-radiomics-for-lung-screening) and additional details of the system architecture will be given in Supplementary Materials.

Methods

Dataset

The primary data source of data is an open access collection from the Lung Image Database Consortium (LIDC-IDRI) [31], accessed at The Cancer Imaging Archive (TCIA) during May 2020 [32] under a Creative Commons

Attribution Non-Commercial 3.0 Unported (CC BY-NC) license. The details of subjects in LIDC-IDRI have been provided elsewhere [31], but briefly: the collection comprises 1018 clinical chest CT examinations from seven disjoint institutions. Radiologists working independently entered 7371 annotations, of which there were 2669 consensus nodules. We excluded subjects with unreported or unknown diagnosis, and excluded nodules below 3mm in diameter according to current diagnosis protocols [34][35]. This resulted in 110 unique subjects with a total of 310 nodules eligible for consideration. Binary masks for the nodules were provided in the data collection as an XML file. Numbers of subjects and nodules excluded, along with reason, are provided in Figure 4-1 below. From the summary of diagnostic findings in Table 4-1, we note that the majority of subjects and lung nodules in the dataset are positive for lung cancer; 75% and 77%, respectively. Index of available patients for experiments in LIDC-IDRI can be found in Supplementary Table 1.

Table 4-1 summarizes the radiological findings available in the selected subset with definitive subject-level diagnosis and nodule-level classification.

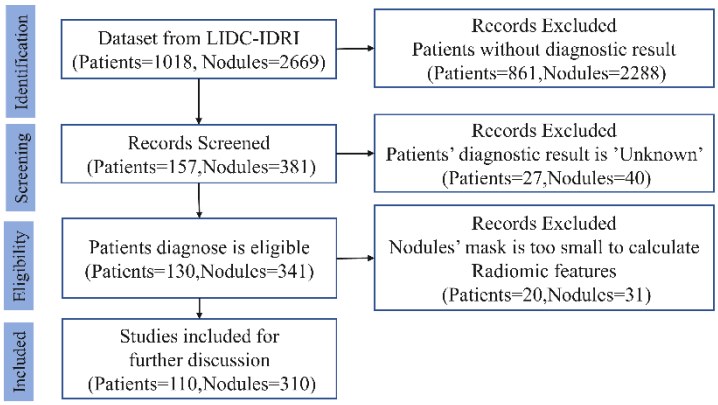


Figure 4-1. Sample selection flowchart describing the number of subjects and the number of nodules selected for this analysis.

Table 4-1. Number of patients and nodules according to ground truth diagnosis in the dataset

	Lung cancer	**Not lung cancer**	Total
Numbers of (% of total) patients	82 (75%)	28 (25%)	110
Numbers of (% of total) nodules	239 (77%)	71 (23%)	310

Image acquisition settings

The LIDC-IDRI contains a heterogeneous set of CT of subjects from different institutions. We used axial CT images with dimension of 512x512 pixels. Radiation exposure of selected samples ranged from 3 milliampere-seconds (mAs) to 534 mAs (median: 147.5 mAs), and reconstructed slice thicknesses ranging from 0.6 mm to 5.0 mm (median: 2.0 mm).

Feature extraction

Radiomics features were extracted using an open-source Python library pyRadiomics (v2.2.0) [36]. Images were resampled to 2 mm isotropic voxels prior to feature extraction. A total of 103 features were extracted. These consisted of 13 morphology (shape) features, 17 intensity-histogram (first-order) features and 73 textural (Haralick) features. Binary masks for the GTV were generated from the XML file in the LIDC-IDRI collection, using an open-access library *pylidc* [37]. DICOM CT images were converted to 3D images by using SimpleITK (v1.2.4) [38] for pyRadiomics feature

extraction. The mathematical definition of each feature has been given in the online documentation. Our pyRadiomics extraction settings (from the params.yaml file) have been included in Supplementary Table 2. All features included in this analysis have been listed in Supplementary Table 3.

Classifier

We used an attention-based MIL for the lung cancer classifier component. This consists of two parts that can be trained end-to-end. First, the transformation network was implemented as three fully connected neural network layers with a dropout rate of 0.5. Additional details about this network are in Supplementary Table 4. To fix the dimension of the input layer of neurons, the 103 features per nodule were duplicated within the same subject until it was the same as the maximum number of nodules per subject, which we found to be 12 in this case. More specifically, each nodule in the same bag should be duplicated with the same probability. For example, if there are 5 nodules in a bag, 3 random nodules need to be duplicates once (i.e., appear twice in total) and 2 random nodules need to be duplicated twice (appear 3 times in total) in the final fixed feature bags. Therefore, the dimension of the input layer should be 103 and one bag consists of 12 vectors (103 x 12). Feature duplication was performed before model training and was also used in model testing.

Second, the attention-based pooling layer implemented the attention mechanism popularized by long short-term memory networks (LSTMs) [39]. The attention mechanism is an important strategy that fits encoder input sequences into a fixed-length internal representation. The architecture of the classifier is illustrated schematically in Figure 4-2.

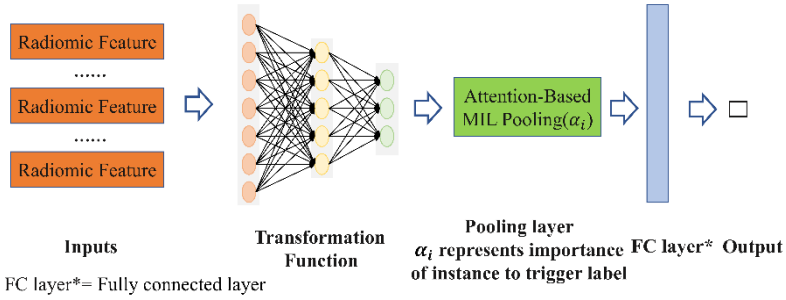


Figure 4-2. Architecture of the Attention-based Deep MIL. Extracted radiomics features are used as the input to the transformation network, which is then pooled with attention. A fully-connected final layer combines the attention-based pooling to give the output probability.

Addressing class imbalance

Imbalance in the outcome frequency (i.e. lung cancer versus not lung cancer) has been known to affect the classifier, biasing this towards the dominant class. Several methods are available to address class imbalance [41] in general, and we applied a novel sampling method to address class imbalance specifically for MIL. It is assumed that all nodules in non-cancer subjects are, by clinical definition, non-cancerous nodules. Synthetic non-cancer patients were thus generated by randomly sampling a finite number of instances out of all the nodules in an aggregated pool of actual non-cancer subjects. On the other hand, synthetic cancer patients could be generated by adding a random number of negative instances sampled from the instances pool (from both negative and positive bags) to the original positive bags. However, we did not simulate cancer patients in our experiments, because positive bags were majority in our dataset. This was only done for the training set; no class imbalance correction was applied in the testing set.

Model development and validation

All work was executed on a Core i7 8565U CPU with 8GB of RAM. The optimizer for network training was stochastic gradient descent (SGD) [42], with batch size 1 and the learning rate fixed at 0.0001. The neural network was trained for 500 epochs (taking 3-4 minutes) per experiment.

We performed experiments for the attention-based MIL in comparison with other MIL approaches - MI-SVM, mi-graph, miVLAD, MI-Net and a naïve MIL algorithm that performs a simple aggregation of the predictions by replacing the attention-based MIL pooling with average MIL pooling [20]. The optimizer, batch size, learning rate and training epochs were set same as attention-based MIL in MI-Net. The setting of hyperparameters in other methods were followed as mentioned in original literatures [23][24][25]. Same training and testing data were used in every running for all methods.

Model training was performed on all the available subjects, taking their respective diagnosis as the “bag label” and the nodules as the instances. We ran 20 repetitions of end-to-end training runs on the hand-crafted features with 5-fold cross-validation in each run and there is no oversampling in testing dataset. For each repetition of 5-fold cross-validation, samples were randomly sorted first and then split into 5 folds, so that each sample was used once for testing and 4 times for training in each repetition. We adjusted for the lower number of non-cancer diagnoses by generating synthetic non-cancer patients as described above (section 2.4). Specifically, we synthesized 60 additional non-cancer subjects from the initial training dataset and added these to the actual 88 training subjects, resulting in a training set containing 148 subjects in total. No synthetic re-sampling was used for positive lung cancer subjects. We further conducted an additional

sensitivity analysis to assess how oversampling to overcome class imbalance might have affected the model's performance by using only the original data of 110 subjects.

The discriminative performance was assessed using the mean and standard error of the mean (SEM) of recall, accuracy, positive predictive value (PPV), negative predictive value (NPV), respectively. For dichotomization of outcome, we used a probability threshold of 0.5 to separate lung cancer from non-lung cancer. The area under the receiver operating characteristic curve (AUC) was computed for each model, the definition of AUC can be found in [43]. Let TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively, then we define recall, accuracy, PPV and NPV as:

$$recall = \frac{TP}{TP+TN} \qquad \text{Equation 4-1}$$

$$accuray = \frac{TP+TN}{TP+TN+FR+FN} \qquad \text{Equation 4-2}$$

$$PPV = \frac{TP}{TP+FP} \qquad \text{Equation 4-3}$$

$$NPV = \frac{TN}{TN+FN} \qquad \text{Equation 4-4}$$

All statistical analysis was done in Python (version 3.6.1).

Results

Figure 4-3 shows the violin plots comparing the results of attention-based MIL with (3a) and without (3b) synthetic minority oversampling. The

estimated mean (with SEM in the parentheses) for recall, accuracy, PPV, NPV and the AUC for the model including the class imbalance correction were: 0.870 (SEM 0.061), 0.807 (SEM 0.069), 0.928 (SEM 0.078), 0.591 (SEM 0.155) and 0.842 (SEM 0.071) respectively. Without the class imbalance correction, these values were: 0.889 (SEM 0.061), 0.768 (SEM 0.059), 0.842 (SEM 0.071), 0.483 (SEM 0.209) and 0.696 (SEM 0.108) respectively. The main effect of the minority oversampling was to improve accuracy, PPV, NPV and AUC. A representative (from a selected repetition) set of AUC curves for the different MIL methods with the same training and testing data can be found in Figure 4-4.

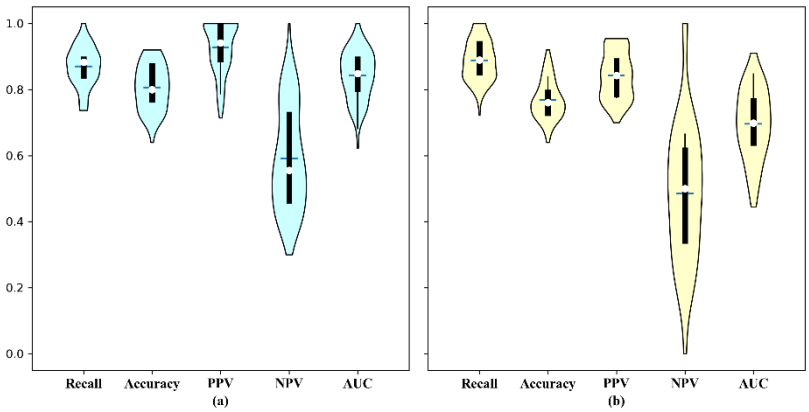


Figure 4-3. Violin plot of the experimental results (a) with oversampling and (b) without oversampling.

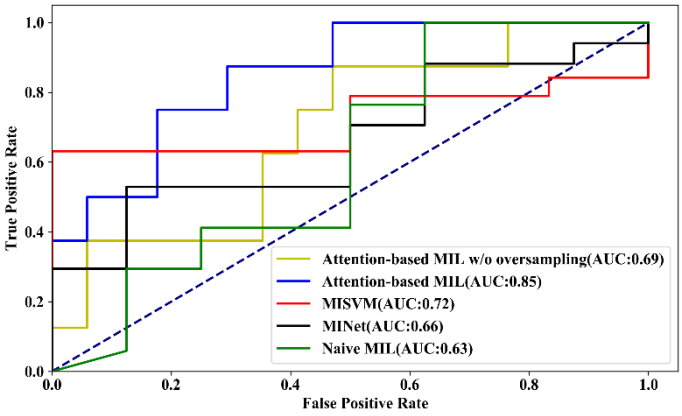


Figure 4-4. An example of AUC curves for different methods with same training and testing data. An AUC curves for Attention-based MIL, Attention-based MIL w/o oversampling, MI-SVM, MI-Net and Naïve MIL.

Table 4-2 summarizes the results of comparing different MIL approaches. Attention based MIL without oversampling achieved the best recall, MI-Net achieved the best PPV and attention based MIL achieved the best accuracy, PPV and AUC. Attention based MIL was better than other methods in PPV and AUC significantly (Wilcoxon test, p <0.01), however, attention based MIL was worse than best result in recall and NPV (Wilcoxon test, p = 0.02 and p <0.01 respectively). Moreover, attention based MIL with oversampling is better than attention-based MIL without oversampling in all metrics significantly except recall (Wilcoxon test, p <0.01 for accuracy, PPV, NPV and AUC, p =0.02 for recall).

The absence of AUCs for mi-graph and miVLAD is due to our reusing of the source code by the LAMDA lab, Nanjing University [44]. Their source code for mi-graph and miVLAD outputs only the classification label (not the probability) and therefore, the AUCs cannot be calculated.

In order to determine the level of oversampling, we ran sensitivity analyses. We gradually increased the number of included simulated non-cancer subjects from 0 to 100 on steps of 20. We ran 20-repeat 5-fold cross-validation for each experiment. The results of sensitivity analysis are shown in Figure 4-5.

As shown in Figure 4-5, including 60 simulation samples results in good performance for all metrics (especially for Recall) with less computation compared with other settings with similar performance.

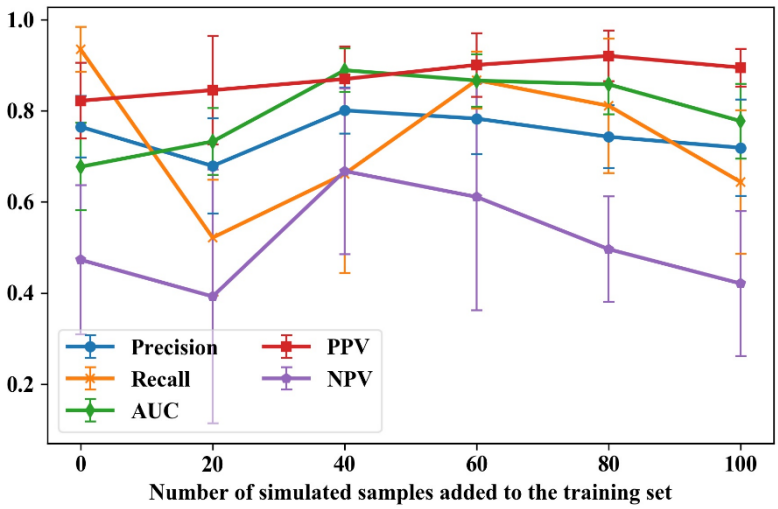


Figure 4-5. Results of sensitivity analysis for different levels of oversampling.

Given how important batch size is for convolutional neural networks training [45], we ran a sensitivity analysis on this parameter. We ran 20-repeat 5-fold cross-validation analyses with different batch sizes (1, 2, 3 and 4) for each experiment. The loss curves for model training with different batch sizes is shown in Figure 4-6 (a) and the performance of models trained with different batch sizes is shown in Figure 4-6 (b).

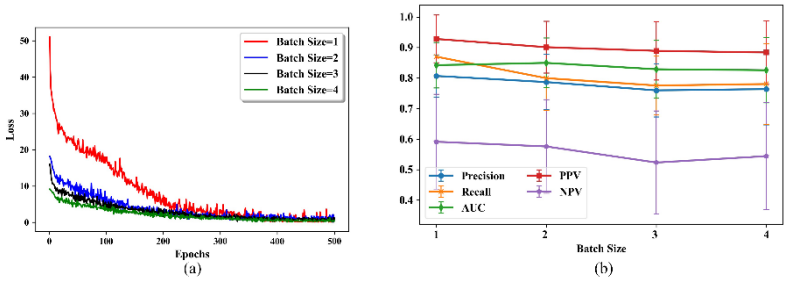


Figure 4-6. Results of sensitivity analysis for different batch sizes. (a) Loss curves for model training with different batch size; (b) performance of models trained with different batch sizes.

As shown in Figure 4-6, the model trained with a batch size of 1 achieved the best performance according to all metrics except AUC (0.842 for batch size 1 vs 0.849 for batch size 2) and the loss of all models converged at the end of the 500 epochs. Therefore, we set the batch size to 1 in this study.

Besides model performance, one of the most appealing aspects that we selected the attention-based MIL method for, was to indicate the instances that might have been strongly influential on the classification. In this case, it would be the relative importance of each nodule when predicting the subject label as either lung cancer or not lung cancer.

A couple of lung cancer examples are shown in Figure 4-7 for two subjects in the dataset, LIDC-IDRI-1004 and LIDC-IDRI-1011. Alpha in Figure 4-7 mean the strength of the attention, value of alpha only meaningful in the same patient and it is meaningless by comparing alphas across patients. The order of nodules was arranged in random way within same patient.

The evaluation of the attention mechanism was performed by one of co-authors -- a radiologist with 3-year experience, who examined some sample patients' weights and agreed with the weighting. In these examples, it is clearly discernable from the weights (α_2 and α_3 larger than α_0 and α_1) that the two rightmost nodules pictured for subject LIDC-IDRI-1004 are much more strongly influential in the diagnostic evaluation compared to the two leftmost nodules. Similarly, for subject LIDC-IDRI-1011, three of the nodules are influential on the subject classification, but the nodule pictured rightmost is not influential at all (alpha < 0.01).

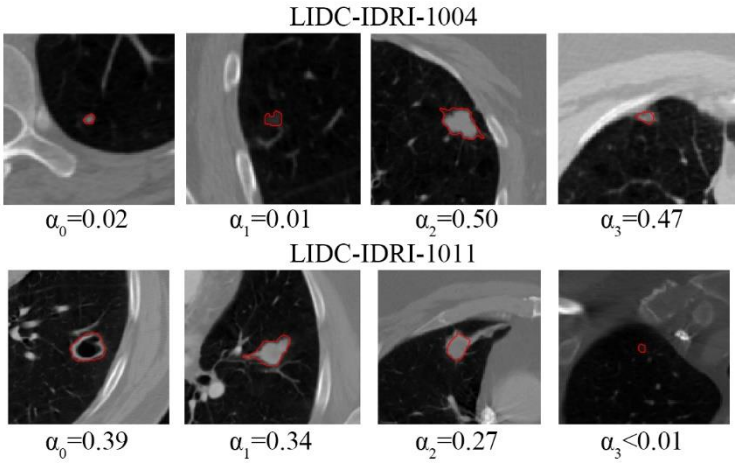


Figure 4-7. An example of attention weights for two positive lung cancer subjects (LIDC-IDRI-1004 and 1011).

Table 4-2. Results of the Attention based Deep MIL approach with class imbalance correction, compared to other MIL methods (Attention-based MIL w/o oversampling, MI-SVM, mi-graph, miVLAD and MI-Net)

Methods	Attention-based MIL	Attention-based MIL w/o oversample	MI-SVM	mi-graph	miVLAD	MI-Net	Traditional MIL
Recall	0.870±0.061	0.889±0.061	0.756±0.084	0.777±0.048	0.871±0.087	0.835±0.109	0.850±0.099
Accuracy	0.807±0.069	0.768±0.059	0.703±0.080	0.749±0.055	0.782±0.063	0.727±0.050	0.748±0.065
PPV	0.928±0.078	0.842±0.071	0.560±0.199	0.772±0.042	0.835±0.059	0.522±0.265	0.835±0.070
NPV	0.591±0.155	0.483±0.209	0.810±0.080	0.713±0.229	0.675±0.160	0.838±0.069	0.478±0.233
AUC	0.842±0.071	0.696±0.108	0.625±0.099	--	--	0.662±0.093	0.681±0.080

Discussion

Our objective was to propose a lung cancer classification at the subject level from multiple examined nodules, with an attention mechanism for improving the interpretability. The results show that our proposed classification achieves good performance compared to other MIL methods and that the unique characteristic of the deep attention-based MIL, namely attention weights, potentially makes our method more interpretable for clinicians.

To see the effect of minority oversampling to overcome class imbalance, we tested the model with and without the oversampling. The results show that the oversampling improved the model's performance significantly in accuracy, PPV, NPV and AUC by comparing Attention-based MIL without oversampling. However, there are seem some decrease in recall.

We observed from Figure 4-3 that minority oversampling has a major effect on the AUC. In fact, the AUC sinks below 0.5 in some experiments without oversampling. This can be explained by the fact that the AUC is more sensitive to the classification performance of the model with the minority class than either accuracy or recall [40].

We proposed a new synthetic subject generation method that can be used to overcome class imbalance by oversampling the minority class. We did this by sampling from an aggregated pool of nodules from patients with the ground truth of "not lung cancer". To the best of our knowledge, such methods have not yet been proposed for MIL. This oversampling technique resulted in significant improvements on accuracy, PPV, NPV and AUC. We

believe this strategy, which is based on the characteristics of MIL, can be used when training any MIL model from a class imbalanced dataset.

The results show that our method could potentially be applied to automated lung cancer diagnosis, subject to further validation and studies in large datasets. However, we acknowledge there are some limitations and weaknesses in the assumptions we had to make. First, due to the need of a mask that delineates the nodules to calculate radiomic features, our method would have to be dependent on lung nodule detection and segmentation methods such as the ones proposed by Huang et al. [46] and Anirudh et al. [47]. This dependence on pre-existing or human expert segmentation is not new and is problem that still affects many aspects of medical image analysis and supervised machine learning. Related to this is the potential for inter-observer disagreement about the external outline of the nodule. This problem is well known and documented for large and locally advanced lung tumors, but with the small nodule volumes involved in this study, we have assumed that the inter-observer problem does not strongly affect the extracted features. A further question we cannot address in this study is the problem of undetected nodules and very small nodules (diameter smaller than 3mm) that were omitted, moreover, images were resampled to 2 mm isotropic voxels prior to feature extraction, which is possibly also a reason why very small nodules are not appropriate. This work has assumed no false positives and no false negatives, so we cannot elucidate what happens with imperfect nodule detection.

The performance of our model appears sensitive to sampling effects, in other words, the performance of model fluctuates across repeated experiments, as shown in Figure 4-3. This is likely a direct consequence of the relatively

small sample size of the dataset. Expanding the sample size by including small nodules is not immediately helpful because they do not add that many subjects and nodules to the sample, whereas hand-crafted features would not be stable when taken from very small volumes. The major root of the problem appears to be the lack of ground truth and annotated images. Related to this fact is that we currently did not find a suitable dataset for external, independent validation. Therefore, our results should be interpreted as preliminary indication of feasibility, and larger datasets need to be used to demonstrate wider generalizability of this work.

Due to the high fitting ability of neural networks and large epochs during training, the model returns 1 or 0 almost all of the time, which means the overall model calibration was generally poor [48][49]. Model calibration plot is shown in Supplementary Figure 1, and it appears that all MIL methods have poor calibration except MI-SVM.

In addition, we have not explored feature dimensionality reduction and applied feature redundancy analysis. This is in part due in part to the transformation network that does not require explicit feature selection steps prior to MIL pooling. The repeatability and reproducibility of handcrafted features are subjects of numerous investigations in radiomics and appears to be highly modality specific. This work has not explored the stability of low-dose CT-derived image features, which tend to have quite a lot of noise present [50]. This could affect the performance of our model in an external validation, and image harmonization or denoising strategies may be needed in future to support general extensibility.

Moreover, we were not able to test the performance of the models in an external dataset, which would have provided more reliable estimates of the models' potential performance in a different setting. On the other hand, the dataset used in this study (LIDC-IDRI), was collected over 10 years ago. With new emerging CT technologies and reconstruction methods, it is possible that different conclusions would be reached if the proposed method is applied to newer images currently being used in clinical practice. Further research on this aspect is required.

Finally, our oversampling strategy is sensitive to the quality of data's label at patient level. More specifically, if labels are incorrect (e.g. if one or two of the nodules has been misclassified by error and the subject is hence a false negative), the noise will be amplified due to oversampling.

For future work, an automated nodule detection and segmentation algorithm could be attached to this attention-based MIL classifier to fully complete the lung cancer diagnosis workflow. Secondly, methods for improving radiomic features' reliability in low dose CT may be necessary for improving model's performance in unseen data. Thirdly, large scale and comprehensive evaluation of the attention mechanism is needed in the future to assess its reliability and reproducibility. Fourthly, a comparison between the proposed method and a traditional deep learning-based image classification algorithm would be of special interest. Finally, the proposed model needs to be externally validated to assess whether the model suffers from overfitting to the training data or whether it is widely generalizable to CT images from different scanners.

Conclusions

We treated computer-aided diagnosis of lung cancer as a multiple instance learning (MIL) problem, such that the classification as lung cancer or not is made at the subject level (i.e. the patient) without relying on classifications at the level of individual nodules (i.e. each of suspicious lung nodules). The addition of the attention mechanism was used to draw the clinician's eye towards features that were important for triggering the recommended diagnosis, with the aim of supporting interpretability and, importantly, verification by human experts of the algorithm's internal logic. We used radiomics as a source of interpretable image-derived features, and deep attention-based MIL was found to be a superior classifier compared to other MIL options with regard to accuracy, NPV and AUC. A novel approach for minority oversampling, adapted for MIL problems, has been used to address the outcome class imbalance in the LIDC-IDRI dataset. We showed how an attention mechanism could be used as an indication of the importance of each nodule for triggering the diagnostic recommendation. Cross-validation was used to check for model performance, but more data is required to provide a robust test of wider generalizability.

Appendix

Support materials of this Chapter can be found in this link.

References:

[1]Cancer. (2018) [WWW document]. URL https://www.who.int/news-room/fact-sheets/detail/cancer. [access on 20 June 2020]

[2]Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: a cancer journal for clinicians*, *70*(1), 7–30. https://doi.org/10.3322/caac.21590

[3]National Lung Screening Trial Research Team. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, *365*(5), 395-409.

[4]Kanazawa, K., Kawata, Y., Niki, N., Satoh, H., Ohmatsu, H., Kakinuma, R., ... & Eguchi, K. (1998). Computer-aided diagnosis for pulmonary nodules based on helical CT images. *Computerized medical imaging and graphics*, *22*(2), 157-167.

[5]Naqi, S. M., Sharif, M., & Jaffar, A. (2020). Lung nodule detection and classification based on geometric fit in parametric form and deep learning. *Neural Computing and Applications*, *32*(9), 4629-4647.

[6]Parveen, S. S., & Kavitha, C. (2012). A review on computer aided detection and diagnosis of lung cancer nodules. *International Journal of Computers & Technology*, *3*(3a), 393-400.

[7]Yang, Y., Feng, X., Chi, W., Li, Z., Duan, W., Liu, H., ... & Liu, B. (2018). Deep learning aided decision support for pulmonary nodules diagnosing: a review. *Journal of thoracic disease*, *10*(Suppl 7), S867.

[8]Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[9]Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015, June). Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging* (pp. 588-599). Springer, Cham.

[10]Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929).

[11]Xie, Y., Zhang, J., Xia, Y., Fulham, M., & Zhang, Y. (2018). Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. *Information Fusion*, 42, 102-110.

[12]Alakwaa, W., Nassef, M., & Badr, A. (2017). Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). *Lung Cancer*, 8(8), 409.

[13]Setio, A. A. A., Traverso, A., De Bel, T., Berens, M. S., van den Bogaard, C., Cerello, P., ... & van der Gugten, R. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Medical image analysis*, *42*, 1-13.

[14]Black, W. C., Gareen, I. F., Soneji, S. S., Sicks, J. D., Keeler, E. B., Aberle, D. R., ... & Gatsonis, C. (2014). Cost-effectiveness of CT screening in the National Lung Screening Trial. *The New England journal of medicine*, *371*(19), 1793-1802.

[15]Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... & Naidich, D. P. (2019). End-to-end lung cancer screening with

three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, *25*(6), 954-961.

[16]Selbst, A., & Powles, J. (2018, January). "Meaningful Information" and the Right to Explanation. In *Conference on Fairness, Accountability and Transparency* (pp. 48-48). PMLR.

[17]Data protection in the EU. (2020) [WWW document]. URL https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en. [access on 5 July 2020]

[18]Dietterich, T. G., Lathrop, R. H., & Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, *89*(1-2), 31-71.

[19]Maron, O., & Lozano-Pérez, T. (1998). A framework for multiple-instance learning. In *Advances in neural information processing systems* (pp. 570-576).

[20]Bhattacharjee, K., Pant, M., Zhang, Y. D., & Satapathy, S. C. (2020). Multiple Instance Learning with Genetic Pooling for medical data analysis. *Pattern Recognition Letters*, *133*, 247-255.

[21]Ilse, M., Tomczak, J. M., & Welling, M. (2018, January). Attention-based deep multiple instance learning. In *35th International Conference on Machine Learning, ICML 2018* (pp. 3376-3391). International Machine Learning Society (IMLS).

[22]Wang, X., Yan, Y., Tang, P., Bai, X., & Liu, W. (2018). Revisiting multiple instance neural networks. *Pattern Recognition*, *74*, 15-24.

[23]Andrews, S., Tsochantaridis, I., & Hofmann, T. (2003). Support vector machines for multiple-instance learning. In *Advances in neural information processing systems* (pp. 577-584).

[24]Zhou, Z. H., Sun, Y. Y., & Li, Y. F. (2009, June). Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1249-1256).

[25]Wei, X. S., Wu, J., & Zhou, Z. H. (2016). Scalable algorithms for multi-instance learning. *IEEE transactions on neural networks and learning systems*, *28*(4), 975-987.

[26]Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112).

[27]Afshar, P., Mohammadi, A., Plataniotis, K. N., Oikonomou, A., & Benali, H. (2019). From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. *IEEE Signal Processing Magazine*, 36(4), 132-160.

[28]Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., ... & Hoebers, F. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, *5*(1), 1-9.

[29]Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R. G., Granton, P., ... & Aerts, H. J. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, *48*(4), 441-446.

[30]Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2014). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, *11*(3), 812-820.

[31]Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F.,

Meyer, C. R., Reeves, A. P., ... & Kazerooni, E. A. (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.

[32]Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., ... & Tarbox, L. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging*, *26*(6), 1045-1057.

[33]Han, Z., Wei, B., Hong, Y., Li, T., Cong, J., Zhu, X., ... & Zhang, W. (2020). Accurate Screening of COVID-19 using Attention Based Deep 3D Multiple Instance Learning. *IEEE Transactions on Medical Imaging*.

[34]Han, F., Wang, H., Zhang, G., Han, H., Song, B., Li, L., ... & Liang, Z. (2015). Texture feature analysis for computer-aided diagnosis on pulmonary nodules. *Journal of digital imaging*, *28*(1), 99-115.

[35]Dhara, A. K., Mukhopadhyay, S., Dutta, A., Garg, M., & Khandelwal, N. (2016). A combination of shape and texture features for classification of pulmonary nodules in lung CT images. *Journal of digital imaging*, *29*(4), 466-475.

[36]Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., ... & Aerts, H. J. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer research*, *77*(21), e104-e107.

[37]Hancock, M. C., & Magnan, J. F. (2016). Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods. *Journal of*

Medical Imaging, *3*(4), 044504.

[38]Yaniv, Z., Lowekamp, B. C., Johnson, H. J., & Beare, R. (2018). SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. *Journal of digital imaging*, *31*(3), 290-303.

[39]Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

[40]He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284.

[41]Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220-239.

[42]Bottou, L. (2012). Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade* (pp. 421-436). Springer, Berlin, Heidelberg.

[43]Huang, Jin, and Charles X. Ling. "Using AUC and accuracy in evaluating learning algorithms." *IEEE Transactions on knowledge and Data Engineering* 17.3 (2005): 299-310.

[44]Learning and Mining from DatA. (2021) [WWW document]. URL http://210.28.132.67/Data.ashx. [accesses on 4 Jun 2021]

[45]Yong, H., Huang, J., Meng, D., Hua, X., & Zhang, L. (2020, August). Momentum batch normalization for deep learning with small batch size. In *European Conference on Computer Vision* (pp. 224-240). Springer, Cham.

[46]Huang, X., Shan, J., & Vaidya, V. (2017, April). Lung nodule detection in CT using 3D convolutional neural networks. In *2017 IEEE 14th*

International Symposium on Biomedical Imaging (ISBI 2017) (pp. 379-383). IEEE.

[47]Anirudh, R., Thiagarajan, J. J., Bremer, T., & Kim, H. (2016, March). Lung nodule detection using 3D convolutional neural networks trained on weakly labeled data. In *Medical Imaging 2016: Computer-Aided Diagnosis* (Vol. 9785, p. 978532). International Society for Optics and Photonics.

[48]Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ... & Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, *21*(1), 128.

[49]Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal*, *35*(29), 1925-1931.

[50]Bagher-Ebadian, H., Siddiqui, F., Liu, C., Movsas, B., & Chetty, I. J. (2017). On the impact of smoothing and noise on robustness of CT and CBCT radiomics features for patients with head and neck cancers. *Medical physics*, *44*(5), 1755-1770.

[51]Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., ... & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine*, 162(1), W1-W73.

Chapter 5

Generative Models Improve Radiomics Performance in Different Tasks and Different Datasets: An Experimental Study

Junhua Chen, Inigo Bermejo, Andre Dekker, Leonard Wee

Adapted from

Junhua Chen, et al. "Generative models improve radiomics performance in different tasks and different datasets: An experimental study." Physica Medica 98 (2022): 11-17.

DOI: https://doi.org/10.1016/j.ejmp.2022.04.008

Abstract

Purpose: Radiomics is an active area of research focusing on high throughput feature extraction from medical images with a wide array of applications in clinical practice, such as clinical decision support in oncology. However, noise in low dose computed tomography (CT) scans can impair the accurate extraction of radiomic features. In this article, we investigate the possibility of using deep learning generative models to improve the performance of radiomics from low dose CTs.

Methods: We used two datasets of low dose CT scans – NSCLC Radiogenomics and LIDC-IDRI – as test datasets for two tasks – pre-treatment survival prediction and lung cancer diagnosis. We used encoder-decoder networks and conditional generative adversarial networks (CGANs) trained in a previous study as generative models to transform low dose CT images into full dose CT images. Radiomic features extracted from the original and improved CT scans were used to build two classifiers – a support vector machine (SVM) and a deep attention based multiple instance learning model – for survival prediction and lung cancer diagnosis respectively. Finally, we compared the performance of the models derived from the original and improved CT scans.

Results: Denoising with the encoder-decoder network and the CGAN improved the area under the curve (AUC) of survival prediction from 0.52 to 0.57 (p-value<0.01). On the other hand, the encoder-decoder network and the CGAN improved the AUC of lung cancer diagnosis from 0.84 to 0.88 and 0.89 respectively (p-value<0.01). Finally, there are no statistically

significant improvements in AUC using encoder-decoder networks and CGAN (p-value=0.34) when networks trained at 75 and 100 epochs.

Conclusion: Generative models can improve the performance of low dose CT-based radiomics in different tasks. Hence, denoising using generative models seems to be a necessary pre-processing step for calculating radiomic features from low dose CTs.

Keyword: Radiomics; Generative Models; Image Denoising; Comparative Study

Introduction

Recent years have seen a dramatic increase in the applications of artificial intelligence in medical imaging [1]. Radiomics, [2] for example, has been applied to clinical-decision support in oncology in a range of cancers (lung cancers, [3] head and neck cancer, [4] rectal cancer [5]) multiple medical imaging modalities (computed tomography (CT), [4] magnetic resonance imaging (MRI), [6] and positron emission tomography (PET)), [3] and applications, such as deriving prognostic models to measure therapeutic plan efficiency [8][9][10]. Radiomics has also garnered attention in the field of radiotherapy, where it is known as dosiomics [7].

Following the ALARA (As Low As Reasonably Achievable) principle [11], low dose CTs has become popular as the preferred imaging method for screening and monitoring populations at risk [12]. As a tradeoff of low radiation exposure, low dose CTs' image quality is inferior to that of full dose CTs', due to the higher noise levels present in low dose CTs. Radiomics applied to low dose CT has already been shown to improve the accuracy of pulmonary nodules analysis for early detection during lung cancer screening [13][14]. In addition, different studies have shown the potential of radiomics on low dose CT for survival prediction [15][16][17][18]. However, image quality and noise impact the repeatability and reproducibility of radiomic features [19] as well as their robustness [20]. In other words, radiomic features extracted from low dose CTs have lower reliability than the counterparts extracted from full dose CTs. Therefore, prediction models or computer aided diagnosis systems based on radiomic features from low dose CTs will likely be less robust and accurate than those based on radiomic features from full dose CT. Improving the performance of radiomics

calculated from low dose CT in different tasks and datasets is therefore a timely and potentially impactful research topic.

One approach is to denoise low dose CT scans [21] and to recalculate the radiomic features based on the denoised CT. The aim of this article, is to answer the question: should we regard denoising as a preprocessing step for radiomic feature extraction from low dose CT? Image denoising can be regarded as a special case of domain adaptation [22], from low dose CT images to full dose style CT images [23]. Many methods have been proposed to perform this transformation [24][25], but recently deep learning [26] based generative models have garnered special attention and achieved state-of-art results [27][28] [29]. We will use generative models to denoise low dose CT scans and improve the reliability of radiomic features[30][31].

In addition, we will explore whether more reliable radiomic features result in models with better performance using two real applications of radiomics: pre-treatment survival prediction [1] and cancer diagnosis [32][33]. The cancer diagnosis task will be based on [34], in which lung cancer diagnosis was approached as a multiple instance learning (MIL) problem [35] where nodules in each CT scan were regarded as instances. The authors used radiomic features as the input and deep attention based MIL [36] as the MIL problem solver for the sake of interpretability. The authors reported a mean precision of 0.807 with a standard error of the mean (SEM) of 0.069, a recall of 0.870 (SEM 0.061), and an area under the curve (AUC) of 0.842 (SEM 0.074) by using this method.

The most related literature to this article is [37], where the authors trained three generative models – encoder-decoder networks [28], conditional

generative adversarial networks (GANs) [38] and cycle GANs [39] – using full dose CTs and simulated paired high-noise low dose CTs. Finally, they showed that radiomic features extracted from low dose CT scans (low-noise CT and high-noise CT) denoised by the models had improved reproducibility. The main differences between [37] and this article is that: 1) we use pre-trained generative models; 2) we use real (not simulated) low dose CTs; and 3) we focus on the improvement in radiomics-based model performance instead of feature reproducibility.

To the authors' best knowledge, this is the first effort to improve the performance of radiomics-based models from features extracted from low dose CT scans. Source code, Radiomics features, data for statistical analysis and supplementary materials of this article are available online at https://gitlab.com/UM-CDS/low-dose-ct-denoising/-/tree/Experimental_Study.

Methods

Institutional Review Board approval was not applicable for this study, since the primary source of data was an open access collection on The Cancer Imaging Archive (National Institutes of Health) [40] and all patients' personal information had been removed from CT scans. This dataset has been used for this study in accordance with the Creative Commons Attribution-NonCommercial 3.0 Unported (CC BY-NC) conditions. The flowchart in Figure 5-1 summarizes our study methodology.

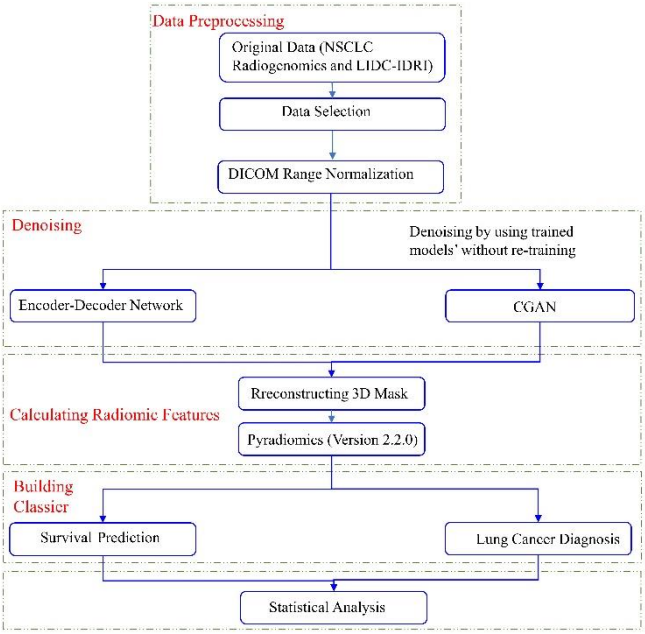


Figure 5-1. Flowchart of methods

Denoising Models' Development

Based on [37], we selected two generative models – encoder-decoder networks and CGANs – that achieved good performance in improving radiomics reproducibility as the experimental models for this study. Moreover, we took the same architecture of encoder-decoder network and CGANs presented in [37].

Training of encoder-decoder networks and CGANs requires paired low dose and full dose versions of the same CT scan. Although there is an open access dataset containing this kind of scans [41], the exposure of low dose CT scans in the dataset is higher - 50 milliampere-seconds (mAs) - than in many low dose CT scanning situations. For example, CT scans in the non-small cell

lung cancer (NSCLC) Radiogenomics dataset were scanned from 1 to 400 mAs [42] and over half CT images scanned with an exposure lower or equal to 5 mAs. Models trained from the dataset described in [41] may have a bad performance in much lower CT scans. The noise power of high noise images (used to train the models) in [37] is 25 times than that in [41]. For this reason, we used trained models from [37] without re-training to denoise low dose CT images. The source code and pre-trained models can be found at https://gitlab.com/UM-CDS/low-dose-ct-denoising/.

Data Acquisition

As mentioned in the Introduction, we will apply pretrained generative models to improve the performance of low CT radiomics-based models in two tasks: pre-treatment survival prediction and lung cancer diagnosis. For this purpose, we chose the NSCLC Radiogenomics dataset [42] for survival prediction and the Lung Image Database Consortium image collection (LIDC-IDRI) for lung cancer diagnosis [43], because they contain the necessary mask of the region of interest (ROI) for calculating the radiomics features and the images were scanned with low radiation exposure.

NSCLC Radiogenomics is a unique radiogenomic dataset from a cohort of 211 patients with NSCLC [44], from which we used low dose CT images, their respective segmentation masks and clinical data for survival prediction. The lung image database consortium and image database resource initiative (LIDC-IDRI) dataset contains 1018 clinical chest CT scans, along with 157 patients' diagnoses. We used the diagnoses and their respective CT scans for the lung cancer diagnosis task. Finally, 106 samples of the NSCLC Radiogenomics were selected for survival prediction and 110 samples from LIDC-IDRI for lung cancer diagnosis. The index of selected samples for

further investigation can be found in Supplementary Tables 1 and 2. The average radiation exposure of selected samples was 38.65 ± 81.97 mAs (±=SEM) in NSCLC Radiogenomics and 145.79 ± 174.57 mAs in LIDC-IDRI. The distributions of radiation exposure for the two datasets are shown in Supplementary Figure 1.

Extraction of Radiomic Features

Before extracting radiomic features from CT images, Hounsfield Unit (HU) value range of CT images were normalized at first. In other words, HU value of pixel in CT images larger than 1000 was set as 1000, and then send the images to extract features.

The masks of the ROIs (tumors) are stored in DICOM format in NSCLC Radiogenomics whilst the segmentation of each nodule is stored in XML file in the LIDC-IDRI dataset. The 3D masks for corresponding ROIs (tumors or nodules) were reconstructed from their corresponding files. We used pyradiomics [45] (version 2.2.0) to calculate 103 radiomic features for further analysis. All features included in the analyses are listed in the Supplementary Table 3.

Radiomics based Models' Development

One of the main tasks in the seminal article on radiomics by Aerts *et al.* [1] is survival prediction. For pre-treatment prediction of survival at 4 years, we used least squares support vector machines (SVMs) [47] with Radial Basis Function (RBF) Kernel as our classifier. SVMs use regularization to prevent overfitting when the number of input variables is high [46]. The input variables for the classifier were age and the 103 radiomic features extracted from the tumor.

For lung cancer diagnosis, we used deep attention-based MIL [36] as the classifier as shown in paper [34]. The main characteristic of this classifier is that it can classify groups of samples (e.g. issue a diagnosis based on a set of CT scans from a patient) and reveal the importance of each sample in determining the diagnosis. The architecture of the method is shown in Supplementary Figure 2. The inputs of the model are the radiomic features and the clinical diagnosis (cancer or not) is the output.

Experiments

We applied the trained generative models to denoise real low dose CT images before extracting the radiomic features. Subsequently, we trained the classification models for survival prediction and lung cancer diagnosis using radiomic features and we compared their performance with that of models trained using radiomic features extracted from low dose CT images.

All denoising experiments for low dose CT images were executed on a Core i7 8565 U CPU with 8GB of RAM based on pre-trained generative models. Based on training specifications described in [37], generative models were trained 25, 50, 75 and 100 epochs. All four trained models were used for denoising. For internal validation, 40 trials of nested cross validation [48] of RBF kernel SVM were executed and the number of GroupKFold in each trial was set as 5 for survival prediction validation. We adopted the minority oversampling strategy described in [49] for lung cancer diagnosis task to improve the model's performance due to our dataset being small and imbalanced.

We assessed the models' performance calculating their area under the receiver-operating characteristics curve (AUC), accuracy and recall (using a

probability threshold of 0.5). Finally, we used Student's *t*-test, after testing the data for normality, to assess the statistical significance of the differences in model performance results.

Results

An example of an original CT image from the NSCLC Radiogenomics dataset and its denoised counterparts are shown in Figure 5-2.

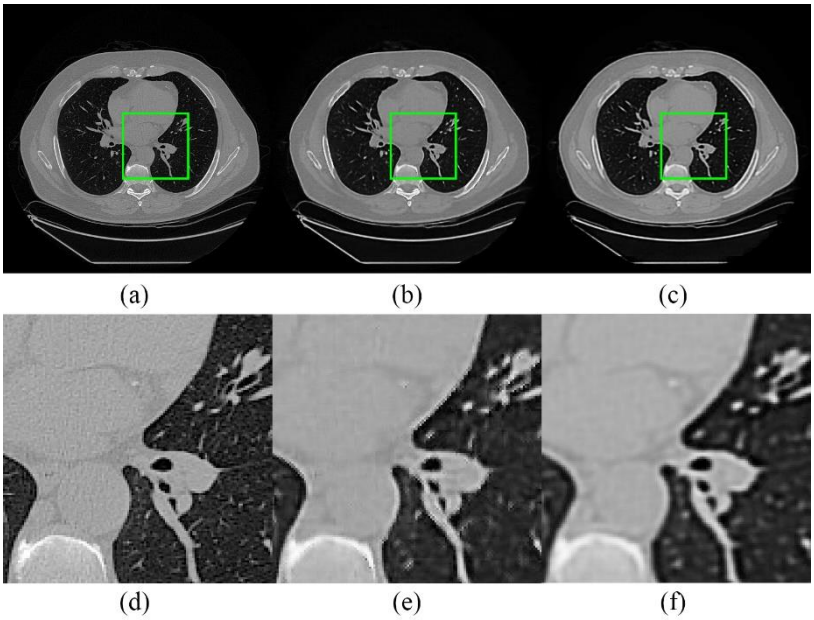


Figure 5-2. Example of low dose CT denoising: (a) original CT Image from NSCLC Radiogenomics (R01-003, radiation exposure: 7 mAs); (b) image denoised by the CGAN (100 epochs); (c) image denoised by the encoder-decoder network (100 epochs); (d) zoomed region of interests (ROI) of (a); (e) zoomed ROI of (b); and (f) zoomed ROI of (c)

Survival Prediction

The 4-year survival prediction model based on radiomic features extracted from low dose CTs achieved an AUC of 0.524 with a standard error of the mean (SEM) of 0.042. On the other hand, the survival prediction models based on radiomic features extracted from denoised low dose CTs achieved AUC ranging between 0.54 and 0.58. As shown in Table 5-1 and Figure 5-3, encoder-decoder networks and CGANs can improve radiomics-based models' performance significantly. The difference between encoder-decoder network and CGAN was not significant when trained for 75 epochs and 100 epochs, similar to what was reported in reference [37].

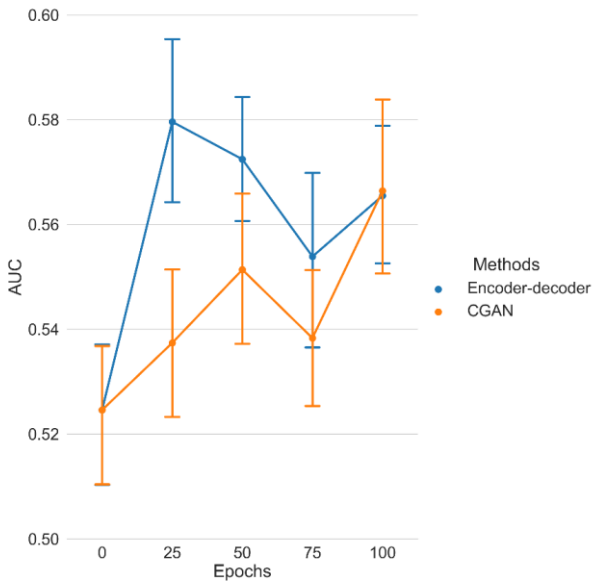


Figure 5-3. Experimental results (AUC) of survival prediction task

Table 5-1. Experimental results for 4-year survival prediction

Metrics \ Training length	Without Denoising	25 Epochs	50 Epochs	75 Epochs	100 Epochs
Encoder-decoder network					
AUC	0.525±0.042	0.580±0.049	0.572±0.040	0.554±0.051	0.566±0.044
p-value *	--	<0.01	<0.01	<0.01	<0.01
CGAN					
AUC	--	0.537±0.045	0.551±0.049	0.538±0.123	0.566±0.053
p-value	--	0.20	<0.01	0.16	<0.01
Encoder-decoder network versus CGAN					
p-value **	--	<0.01	0.04	0.15	0.93

*compared with results from original radiomics; ** comparing encoder-decoder network and CGAN

Lung Cancer Diagnosis

As shown in [34], our method can achieve an AUC of 0.842 (SEM 0.074) based on radiomic features extracted from the original low dose CT scans from the LIDC-IDRI dataset. The AUCs of the classification models based on radiomics extracted from denoised images range between 0.84 and 0.89 as shown in Table 5-2 and Figure 5-4 (c). Models built using radiomic features calculated from denoised images outperformed models developed from the original radiomic features in most experiments. Similarly to survival prediction, the difference between encoder-decoder network and CGAN was not significant when trained for 75 and 100 epochs.

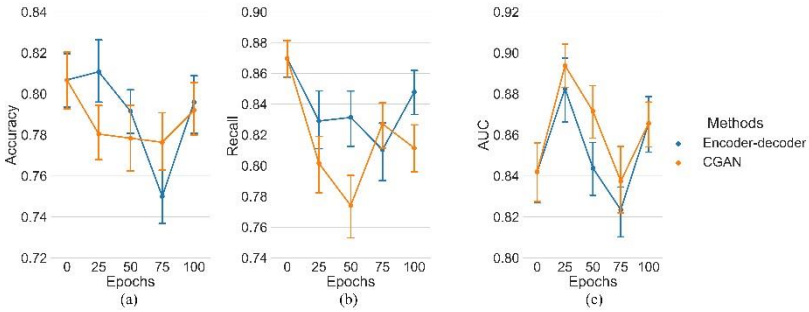


Figure 5-4. Experimental results of lung cancer diagnosis: (a) Accuracy, (b) recall and (c) AUC.

Figure 5-4 (a) and (b) and Table 5-3 show that denoising had a negative impact in the accuracy and recall of the lung cancer diagnosis classification models, when using a threshold of 0.5.

Table 5-2. The AUCs of different models for lung cancer diagnosis

Training length / Metrics	Without Denoising	25 Epochs	50 Epochs	75 Epochs	100 Epochs
Encoder-decoder Network					
AUC	0.84±0.07	0.88±0.08	0.84±0.07	0.82±0.07	0.87±0.07
p-value*	--	<0.01	0.86	0.07	0.02
CGAN					
AUC	--	0.89±0.06	0.86±0.07	0.84±0.09	0.87±0.06
p-value*	--	<0.01	0.06	0.49	0.01
Differences of results by comparing Encoder-decoder network and CGAN					
p-value*	--	0.31	0.07	0.75	0.52

*compared with results from original radiomics;

Table 5-3. Accuracy and recall for lung cancer diagnosis

Training length / Metrics	0 Epochs	25 Epochs	50 Epochs	75 Epochs	100 Epochs
Encoder-decoder network					
Acc	0.81±0.07	0.82±0.08	0.79±0.06	0.75±0.07	0.80±0.07
p-value*	--	0.70	0.10	<0.01	0.26
Recall	0.87±0.06	0.83±0.10	0.83±0.09	0.81±0.10	0.85±0.07
p-value*	--	<0.01	<0.01	<0.01	0.02

CGAN					
Acc	--	0.78±0.07	0.78±0.09	0.78±0.07	0.80±0.06
p-value*	--	0.01	0.01	< 0.01	0.12
Recall	--	0.80±0.09	0.77±0.10	0.83±0.08	0.81±0.08
p-value*	--	< 0.01	< 0.01	< 0.01	< 0.01
Encoder-decoder network versus CGAN (p-values)					
Acc	--	< 0.01	0.21	0.01	0.67
Recall	--	0.04	< 0.01	0.18	< 0.01

*compared with results from original radiomics;

Discussion

In this study, we aimed to assess the potential of generative models to improve the performance of prediction models based on radiomic features extracted from low dose CT scans. The results show that encoder-decoder networks and CGANs can improve the AUC of radiomics for survival prediction and lung cancer diagnosis based on different low dose CT datasets. These findings imply that denoising low dose CT scans using generative models is a convenient pre-processing step before calculating radiomic features to train a predictive or diagnostic model.

The results also show that denoising using generative models might lead to a decrease in accuracy and recall. This might be caused by a shift in the receiver operating characteristic (ROC) curve as a result of the denoising. However, a higher AUC implies that there are other thresholds for which the accuracy and recall are higher with the denoised images. The threshold will

differ for each possible application of these models, and a model with a higher AUC will be more likely to have a better accuracy/recall combination.

Another interesting aspect of the results is the variability of the models' AUCs for different numbers of training epochs. As shown in Figure 5-3 and Figure 5-4 (c), the performance of the models improves after the first epochs, then deteriorates when training for a higher number of epochs, and finally it seems to improve again after a particular number of training epochs. This tendency seems more significant in Figure 5-4 (c) than Figure 5-4. This might be explained by a phenomenon that has attracted considerable attention in the deep learning research domain in last few years -- deep double descent [50][51]. Unfortunately, the mechanisms of this phenomenon are still unclear, and more research on this topic is needed.

It is worth delving into the cause for the observed improvement using generative models. As mentioned previously, we think this improvement is brought on by the denoising effect of generative models to low dose CT. However, as shown in Supplementary Figure 1 (b), 40% CT images in the LIDC-IDRI dataset were not noisy (since they were scanned with over 200 mAs). Denoising these images using generative models would decrease images' quality. Therefore, there must be another source of improvement. Our hypothesis for this alternative source of improvement is dose normalization. In other words, generative models not only improved image quality of low dose CT images in dataset but also transfer the imaging exposure of the whole dataset from a wide range to a more compact but unknown range.

One potential limitation of our study is the low AUCs achieved by the models for pre-treatment survival prediction for lung cancer based on radiomic features. However, these are in line with results reported elsewhere. For example, Isensee *et al.* [52] reported an accuracy of 52.6% based on the BraTS 2017 dataset [53] for brain tumor by using radiomics; Choi *et al.* [54] reported an integrated AUC (iAUC) of 0.620 [95% CI: 0.501–0.756] in TCGA/TCIA dataset using random survival forest to derive a prediction model; Finally, Bae *et al.* [55] reported an iAUC of 0.590 [95% CI: 0.502, 0.689] for overall survival prediction in Glioblastoma using MRI radiomic features. These relatively low AUCs can be partly explained by the difficulty of pre-treatment survival prediction, especially over a long term (over 2 years). In addition to the information available in the medical image, many other factors can affect survival. In fact, some researchers claim that any AUC over 0.80 is suspect [56][57]. As a system of hand-crafted features with higher interpretability but lower information representation ability (compared with deep features), it is not surprising that radiomics has a relatively poor performance in survival prediction. Some studies have proposed techniques to improve the performance of radiomics in survival prediction, such as Jia *et al.* [58], that managed to increase the concordance index (C-index) from 0.6 to 0.67 or Wang *et al.* [59] who combined radiomics with deep features to improve the C-index from 0.68 to 0.72. However, these improved results are still low compared to those achieved in diagnosis, and even further developments might still drive the performance up, the performance in survival will remain relatively low due to inherent uncertainty.

Regarding future work, we believe generative models should be trained to keep more information from the original domain. More specific, low level domain adaptation such as denoising for medical images should focus on keeping content information from original domain in the target domain. For example, by adding a content loss term in the cost function, adjusting generative models training method as shown in [60]. Second, more generative models with different architectures should be considered as the test models to find better models for this task. Thirdly, given the important fluctuations in performance across different numbers of training epochs, it is not possible to provide an optimal number of epochs based on our experiments. For consistency, we reported the results from the model trained model for 100 epochs as our final results. However, more studies about the optimal number of epochs are needed in the future. Finally, since the validity of the results of this study are limited to our selected datasets and tasks, further application to more datasets and tasks could reinforce or disprove our findings.

Conclusions

In this study, we assessed the potential of generative models (CGANs and encoder-decoder networks) to improve the performance of low dose CT scan radiomics-based models in two tasks – survival prediction and lung cancer diagnosis – and two datasets – NSCLC Radiogenomics and LIDC-IDRI. SVM and deep attention based MIL were used classifiers in survival prediction and lung cancer diagnosis respectively. The results support the hypothesis that generative models can improve radiomics performance in different tasks and datasets. In conclusion, denoising using generative models is an effective pre-processing step for calculating radiomic features from low dose CT.

Appendix

Support materials of this Chapter can be found in this link.

References:

[1]Avanzo, Michele, Massimiliano Porzio, Leda Lorenzon, Lisa Milan, Roberto Sghedoni, Giorgio Russo, Raffaella Massafra et al. "Artificial intelligence applications in medical imaging: A review of the medical physics research in Italy." Physica Medica 83 (2021): 221-241. https://doi.org/10.1016/j.ejmp.2021.04.010

[2]Aerts, Hugo JWL, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." Nat. Commun. 5, no. 1 (2014): 1-9. https://doi.org/10.1038/ncomms5006

[3]Desseroit, Marie-Charlotte, Florent Tixier, Wolfgang A. Weber, Barry A. Siegel, Catherine Cheze Le Rest, Dimitris Visvikis, and Mathieu Hatt. "Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non–small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort." J. Nucl. Med. 58, no. 3 (2017): 406-411. https://doi.org/10.2967/jnumed.116.180919

[4]Bogowicz, Marta, O. Riesterer, R. A. Bundschuh, P. Veit-Haibach, M. Hüllner, G. Studer, S. Stieb et al. "Stability of radiomic features in CT perfusion maps." Phys. Med. Biol. 61, no. 24 (2016): 8736. https://doi.org/10.1088/1361-6560/61/24/8736

[5]Tixier, Florent, Mathieu Hatt, Catherine Cheze Le Rest, Adrien Le Pogam, Laurent Corcos, and Dimitris Visvikis. "Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET." J. Nucl. Med. 53, no. 5 (2012): 693-700. https://doi.org/10.2967/jnumed.111.099127

[6]Zhang, Bin, Jie Tian, Di Dong, Dongsheng Gu, Yuhao Dong, Lu Zhang, Zhouyang Lian et al. "Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma." Clin. Cancer Res. 23, no. 15 (2017): 4259-4269. https://doi.org/10.1158/1078-0432.CCR-16-2910

[7]Placidi, Lorenzo, Eliana Gioscio, Cristina Garibaldi, Tiziana Rancati, Annarita Fanizzi, Davide Maestri, Raffaella Massafra et al. "A Multicentre Evaluation of Dosiomics Features Reproducibility, Stability and Sensitivity." Cancers 13, no. 15 (2021): 3835. https://doi.org/10.3390/cancers13153835

[8]Comes, Maria Colomba, Annarita Fanizzi, Samantha Bove, Vittorio Didonna, Sergio Diotaiuti, Daniele La Forgia, Agnese Latorre et al. "Early prediction of neoadjuvant chemotherapy response by exploiting a transfer learning approach on breast DCE-MRIs." Sci. Rep. 11, no. 1 (2021): 1-12. https://doi.org/10.1038/s41598-021-93592-z

[9]Comes, Maria Colomba, Daniele La Forgia, Vittorio Didonna, Annarita Fanizzi, Francesco Giotta, Agnese Latorre, Eugenio Martinelli et al. "Early prediction of breast cancer recurrence for patients treated with neoadjuvant chemotherapy: a transfer learning approach on DCE-MRIs." Cancers 13, no. 10 (2021): 2298. https://doi.org/10.3390/cancers13102298

[10]La Forgia, Daniele, Angela Vestito, Maurilia Lasciarrea, Maria Colomba Comes, Sergio Diotaiuti, Francesco Giotta, Agnese Latorre et al. "Response predictivity to neoadjuvant therapies in breast cancer: A qualitative analysis of background parenchymal enhancement in DCE-MRI." J. Pers. Med 11, no. 4 (2021): 256. https://doi.org/10.3390/jpm11040256

[11]Musolino, Stephen V., Joseph DeFranco, and Richard Schlueck. "The ALARA principle in the context of a radiological or nuclear emergency." Health Phys. 94, no. 2 (2008): 109-111. https://doi.org/10.1097/01.HP.0000285801.87304.3f

[12]Bi, Wenya Linda, Ahmed Hosny, Matthew B. Schabath, Maryellen L. Giger, Nicolai J. Birkbak, Alireza Mehrtash, Tavis Allison et al. "Artificial intelligence in cancer imaging: clinical challenges and applications." Ca-Cancer J. Clin 69, no. 2 (2019): 127-157. https://doi.org/10.3322/caac.21552

[13]Gillies, Robert J., and Matthew B. Schabath. "Radiomics improves cancer screening and early detection." Cancer Epidemiol., Biomarkers Prev. 29, no. 12 (2020): 2556-2567. https://doi.org/10.1158/1055-9965.EPI-20-0075

[14]Choi, Wookjin, Jung Hun Oh, Sadegh Riyahi, Chia-Ju Liu, Feng Jiang, Wengen Chen, Charles White et al. "Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer." Med. Phys. 45, no. 4 (2018): 1537-1549. https://doi.org/10.1002/mp.12820

[15]Homayounieh, Fatemeh, Pingkun Yan, Subba R. Digumarthy, Uwe Kruger, Ge Wang, and Mannudeep K. Kalra. "Prediction of coronary calcification and stenosis: role of radiomics from Low-Dose CT." Academic Radiology 28, no. 7 (2021): 972-979. https://doi.org/10.1016/j.acra.2020.09.021

[16]van Timmeren, Janna E., Ralph TH Leijenaar, Wouter van Elmpt, Bart Reymen, Cary Oberije, René Monshouwer, Johan Bussink, Carsten Brink, Olfred Hansen, and Philippe Lambin. "Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images." Radiother. Oncol. 123, no. 3 (2017): 363-369.

https://doi.org/10.1016/j.radonc.2017.04.016

[17]Nawa, Takeshi, Tohru Nakagawa, Tetsuya Mizoue, Suzushi Kusano, Tatsuya Chonan, Shimao Fukai, and Katsuyuki Endo. "Long-term prognosis of patients with lung cancer detected on low-dose chest computed tomography screening." Lung Cancer 75, no. 2 (2012): 197-202. https://doi.org/10.1016/j.lungcan.2011.07.002

[18]Ayati, Narjess, Sze Ting Lee, S. Rasoul Zakavi, Melissa Cheng, WF Eddie Lau, Sagun Parakh, Kunthi Pathmaraj, and Andrew M. Scott. "Response evaluation and survival prediction after PD-1 immunotherapy in patients with non–small cell lung cancer: comparison of assessment methods." J. Nucl. Med. 62, no. 7 (2021): 926-933. https://doi.org/10.2967/jnumed.120.254508

[19]Traverso, Alberto, Leonard Wee, Andre Dekker, and Robert Gillies. "Repeatability and reproducibility of radiomic features: a systematic review." Int. J. Radiat. Oncol., Biol., Phys. 102, no. 4 (2018): 1143-1158. https://doi.org/10.1016/j.ijrobp.2018.05.053

[20]Bagher-Ebadian, Hassan, Farzan Siddiqui, Chang Liu, Benjamin Movsas, and Indrin J. Chetty. "On the impact of smoothing and noise on robustness of CT and CBCT radiomics features for patients with head and neck cancers." Med. Phys. 44, no. 5 (2017): 1755-1770. https://doi.org/10.1002/mp.12188

[21]Kelm, Zachary S., Daniel Blezek, Brian Bartholmai, and Bradley J. Erickson. "Optimizing non-local means for denoising low dose CT." In 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 662-665. IEEE, 2009. https://doi.org/10.1109/ISBI.2009.5193134

[22]Chen, Minmin, Zhixiang Xu, Kilian Weinberger, and Fei Sha.

"Marginalized denoising autoencoders for domain adaptation." arXiv preprint arXiv:1206.4683 (2012).

[23]Yang, Qingsong, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K. Kalra, Yi Zhang, Ling Sun, and Ge Wang. "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss." IEEE Trans. Med. Imaging 37, no. 6 (2018): 1348-1357. https://doi.org/10.1109/TMI.2018.2827462

[24]Sharma, Abhishek, and Vijayshri Chaurasia. "A review on magnetic resonance images denoising techniques." In Machine Intelligence and Signal Analysis, pp. 707-715. Springer, Singapore, 2019. https://doi.org/10.1007/978-981-13-0923-6_60

[25]Kollem, Sreedhar, Katta Rama Linga Reddy, and Duggirala Srinivasa Rao. "A review of image denoising and segmentation methods based on medical images." Int. J. Mach. Learn. Comput. 9, no. 3 (2019): 288-295. https://doi.org/10.18178/ijmlc.2019.9.3.800

[26]LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521, no. 7553 (2015): 436-444. https://doi.org/10.1038/nature14539

[27]Shan, Hongming, Yi Zhang, Qingsong Yang, Uwe Kruger, Mannudeep K. Kalra, Ling Sun, Wenxiang Cong, and Ge Wang. "3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network." IEEE Trans. Med. Imaging 37, no. 6 (2018): 1522-1534. https://doi.org/10.1109/TMI.2018.2832217

[28]Chen, Hu, Yi Zhang, Mannudeep K. Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. "Low-dose CT with a residual encoder-decoder convolutional neural network." IEEE Trans. Med. Imaging 36,

no. 12 (2017): 2524-2535. https://doi.org/ 10.1109/TMI.2017.2715284

[29]Kang, Eunhee, Won Chang, Jaejun Yoo, and Jong Chul Ye. "Deep convolutional framelet denosing for low-dose CT via wavelet residual network." IEEE Trans. Med. Imaging 37, no. 6 (2018): 1358-1369. https://doi.org/10.1109/TMI.2018.2823756

[30]Lucia, François, Dimitris Visvikis, Marie-Charlotte Desseroit, Omar Miranda, Jean-Pierre Malhaire, Philippe Robin, Olivier Pradier, Mathieu Hatt, and Ulrike Schick. "Prediction of outcome using pretreatment 18F-FDG PET/CT and MRI radiomics in locally advanced cervical cancer treated with chemoradiotherapy." Eur. J. Nucl. Med. Mol. Imaging 45, no. 5 (2018): 768-786. https://doi.org/10.1007/s00259-017-3898-7

[31]Parmar, Chintan, Ralph TH Leijenaar, Patrick Grossmann, Emmanuel Rios Velazquez, Johan Bussink, Derek Rietveld, Michelle M. Rietbergen, Benjamin Haibe-Kains, Philippe Lambin, and Hugo JWL Aerts. "Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer." Sci. Rep. 5, no. 1 (2015): 1-10. https://doi.org/10.1038/srep11044

[32]Chen, Chia-Hung, Chih-Kun Chang, Chih-Yen Tu, Wei-Chih Liao, Bing-Ru Wu, Kuei-Ting Chou, Yu-Rou Chiou, Shih-Neng Yang, Geoffrey Zhang, and Tzung-Chi Huang. "Radiomic features analysis in computed tomography images of lung nodule classification." PloS one 13, no. 2 (2018): e0192002. https://doi.org/10.1371/journal.pone.0192002

[33]Li, Hui, Yitan Zhu, Elizabeth S. Burnside, Erich Huang, Karen Drukker, Katherine A. Hoadley, Cheng Fan et al. "Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in

the TCGA/TCIA data set." NPJ breast cancer 2, no. 1 (2016): 1-10. https://doi.org/10.1038/npjbcancer.2016.12

[34]Chen, Junhua, Zeng, Haiyan, Zhang, Cong, et al. “Lung cancer diagnosis using deep attention based multiple instance learning and radiomics.” Med. Phys.. 2022; 00: 00- 00. https://doi.org/10.1002/mp.15539

[35]Maron, Oded, and Tomás Lozano-Pérez. "A framework for multiple-instance learning." Advances in neural information processing systems 10 (1997).

[36]Ilse, Maximilian, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning." In International conference on machine learning, pp. 2127-2136. PMLR, 2018.

[37]Chen, Junhua, Chong Zhang, Alberto Traverso, Ivan Zhovannik, Andre Dekker, Leonard Wee, and Inigo Bermejo. "Generative models improve radiomics reproducibility in low dose CTs: a simulation study." Phys. Med. Biol. 66, no. 16 (2021): 165002. https://doi.org/10.1088/1361-6560/ac16c0

[38]Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. "Image-to-image translation with conditional adversarial networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125-1134. 2017.

[39]Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In Proceedings of the IEEE international conference on computer vision, pp. 2223-2232. 2017.

[40]Clark, Kenneth, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore et al. "The Cancer Imaging Archive

(TCIA): maintaining and operating a public information repository." J. Digit. Imaging 26, no. 6 (2013): 1045-1057. https://doi.org/10.1007/s10278-013-9622-7

[41]McCollough, Cynthia H., Adam C. Bartley, Rickey E. Carter, Baiyu Chen, Tammy A. Drees, Phillip Edwards, David R. Holmes III et al. "Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge." Med. Phys. 44, no. 10 (2017): e339-e352. https://doi.org/10.1002/mp.12345

[42]Bakr, Shaimaa, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng et al. "A radiogenomic dataset of non-small cell lung cancer." Sci. Data 5, no. 1 (2018): 1-9. https://doi.org/10.1038/sdata.2018.202

[43]Armato III, Samuel G., Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao et al. "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans." Med. Phys. 38, no. 2 (2011): 915-931. https://doi.org/10.1118/1.3528204

[44]Mazurowski, Maciej A. "Radiogenomics: what it is and why it is important." J. Am. Coll. Radiol. 12, no. 8 (2015): 862-866. https://doi.org/10.1016/j.jacr.2015.04.019

[45]Van Griethuysen, Joost JM, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. "Computational radiomics system to decode the radiographic phenotype." Cancer Res. 77, no. 21 (2017): e104-e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

[46]Xu, Huan, Constantine Caramanis, and Shie Mannor. "Robustness and Regularization of Support Vector Machines." J. Mach. Learn. Res.10, no. 7 (2009).

[47]Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." Neural Process. Lett. 9, no. 3 (1999): 293-300.

[48]Cawley, Gavin C., and Nicola LC Talbot. "On over-fitting in model selection and subsequent selection bias in performance evaluation." J. Mach. Learn. Res. 11 (2010): 2079-2107.

[49]Zhu, Tuanfei, Yaping Lin, Yonghe Liu, Wei Zhang, and Jianming Zhang. "Minority oversampling for imbalanced ordinal regression." Knowledge-Based Syst. 166 (2019): 140-155. https://doi.org/10.1016/j.knosys.2018.12.021

[50]Nakkiran, Preetum, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. "Deep double descent: Where bigger models and more data hurt." J. Stat. Mech.: Theory Exp 2021, no. 12 (2021): 124003. https://doi.org/10.1088/1742-5468/ac3a74

[51]d'Ascoli, Stéphane, Maria Refinetti, Giulio Biroli, and Florent Krzakala. "Double trouble in double descent: Bias and variance (s) in the lazy regime." In International Conference on Machine Learning, pp. 2280-2290. PMLR, 2020.

[52]Isensee, Fabian, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. "Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge." In International MICCAI Brainlesion Workshop, pp. 287-297. Springer, Cham, 2017. https://doi.org/10.1007/978-3-319-75238-9_25

[53]Menze, Bjoern H., Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren et al. "The multimodal brain tumor image segmentation benchmark (BRATS)." IEEE Trans. Med. Imaging 34, no. 10 (2014): 1993-2024. . https://doi.org/10.1109/TMI.2014.2377694

[54]Choi, Yoon Seong, Sung Soo Ahn, Jong Hee Chang, Seok-Gu Kang, Eui Hyun Kim, Se Hoon Kim, Rajan Jain, and Seung-Koo Lee. "Machine learning and radiomic phenotyping of lower grade gliomas: improving survival prediction." Eur. Radiol. 30, no. 7 (2020): 3834-3842. https://doi.org/10.1007/s00330-020-06737-5

[55]Bae, Sohi, Yoon Seong Choi, Sung Soo Ahn, Jong Hee Chang, Seok-Gu Kang, Eui Hyun Kim, Se Hoon Kim, and Seung-Koo Lee. "Radiomic MRI phenotyping of glioblastoma: improving survival prediction." Radiology 289, no. 3 (2018): 797-806. https://doi.org/10.1148/radiol.2018180200

[56]Bahn, Emanuel, and Markus Alber. "On the limitations of the area under the ROC curve for NTCP modelling." Radiother. Oncol. 144 (2020): 148-151. https://doi.org/10.1016/j.radonc.2019.11.018

[57]Cook, Nancy R. "Use and misuse of the receiver operating characteristic curve in risk prediction." Circulation 115, no. 7 (2007): 928-935. https://doi.org/10.1161/CIRCULATIONAHA.106.672402

[58]Wu, Jia, Chao Li, Michael Gensheimer, Sukhmani Padda, Fumi Kato, Hiroki Shirato, Yiran Wei et al. "Radiological tumour classification across imaging modality and histology." Nat. Mach. Intell. 3, no. 9 (2021): 787-798. https://doi.org/10.1038/s42256-021-00377-0

[59]Wang, Yutao, Qian Shao, Shuying Luo, and Randi Fu. "Development of a nomograph integrating radiomics and deep features based on MRI

to predict the prognosis of high grade Gliomas." Math. Biosci. Eng. 18, no. 6 (2021): 8084-8095. https://doi.org/10.3934/mbe.2021401

[60]Yang, Heran, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Jerry L. Prince, and Zongben Xu. "Unsupervised MR-to-CT synthesis using structure-constrained CycleGAN." IEEE Trans. Med. Imaging 39, no. 12 (2020): 4249-4261. https://doi.org/ 10.1109/TMI.2020.3015379

Chapter 6
Improving Reproducibility and Performance of Radiomics in Low Dose CT using Cycle GANs

Junhua Chen, Leonard Wee, Andre Dekker, Inigo Bermejo

Adapted from

Junhua Chen et al. Improving reproducibility and performance of radiomics in low-dose CT using cycle GANs. Journal Of Applied Clinical Medical Physics. 2022;23(8):e13739.

DOI: https://doi.org/10.1002/acm2.13739

Abstract

Background: As a means to extract biomarkers from medical imaging, radiomics has attracted increased attention from researchers. However, reproducibility and performance of radiomics in low dose CT scans are still poor, mostly due to noise. Deep learning generative models can be used to denoise these images and in turn improve radiomics' reproducibility and performance. However, most generative models are trained on paired data, which can be difficult or impossible to collect.

Purpose: In this article, we investigate the possibility of denoising low dose CTs using cycle generative adversarial networks (GANs) to improve radiomics reproducibility and performance based on unpaired datasets.

Methods and Materials: Two cycle GANs were trained: 1) from paired data, by simulating low dose CTs (i.e., introducing noise) from high dose CTs; and 2) from unpaired real low dose CTs. To accelerate convergence, during GAN training, a slice-paired training strategy was introduced. The trained GANs were applied to three scenarios: 1) improving radiomics reproducibility in simulated low dose CT images and 2) same-day repeat low dose CTs (RIDER dataset) and 3) improving radiomics performance in survival prediction. Cycle GAN results were compared with a conditional GAN (CGAN) and an encoder-decoder network (EDN) trained on simulated paired data.

Results: The cycle GAN trained on simulated data improved concordance correlation coefficients (CCC) of radiomic features from 0.87 [95%CI, (0.833,0.901)] to 0.93 [95%CI, (0.916,0.949)] on simulated noise CT and from 0.89 [95%CI, (0.881,0.914)] to 0.92 [95%CI, (0.908,0.937)] on the

RIDER dataset, as well improving the area under the receiver operating characteristic curve (AUC) of survival prediction from 0.52 [95%CI, (0.511,0.538)] to 0.59 [95%CI, (0.578,0.602)]. The cycle GAN trained on real data increased the CCCs of features in RIDER to 0.95 [95%CI, (0.933,0.961)] and the AUC of survival prediction to 0.58 [95%CI, (0.576,0.596)].

Conclusion: The results show that cycle GANs trained on both simulated and real data can improve radiomics' reproducibility and performance in low dose CT and achieve similar results compared to CGANs and EDNs.

Keyword: Radiomics, Denoising, Reproducibility, Cycle GAN, Computed Tomography

Introduction

Biomarkers from medical imaging can provide a macroscopic view of the tissue of interest and can be an effective tool to accurately diagnose disease in precision medicine [1]. Radiomics features [7] have shown value as potential imaging biomarkers in various tumor and neurodegenerative diseases, such as lung cancer [2], head and neck cancer [3], rectal cancer [4], breast cancer [6], Alzheimer disease [7], autism spectrum disorder [8].

However, in Computed Tomography (CT) the repeatability and reproducibility of radiomics has been challenged in multiple published studies [9][10][11][12]. The reproducibility of radiomics can be impacted by various CT parameters such as radiation dose, slice thicknesses, and reconstruction algorithm settings. More specifically, it has been reported that only 11.3% (12 of 106) of radiomics features are robust to these technical parameters [12]. In fact, slice thickness ranks first on impact on radiomics' reproducibility while signal-to-noise ratio ranks second. Intensity and texture radiomic features are especially sensitive to radiation dose and the associated signal to noise ratio [12]. Therefore, it is likely that radiomic features extracted from low dose CT are less accurate than features from high dose CT. In other words, radiomics applied to low dose CT will likely have low reliability and thus the established radiomics signature or models are likely to have worse performance compared to high dose CT [37].

In this study, we aim to use denoising [14] to improve the reliability of radiomics in low dose CT. A variety of image denoising methods have been proposed in the past several decades, and these methods can be divided into two classes -- model based denoisers [15][16] and data driven denoisers

[17][18]. Multiple published studies [18][19] have demonstrated that data driven denoisers outperform model based denoisers and achieve state-of-art denoising quality when suitable training datasets are available.

Most data driven denoisers are based on deep convolutional neural networks (DCNNs) [20] in which this denoising task is posed as an image-to-image translation problem. The popular architectures for medical image denoising are full convolutional network (FCN) [21], encoder-decoder network (EDN) [22] and generative adversarial networks (GAN) [23] which were described in detail recently reviews [14][24]. An important characteristic of most data driven denoisers is that datasets consisting of paired low-high dose CTs from the same subjects are needed to train the deep neural networks. However, collecting paired low-high dose CT is time-consuming, expensive, and impossible in many cases e.g., in patient studies.

Therefore, it is the aim of this study to establish a CT denoiser based on unpaired datasets to improve radiomics performance. The related literature is divided into two topics -- low dose CT denoising and radiomics normalization. In this section, we review these two topics briefly.

a) Low Dose CT Denoising

As mentioned above, most data-driven denoisers are based on one of three backbones – FCN, EDN and GAN – and all of them are used in low dose CT denoising tasks. More specifically, Yang et al. [28] used a 3D residual network as the denoising network architecture with a loss function based on differences between the ground truth residual image and reconstructed residual image. Moreover, pool layers were removed from the network to generate denoised residual images because there is no size or resolution

change between input and output. The results show that the network can reduce noise effectively while preserving tissue details. Chen et al [29] adapted an EDN as the backbone of their denoiser and two residual shortcuts were added into the network to keep details of the image from encoder to decoder. Models were trained by using simulation data and the trained denoiser achieved a competitive performance in both simulation and clinical cases. Yang et al. [23] took conditional GAN (CGAN) [31] as the backbone where they replaced Jensen–Shannon divergence [32] with Wasserstein distance [33] to measure the differences in the data distribution. Moreover, Yang et al. replaced the mean squared error (MSE) loss function with Perceptual Loss [34] to keep more texture information from low dose CT to high dose CT. They proposed a method to not only reduce the image noise level but also tried to keep the critical information at the same time.

One of the biggest shortcomings of these aforementioned denoisers is that paired low-high dose datasets are needed in denoiser training. However, collecting this kind of datasets is time-consuming and expensive. As an alternative a few simulation paired low-high dose CT datasets are publicly available, such as the dataset from 2016 NIH-AAPM-Mayo Clinic Low Dose CT Grand Challenge (LDGC) [35]. The low dose CT images in this dataset are simulation data with a simulated low radiation dose of 50 milliampere-seconds (mAs). The characteristics of LDGC dataset decrease the value for network training as the generalization of models trained from the LDGC to real low dose CT is questionable because the exposure in real low dose CT datasets will much lower than the simulation data in LDGC. For example, radiation dose in The Reference Image Database to Evaluate Therapy Response (RIDER) [36] ranged from 7 to 13 mAs.

Therefore, we believe that implementing a denoiser based on unpaired datasets could help to relieve the problem of data collection and make unsupervised CT denoising for quantitative medical image analysis possible. There are a few studies that used this strategy, Kang et al. [37] used cycle GAN as the backbone for multiphase coronary CT angiography correction where they took routine-dose CT from multiphase coronary CT angiography as the target domain data and low-dose CT as the original domain data to build a training dataset. The results show that visual grading and quality evaluation of low-dose CT are improved, however, they did not investigate the effect of Cycle GAN into deeper quantitative metrics such as radiomics.

However, to the best of our knowledge there are no studies that apply unsupervised CT denoising to improve radiomics reliability and reproducibility in low dose CT.

b) Radiomics Normalization

Berenguer et al. [10] have shown that over half of radiomics features are nonreproducible when images scanned from different scanners even when using the same CT parameters. The results of radiomics signatures or models which based on nonreproducible features are thus unreliable. Li et al. [25] used cycle GAN to normalize CT images from multiple centres and multiple scanners, and then they extracted features from normalized images and established radiomics signatures. They found the average improvement of a classifier based on normalized radiomics features in the area under the receiver operating characteristic curve (AUC) to be 11%. Yang et al. [38] integrated adaptive instance normalization (AdaIN) into cycle GAN for continuous CT kernel conversion, introduced AdaIN kept more content

information from original domain to target domain. The proposed method is promising for radiomics normalization in different CT kernels. The major difference between previous studies and our study is that this paper focused on using cycle GAN to improve radiomics reproducible and performance in low dose CTs.

In previous work [37], we used EDN and CGAN [39] as testing backbones to denoise low-dose CT. Our training datasets consisted of paired simulated low-dose CT and high-dose CTs. Radiomics features reproducibility from noisy images and denoised images were measured using concordance correlation coefficients (CCC) [43]. The results showed that EDN and CGAN can improve CCC of noisy images significantly. Moreover, when we applied our trained denoisers to real low-dose CT images (RIDER dataset), the results showed that this denoiser can improve radiomics reproducibility in realistic low-dose CTs.

In another study [26], we applied the trained denoisers to improve radiomics performance in realistic applications. The results showed that generative models based denoisers can improve the AUC of a lung cancer survival prediction from 0.52 [95%CI, (0.511,0.538)] to 0.58 [95%CI, (0.564,0.596)] and a multiple instance learning based lung cancer diagnostic [41] from 0.84 [95%CI, (0.828,0.856)] to 0.88 [95%CI, (0.866,0.892)].

The major shortcoming of our previous studies is that denoising models were exclusively dependent on paired simulation data which may cause the trained denoiser to not generalize well to real data. In this paper, we took cycle GAN as basic denoising model to train a denoiser using unpaired low-high dose CT. These low and high dose CT images were collected from

different centres and scanners. We evaluated this new denoiser for its ability to improve radiomics reproducibility and performance in realistic applications.

In comparison with previous studies, the major contribution of this study is that we assess the potential of denoising low dose CTs using cycle GANs based on unpaired data to improve radiomics reproducibility and performance. The results show that cycle GANs can improve radiomics' reproducibility and performance in low dose CT and achieve similar results compared to CGANs and encoder-decoder networks. Source code, Radiomics features, data for statistical analysis and supplementary materials of this article are available online at https://gitlab.com/UM-CDS/low-dose-ct-denoising/-/tree/Cycle_GAN_Improve_Radiomics.

Materials and Methods

In this section, we describe the architecture and technical details of our cycle GAN. Then, we introduce our training strategy to improve the speed of convergence. Next, we describe the design of the experiments and datasets used for training and testing. Finally, we describe the extraction of the radiomics features and the evaluation metrics used.

Cycle GAN

We use cycle-consistent GANs, proposed by Zhu et al. [27]. As shown in Figure 6-1 (a), the cycle GAN consist of two generators and two discriminators. The generator G_{LH} maps from low dose CT domain (L) to full dose CT domain (H) while G_{HL}maps from H to L. The loss function of the cycle GAN consists of two parts -- adversarial loss and cycle consistency loss, represented with L_{adv} and L_{cyc} respectively (and each of them can be

broken down into L_{adv1}, L_{adv2} and L_{cyc1}, L_{cyc2}, one for each generator). The adversarial loss for mapping from low dose to full dose CT is defined as follows:

$$\mathcal{L}_{adv1}(G_{LH}, D_H, L, H) = \mathbb{E}_{x_h \sim pdata(x_h)}[logD_H(x_h)] + \mathbb{E}_{x_l \sim pdata(x_l)}[log(1 - D_H(G_{LH}(x_l))]$$

(Equation 6-1)

where G_{LH} is trained to transform low dose CT image x_l to into high dose CT image x_h (denoising), while D_H is trained to discriminate between denoised CT images $G_{LH}(x_l)$ (x_{LH} in Figure 6-1 (a)) and real high dose CT image x_H. During the training, G aims to minimize this loss function against an adversary D that tries to maximize it; therefore, equation (1) can be rewritten as follows:

$$min_G max_D \mathcal{L}_{adv1}(G_{LH}, D_H, L, H) = \mathbb{E}_{x_h \sim pdata(x_h)}[logD_H(x_h)] + \mathbb{E}_{x_l \sim pdata(x_l)}[log(1 - D_H(G_{LH}(x_l))]$$

(Equation 6-2)

The definition of adversarial loss for mapping from high dose CT to low dose CT is defined in similar way and we denote it as $min_G max_D \mathcal{L}_{adv2}(G_{HL}, D_L, H, L)$. Moreover, we denote the adversarial loss for the whole network as $\mathcal{L}_{adv}(G, D) = L_{adv1} + L_{adv2}$.

Regarding the cycle consistency loss of our cycle GAN, we replace the mean squared error (MSE) loss function used in the original cycle GAN with a perceptual loss-based loss function. The definition of cycle consistency loss is as follows:

$$\mathcal{L}_{cyc1} = \mathbb{E}(x_l, x_{lhl}) \left[\frac{1}{wed} \left\| VGG\left(G_{HL}(x_{lh})\right) - VGG(x_l) \right\|^2\right]$$ (Equation 6-3)

Where x_l represents low dose CT image and x_{lhl} represents reconstructed low dose CT image from fake synthetic high dose CT image, w, e, and d represent width, height, and depth of the feature map, and $VGG(.)$ represents feature maps from a pre-trained VGG-16 at a specific convolutional layer. VGG-16 is pre-trained on ImageNet [42], a dataset of over 14 million images belonging to 1000 classes. In order to feed CT images into a model pre-trained on color images, they need to be triplicated into RGB channels before cycle consistency loss calculation. In our implementation, we select feature maps from $conv2\_1$ to calculate perceptual loss. $\mathcal{L}_{cyc2}$ can be defined in similar way with G_{LH}. We denote $\mathcal{L}_{cyc1} + \mathcal{L}_{cyc2}$ as $\mathcal{L}_{cyc}(G)$.

Combining Equation 6-2 and Equation 6-3, the overall loss function is expressed as:

$$min_G max_D \mathcal{L}_{adv}(G, D) + \lambda \mathcal{L}_{cyc}(G) \qquad \text{(Equation 6-4)}$$

where λ is a parameter to control the trade-off between the adversarial and perceptual loss.

More details about the architecture of generators and discriminators can be found in Figure 6-1(b) and (c) respectively.

Slice-paired Training Strategy

Randomly chosen samples from two domains are fed to the networks in training a cycle GAN. However, as mentioned in the original cycle GAN article [27], the training will be more successful and stable when focusing on pairs of visually similar images.

In the case of CT scans, assuming all scans belong to the same organ (the lung in our case), we can expect that images belonging to the same slice number will be more similar to each other than images from different slices. Hence, the first slice of a low dose CT scan will have higher similarity with the first slice of a high dose CT scan.

Therefore, CT based cycle GAN training should be fed with pairs of the same (randomly chosen) slice rather than images of different slices. This could be seen as weakly supervised learning. We call this strategy as slice-paired training strategy hereafter, the similar training strategy can be found in paper [43].

Data Acquisition

In order to compare results of cycle GANs with our previous work (CGAN and EDN) [37][26], we trained networks on the same data as used in [37][26] and applied the trained models to the same applications on the same datasets. In total, we used five datasets in this study.

We used a phantom dataset to test whether our GANs generated artifacts when denoising. [44] This phantom dataset is a collection of phantoms CTs by scanning a Gammex 467 CT phantom (Middleton, WI, USA) using a Philips Brilliance Big Bore CT with different doses (50 mAs, 400 mAs). CT images scanned at 50 mAs are referred to as low dose CT and 400 mAs referenced as high dose CT. We used 52 paired images from two scans for testing.

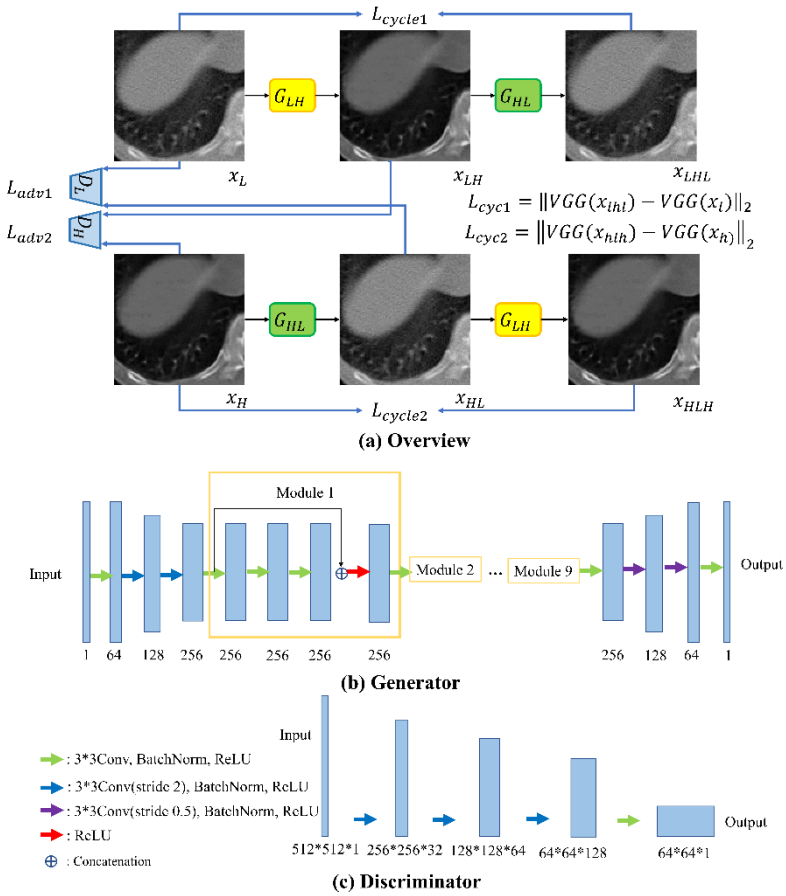


Figure 6-1. Overview of Network, Architecture of Generator and Discriminator

The second is based on the NSCLC-Radiomics dataset (hereafter called LUNG 1) [45]. We selected only the high dose CT scans, those scanned at 400 mAs or more (n=157, indices in Supplementary Table 1) and added noise to the sinograms to simulate low dose CTs with two different levels of noise: low-noise CT and high-noise CT. The specific methods used to add noise are described in [37] section 2.3 and in the Supplementary Method 1. We used a subset of these high-noise CTs and their corresponding high dose

CTs (40 subjects, 4260 images) to train a cycle GAN and we used the remaining images to assess the reproducibility of radiomics features in the original high dose CT versus those in the denoised images.

The third and fourth datasets were used to train the cycle GAN with real low dose CT scans. We used low dose CT scans from the Lung Image Database Consortium dataset (LIDC-IDRI) [47], and high dose CT scans from The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) dataset [48]. We used two inclusion criteria for CTs in both datasets to increase the visual similarity across the two domains: the use of SIEMENS scanner; table height ranging from 150 to 160 mm. As low dose CTs we included those with a radiation exposure lower than 10 mAs and as high dose CTs those with and exposure higher than 100 mAs (list of indices of selected samples is in Supplementary Tables 2 and 3 respectively). Examples of selected samples from LIDC-IDRI and TCGA-LUAD are shown in Supplementary Figure 1.

The final two datasets, used for the two radiomics-based applications, are RIDER [36] and NSCLC Radiogenomics [42]. RIDER is a collection of same day repeat CT scans collected to assess the variability of tumor measurements, which makes it particularly useful to assess the reproducibility of radiomics across pairs of similar CT scans. We use the trained cycle GAN to denoise the images in RIDER to assess the impact of denoising on the reproducibility of radiomic features. NSCLC Radiogenomics is a radiogenomic dataset from a cohort of 211 patients with non-small cell lung cancer [42], from which we selected the low dose CT images, their respective segmentation masks and clinical data for survival prediction (n=106), the indices of the included samples are included in the supplementary Table 4. The average radiation exposure of samples selected

from NSCLC Radiogenomics is 38.65±81.97 mAs (±standard error of the mean, SEM) (the distribution of radiation exposure for selected samples can be found in Supplementary Figure 2).

A summary of scanning parameters of included datasets is shown in Table 6-1.

Experiments

We trained three cycle GANs to denoise low CT scans: on a paired dataset with low dose CT scans simulated from high dose CT scans with and without the Slice-paired training strategy (referred to as ablation study hereafter) and on unpaired real low and high dose CT scans. Regarding CT normalization, the CT HU was set to -1000 when it was lower than -1000 and to 1000 when it was higher than 1000, and then normalized to intensity [0,1] for network training and image denoising.

Then, we assessed the performance of the denoising using Root Mean Square Error (RMSE) and perceptual loss as evaluation metrics. The definition of perceptual loss can be found in equation (3) and definition of RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2} \qquad \text{Equation 6-5}$$

Where y_i and $\hat{y}_i$ represent the image value in position i for the original high dose CT and denoised CT, respectively. Image values were normalized to 0-1 before calculating RMSE. M represents the number of pixels in one image, 512*512 in our case.

We also assessed the impact of denoising on reproducibility of radiomic features by calculating the concordance correlation coefficients (CCC) -- a

metric that measures the degree of agreement between two variables (e.g., to evaluate reproducibility or for inter-rater reliability) as defined in [40]. Several arguments support our choice of CCC as the reproducibility metric: according to a recent systematic review [52], CCC is the most common metric used to measure the reproducibility of radiomics. Moreover, the seminal article that introduced the CCC [40] has shown the clear advantages of using this metric in testing reproducibility in comparison with other methods. On the simulated paired data, we calculated the CCCs of the radiomic features extracted in the original high dose CT and the denoised CT. In RIDER, we calculated the CCC of the same day denoised CT scans.

In the ablation study, we assessed the impact of using the position-based training strategy comparing the performance in terms of RMSE, perceptual loss and CCC on synthetic data.

Next, we applied the trained cycle GAN to two applications -- radiomics reproducibility in same-day repeat CT scans and pre-treatment survival prediction – without retraining. Pre-treatment survival prediction of cancer patients is a typical application of radiomics since it appeared in the seminal article by Aerts et al.[2]. We predicted pre-treatment survival in two different ways: as a binary outcome on 4-year survival and as time-to-event continuous outcome. For the first, we used least squares support vector machines (SVMs) with Radial Basis Function (RBF) Kernel as our classifier. For hyperparameter search and internal validation, we used 40-repeat nested 5-fold cross validation. [54] More details on the survival prediction modelling can be found in [26]. The metric used for measuring the performance this model is the area under the receiver operating characteristic curve (AUC) [49]. For the time-to-event survival analysis we

fitted a Cox proportional hazards models, using the radiomics features (103 features) as predictors. To ensure convergence during parameter fitting, we used penalized Cox regression with a penalty coefficient of 0.01. The discriminative performance of this model was measured using the concordance index (C-index).

All experiments were implemented in Python 3.6 and TensorFlow 1.13.1. The training was run on one Nvidia Tesla V100 GPU 30.5GB of memory and 4 CPUs. We set λ in equation (4) to 10 and the batch size to 1. The discriminator and the denoiser both used the Adam optimizer [50] and shared the same learning rate. The initial learning rate was set to 0.0002 with a decay factor of 0.8 every 10 epochs. Training runs were stopped at 100 epochs and radiomics features were extracted every 25 epochs (i.e.., at 25, 50, 75 and 100 epochs). No early stopping was adopted for terminating the model training. Table 6-2 offers a concise summary of our experiments.

Radiomics Extraction

The masks of the regions of interest (ROIs) are stored in DICOM format in 3D in the Lung 1, RIDER and NSCLC Radiogenomics datasets. The modality of these files is 'SEG'. DICOM CT images were converted to 3D images using the SimpleITK (v1.2.4) software. We resampled the images to 2 mm isotropic voxels prior to feature extraction. Radiomics features were extracted using the pyRadiomics open-source Python library [51] (v2.2.0). A total of 103 features were extracted. These consisted of 13 morphology (shape) features, 17 intensity-histogram (first-order) features and 73 textural (Haralick) features. The full list of features and the settings used for pyRadiomics can be found in the supplementary Table 5.

Table 6-1. Scanning parameters of included datasets

Parameters / Datasets	Scanner	Radiation Dose (mAs)	Slice Thickness (mm)	Spatial Resolution (mm)
Phantom Dataset	Philips(4*)	50(2),400(2)	3(4)	[0.77,0.77]
LUNG 1	Siemens(157)	400(157)	3(157)	[0.98,0.98]
TCGA-LUAD	Siemens(14)	110(3),120(8),140(2),210(1)	1(2),5(8),8(4)	[0.59,0.59]-[0.74,0.74]
LIDC-IDRI	GE(12)	<1(12)	1.25(7),2.5(5)	[0.53,0.53]-[0.70,0.70]
RIDER	N/A**(56)	4(4),5(4),6(6),7(13),8(13),9(9),10(7)	1.25(56)	[0.51,0.51]-[0.82,0.82]
NSCLC Radiogenomics	N/A(5), Philips(1), GE(90), Siemens(10)	38.65±81.97	0.625(7), 1(11),2(1), 1.25(75), 2.5(9),3(2)	[0.59,0.59]-[0.98,0.98]

± standard error of the mean; *number of included scans; ** manufacturer not mentioned in DICOM metadata

Table 6-2. Summary of Experiment and Corresponding Datasets

Experiment	Training Strategy	Training Dataset	Testing Dataset
Simulated data based	With slice-	Part of paired high-noise	The rest of **high-noise CTs**

Training	pairing	and full dose **Lung 1** dataset (n=40, 4260 Frames)	(n=117, 13423 Frames), **low-noise CTs** (n=157, 17683 Frames), **phantom dataset CTs** (n=2, 104 Frames)
Ablation Study	Without slice-pairing	Part of paired high-noise and full dose **Lung 1** dataset (n=40, 4260 Frames)	The rest of **high-noise CTs** (n=117, 13423 Frames), **low-noise CTs** (n=157, 17683 Frames)
Applications with simulated data training-based networks	With slice-pairing	Training finished at first part of experiment without re-training	**RIDER** (n=31, 14875 Frames), **NSCLC Radiogenomics** (n=106, 28404 Frames)
Applications with real data training-based networks	With slice-pairing	Low dose CTs from **LIDC-IDRI** (n=12, 3144 Frames), Full dose CTs from **TCGA-LUAD** (n= 14, 3307 Frames)	**RIDER** (n=31, 14875 Frames), **NSCLC Radiogenomics** (n=106, 28404 Frames)

The shape-related features are not affected by denoising and therefore were excluded from feature reproducibility analysis, resulting in 90 included features. All 103 features were used to derive the 4-year pre-treatment survival prediction model and time-to-event survival analysis.

Result

Training the cycle GAN from simulated and real data took 96 and 72 hours respectively. The loss of the generator during training is shown in Figure 6-2. We choose to plot steps rather than epochs in loss curves because, plotting epochs would make it harder to observe the turbulence of model training and the faster convergence of the slice-paired training strategy. Moreover, the size of the training dataset in the simulated dataset is different to the real dataset and plotting epochs would be an unfair comparison.

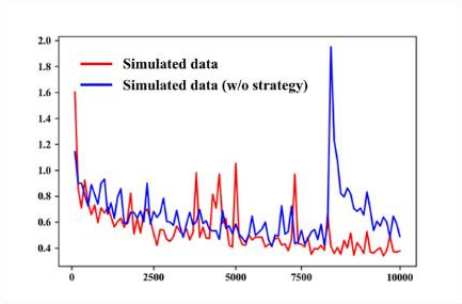


Figure 6-2. Generator loss over time for cycle GAN training runs with and without slice-pairing strategy.

Reproducibility of Radiomic Features on Simulated Paired Data

In the phantoms' dataset, the RMSEs of denoised versus high dose CT scans using the cycle GAN trained on simulated data and the cycle GAN trained on real data were 0.0187 and 0.0226 respectively, compared with 0.0231 in

the original low dose CTs. The encoder-decoder network and CGAN trained on simulated data achieved RMSEs of 0.0182 and 0.0140 respectively in same dataset. Based on visual inspection, we did not detect any image artifacts introduced during the cycle GAN based denoising.

An example of an original, noisy and denoised CT scan is shown in Figure 6-4. We reuse results of CGAN and EDN from [13] for better comparison with the cycle GAN (corresponding Figure for high noise image is Supplementary Figure 3). In addition, Table 6-3 shows the RMSE, perceptual loss, signal-to-noise ratio (SNR), and ratio of radiomic features with poor (CCC<0.65), medium (0.65⩽CCC<0.85), and good (CCC ⩾0.85) reproducibility [8]. The full result of CCC for every feature at different training epochs can be found in Supplementary Table 6-7.

As shown in Table 6-3, the RMSE and perceptual loss of low-noise and high-noise images (before denoising) are 0.0225/0.0706 and 0.0237/0.0781 respectively. The cycle GAN trained on simulated data reduced the RMSE and perceptual loss to 0.170/0.216 and 0.0181/0.0245 for low-noise and high-noise images; the cycle GAN trained on real data increased RMSE and perceptual loss to 0.0229/0.0531 for low-noise images and decreased RMSE and perceptual loss to 0.0230/0.0501 for high-noise images. The cycle GAN trained on simulated data resulted in higher RMSE than the CGAN but lower perceptual loss and outperformed the encoder-encoder network in both metrics. The cycle GAN trained on real data has a worse performance in denoising simulated noisy images compared to other networks.

The mean CCCs for cycle GAN trained on simulated data denoised images improved from 0.87 [95% CI, (0.833,0.901)] and 0.68 [95% CI,

(0.617,0.745)] to 0.93 [95% CI, (0.916, 0.949)] and 0.94 [95% CI, (0.928,0.954)] for low-noise images and high-noise images, respectively (Wilcoxon rank-sum test for the CCC from noisy images and denoised images, p-value<0.01 for both experiments). The mean CCCs of low noise images denoised with the cycle GAN trained on real data decreased to 0.81 [95% CI, (0.788,0.834)] and the mean CCCs of denoised high noise images increased to 0.80 [95% CI, (0.779,0.827)] (Wilcoxon rank-sum test comparing CCC of noisy images and denoised images: p-value<0.01 for both experiments). A heatmap of radiomics improvement from denoised low-noise images by comparing with original noisy images is shown in Figure 6-3.

In contrast, EDN and CGANs were able to improve the mean CCC of radiomic features to 0.92 [95%CI, (0.909,0.936)] for low and high-noise images. The cumulative distribution function (CDF) of CCCs for different models when trained for 100 epochs is shown in Figure 6-5 (a-b). The cycle GAN trained on real data did not manage to improve radiomics features' reproducibility on simulated noisy images. However, it still achieved a significant improvement in the reproducibility of radiomics features of simulated high noise images.

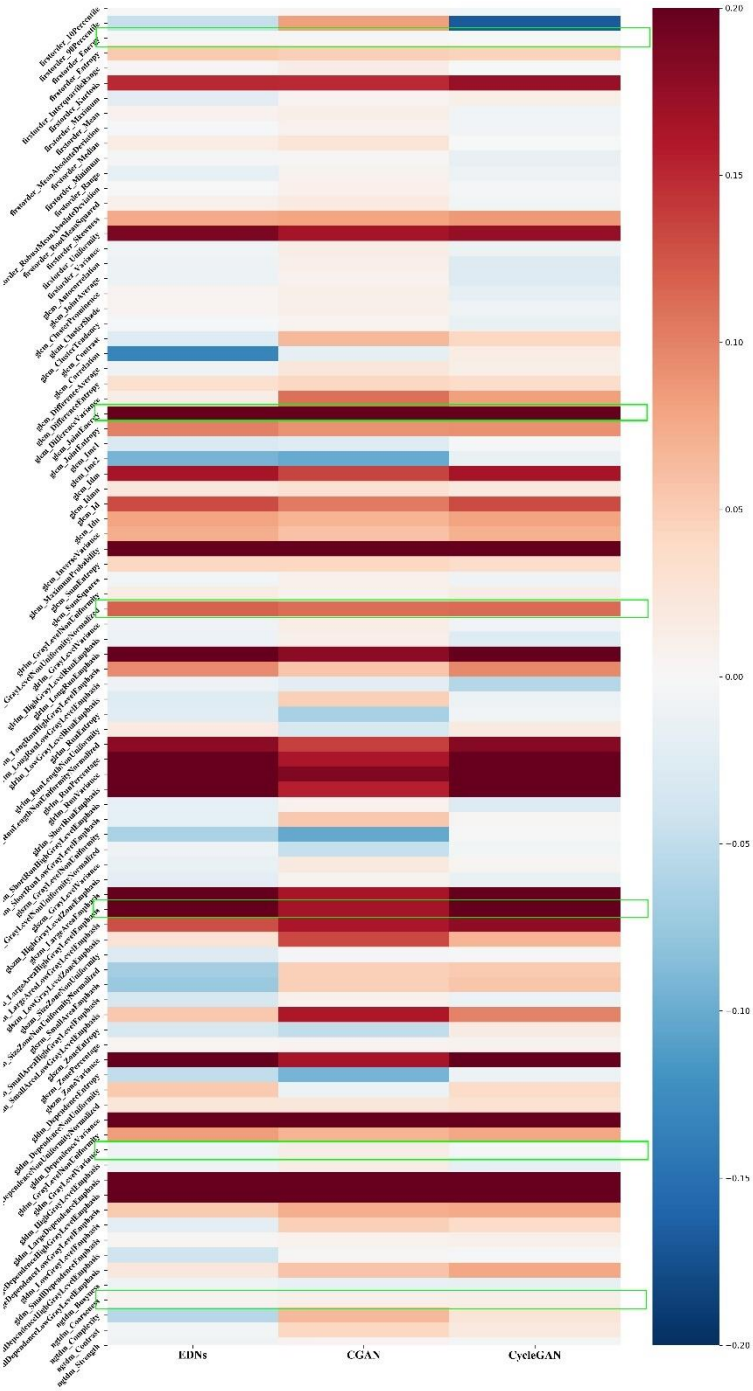

EDNs
CGAN
CycleGAN
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20

Figure 6-3. A heatmap of radiomics improvement from denoised low-noise images, results on EDN and CGAN are reproduced from [13]

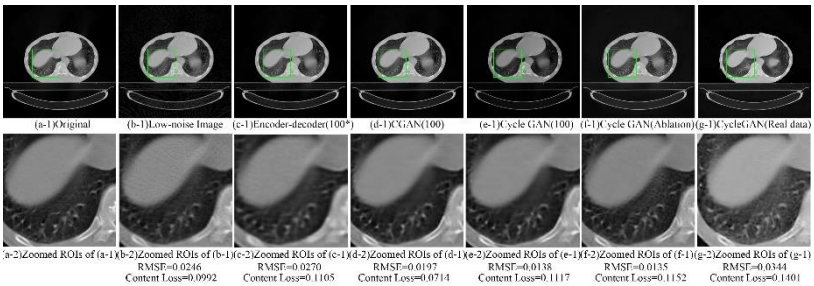
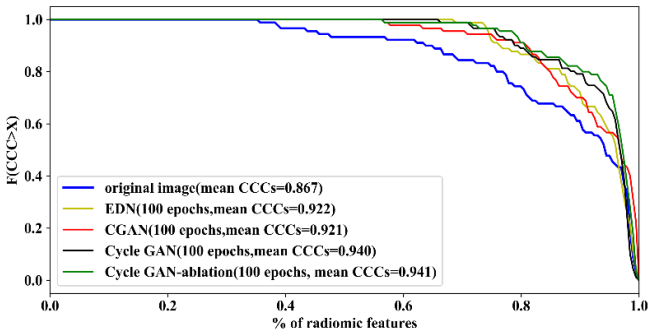
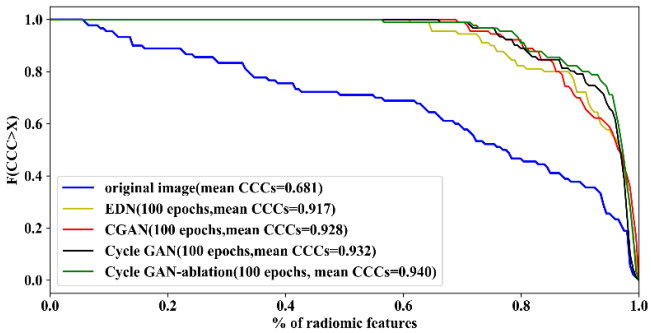


Figure 6-4. Example of low dose CT denoising. (a-1) The original full dose CT image; (b-1) Low-noise image; (c-1) Image denoised by EDN (*Training at 100 epochs); (d-1) Image denoised by CGAN; (e-1) Image denoised by a cycle GAN; (f-1) Image denoised by a cycle GAN (ablation study); (g-1) Image denoised by cycle GAN trained on real data; (a-2) to (g-2) Zoomed ROIs for (a-1) to (g-1).

The second investigation of the simulation study was the effect of different training epochs to radiomics reproducibility. The CDF of CCCs for cycle GAN trained at 25, 50, 75 and 100 epochs are shown in Supplementary Figure 4 (a-b). Summary of RMSE, perceptual loss and CCCs of cycle GAN trained at different epochs can be found in Supplementary Table 8. We compared the CCC distributions of radiomic features calculated on images denoised from high-noise images with those of images denoised from low-noise images using the Wilcoxon rank-sum test resulting in a p-value of 0.94. The results show that a cycle GAN trained to denoise high-noise images can be applied to denoise images with different levels of noise and achieve similar results to a CGAN and EDN based denoiser [13].



(a)



(b)

Figure 6-5. CDF of CCC of radiomic features denoised with different models. (a) low-noise images; (b) high-noise images.

Moreover, we compared the CCC distributions from cycle GAN with CGAN and EDN by using the Wilcoxon rank-sum test which resulted in p-values of 0.73 and 0.07, respectively. The results show that a cycle GAN achieved similar results to CGAN and EDN, and that in some cases, Cycle Gan even received better results.

Ablation Study for the Training Strategy

An example of denoised images from cycle GAN ablation study can be found in Figure 6-4 (f-1) and Figure 6-4 (f-2).

Table 6-3 and Supplementary Table 9 shows the RMSE, perceptual loss and ratio of poor, medium, and good reproducibility radiomic features about ablation study of cycle GAN. The cycle GAN trained without our training strategy can also reduce the RMSE and perceptual loss of low-noise and high-noise images to 0.0167/0.0258 and 0.0188/0.0256 respectively. Moreover, it can increase the average CCC to 0.94 [95%CI, (0.924,0.957)] and 0.93 [95%CI, (0.917,0.953)] for low and high-noise images respectively. The CDF of CCCs for ablation study when trained for 100 epochs is shown in Figure 6-5 (a-b) and the differences among epochs can be found in Supplementary Figure 4 (c-d).

The distribution of CCCs from ablation study trained at 100 epochs was compared with results from a network trained with training strategy and we found no signification differences (Wilcoxon rank-sum test, p-value=0.13). Figure 6-2 shows that training the cycle GAN with the training strategy might speed up convergence slightly. On the other hand, without the training strategy, the generator's loss function increases beyond 60000 steps. Finally, the cycle GAN trained with our training strategy led to significantly higher CCCs when trained for only 25 epochs (Wilcoxon rank-sum test, p-value < 0.01), as shown comparing Supplementary Figure 4(a) to (c) and Figure 4(b) to 4(d).

Reproducibility on Real Data

We now focus on the impact of denoising on the reproducibility of radiomic features in same day repeat low dose CT scans (RIDER dataset). An example of an original image and its denoised counterparts denoised using a CGAN, an EDN and the cycle GANs trained on simulated and real data are shown in Figure 6-6. Figure 6-7 shows the CDF of the CCCs for the radiomic features extracted from the original and denoised CT images. The cycle GAN trained on real data outperforms the rest of generative models (Wilcoxon rank-sum test, p-value < 0.01). On the other hand, the performance of the cycle GAN trained on simulated data is similar to that of the EDN and CGAN (p-value = 0.87 and 0.40 for respectively).

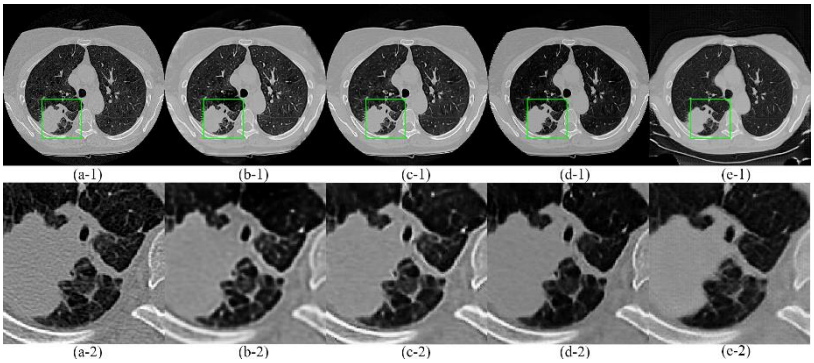


Figure 6-6. Example of RIDER denoising. (a-1) One original image from RIDER; (b-1) Image denoised by EDN (Training at 100 epochs); (c-1) Image denoised by CGAN (Training at 100 epochs); (d-1) Image denoised by cycle GAN trained on simulated data (100 epochs); (e-1) Image denoised by cycle GAN trained on real data (100 epochs); (a-2) to (e-2) Zoomed ROIs for (a-1) to (e-1).

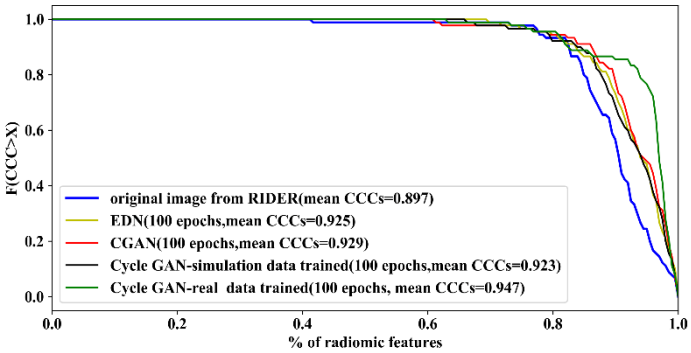


Figure 6-7. CDF of CCCs and for denoised CT scans in the RIDER dataset.

Survival prediction on Real Data

An example of an original NSCLC Radiogenomics image, and its denoised counterparts based on CGAN, EDN and cycle GANs trained from simulated and real data can be found in Supplementary Figure 5.

Figure 6-8 (a) illustrates the results of the of 4-year pre-treatment survival prediction experiment showing the AUC for each generative model across different number of epochs. We achieved an AUC for survival prediction based on radiomics extracted from the original NSCLC Radiogenomics dataset of 0.52 [95%CI, (0.511,0.538)] at 100 epochs. Denoising the CT scans using a CGAN or an EDN led to models with an increased AUC of 0.57 [95%CI, (0.551, 0.580)] (at 100 epochs) as shown in [26].

Table 6-3. Summary of RMSE, perceptual loss and distribution of CCCs of radiomic features based on denoising simulated datasets.

Models \ Distribution	RMSE	Perceptual loss	SNR (dB)	CCCs <0.65	0.65≤ CCCs<0.85	CCCs ≥0.85	95%CI of CCC
Low-noise Images							
Without denoising	0.0225	0.0706	18.3	10%	22%	68%	(0.833, 0.901)
Encoder-decoder	0.0173	0.0427	19.6	0%	19%	81%	(0.901, 0.935)*
CGAN	0.0143	0.0290	21.0	3%	17%	80%	(0.905, 0.939)*
Cycle GAN	0.0170	0.0216	24.6	0%	16%	84%	(0.916, 0.949)
Cycle GAN (w/o slice pairing)	0.0167	0.0258	20.8	1%	13%	86%	(0.924, 0.957)
Cycle GAN (real data)	0.0229	0.0531	15.7	6%	52%	42%	(0.788, 0.834)
High-noise Images							

Without denoising	0.0237	0.0781	6.1	36%	23%	41%	(0.617, 0.745)
Encoder-decoder	0.0175	0.0443	19.3	4%	16%	80%	(0.901, 0.935)*
CGAN	0.0146	0.0305	20.8	0%	16%	84%	(0.905, 0.939)*
Cycle GAN	0.0181	0.0245	20.3	0%	14%	86%	(0.928, 0.954)
Cycle GAN (w/o slice pairing)	0.0188	0.0256	19.4	3%	12%	84%	(0.917, 0.953)
Cycle GAN (real data)	0.0230	0.0501	15.4	4%	54%	42%	(0.779, 0.827)

* results reproduced from [13]

The cycle GANs trained on simulated and real data resulted in a higher mean AUC of around 0.58 [95%CI, (0.576,0.596)] but the difference between models was not statistically significant (Student's t-test, all p-values > 0.10). Figure 6-8 (b) illustrates the results of the time-to-event survival analysis experiment showing the C-index for each generative model across different numbers of epochs. EDN, CGAN, and the cycle GAN trained on simulated data improved C-index of survival analysis from 0.73 to around 0.76 while the cycle GAN trained on real data improved the C-index to 0.78.

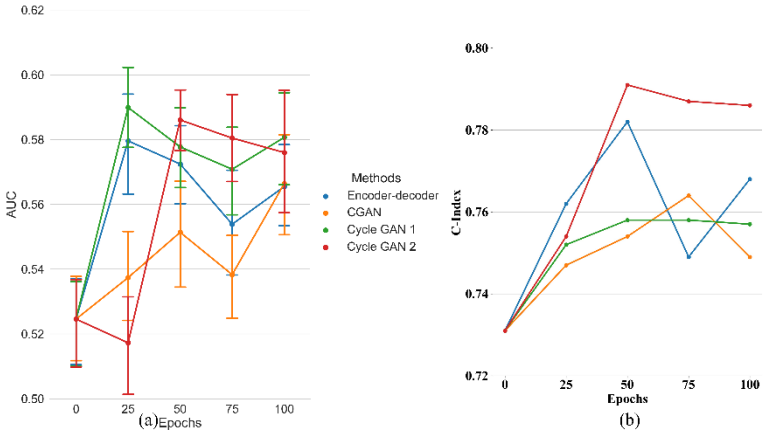


Figure 6-8. Results of 4-year pre-treatment survival prediction (a) and time-to-event survival analysis (b) C-index of survival analysis.

To interpret the improvement of AUC in 4-year survival prediction tasks, we used an RBF kernel based SVM Recursive Feature Elimination algorithm [58] to assess the importance of features in the prediction model.

Table 6-4 shows the top eight most important features in the models trained on the radiomic features from the original images and those from denoised images (The table with all features can be found in Supplementary Table 10-12). Six features appeared in all four models (highlighted in green in Figure

6-3). These features' CCC improved by denoisers, most of them improved significantly, which might explain how denoising can improve the AUC of survival prediction models.

Shape features, which were previously excluded from denoising analyses, were included as candidate predictors for the survival prediction model. However, as shown in Table 6-4, there are no shape features among the top eight most important predictors.

Discussion

The objective of our study was to investigate the potential of cycle GANs for denoising low dose CTs to improve the reproducibility of radiomics features and the performance of radiomics-based models. For this purpose, we trained two cycle GANs, one with simulated paired data and the other one with real data, to denoise low dose CT scans. In order to measure the performance of our denoising models, we ran experiments and compared the results of our method with those of CGANs and EDNs trained on simulated paired data. The results show that both cycle GANs trained on simulated and on real data can improve radiomics' reproducibility and performance in low dose CT and achieve similar results compared to CGANs and EDNs.

The main advantage of cycle GANs over CGANs and EDNs is that they do not require paired images, which are hard to collect. For CGANs and EDNs we overcame this issue by generating simulated low dose CTs by introducing noise into high dose CTs [13]. However, simulated noise might differ from noise encountered in low dose CTs. Hence, being able to train a model on real low dose CT scans is a significant advantage.

Table 6-4 Top eight most important features in the survival prediction model trained on noisy images and images denoised using different generative models

Rank	Original images	Denoised with EDN	Denoised with CGAN	Denoised with cycle GAN
1	glszm_LargeArea LowGrayLevelEmphasis	glszm_LargeArea LowGrayLevelEmphasis	glszm_LargeArea LowGrayLevelEmphasis	glrlm_GrayLevel NonUniformityNormalized
2	ngtdm_Coarseness	gldm_GrayLevelVariance	glrlm_GrayLevel NonUniformityNormalized	glszm_LargeArea LowGrayLevelEmphasis
3	gldm_GrayLevelVariance	glszm_LargeArea LowGrayLevelEmphasis	gldm_GrayLevelVariance	gldm_GrayLevelVariance
4	firstorder_Energy	gldm_LargeDependence HighGrayLevelEmphasis	firstorder_Energy	firstorder_Energy
5	shape_MinorAxisLength	gldm_GrayLevel NonUniformity	gldm_GrayLevel NonUniformity	shape_MinorAxisLength
6	glrlm_GrayLevel NonUniformityNormalized	firstorder_Energy	ngtdm_Coarseness	ngtdm_Coarseness
7	glszm_LargeArea HighGrayLevelEmphasis	glcm_JointEntropy	glcm_JointEntropy	glcm_JointEntropy
8	glcm_JointEntropy	ngtdm_Coarseness	shape_MinorAxisLength	glrlm_RunLength NonUniformityNormalized

However, training cycle GANs is volatile, especially when the target domain and the source domain differ, as documented elsewhere [27][55]. Ideally, in order to maximize the chances of success for the training process, training data would be collected from the same scanner, with the same protocol (except radiation exposure), and from the same group of patients for the two domains (low and high dose CT). However, such a dataset is not available to us. Hence, we defined selection criteria for the training data so that the source and target image domains kept certain similarities. We chose scanner manufacturer and table height (which determines field of view and the height of human body) based on [12]. These inclusion criteria were introduced after several failed attempts at training a cycle GAN with the full dataset. Examples of failed training runs are shown in Figure 6-9. However, trained models retain certain generalizability and can achieve good results across different scanners with different parameter settings as shown in the results (images in the RIDER and NSCLC Radiogenomics datasets were scanned from multiple types of scanners with different protocols).

As shown in Table 6-1, the Lung 1 dataset differs more in terms of scanning parameters from the RIDER and NSCLC Radiogenomics datasets compared to the LIDC-IDRI and TCGA-LUAD datasets. It is therefore possible that the conditional GAN trained on simulated paired TCGA-LUAD data achieved a similar performance as the cycle GAN trained on real data. Future studies may confirm this hypothesis.

Our ablation study results seem different from research reported elsewhere [43] which found that a slice-based training strategy can improve denoising performance. The slice-paired training strategy we proposed seems to lead to slightly faster convergence as hinted by the loss plot and the models' results at 25 epochs. However, this strategy did not lead to significant

improvement of the networks' denoising performance at 100 epochs. One possible explanation is that the training strategy cannot make the resulting network a better approximator of the mapping from low dose CT domain to high dose. Figure 6-2 and the comparisons between Supplementary Figure 4 (a) to (c) and (b) to (d) seem to support this view. Another possible hypothesis for this phenomenon is that reproducibility and performance of radiomics may not be so sensitive to the quality of images when the quality reaches a certain threshold. We did not report results of the ablation study for the slice-paired training strategy when training on real data because the training of the cycle GAN failed to converge multiple times without slice pairing. The failure to converge was probably due to a higher heterogeneity in real data compared to simulated data (simulated data were collected from the same scanners while real data were collected from different scanners). Thus, the slice-pairing strategy seems to have made the network training more stable in our study.

As mentioned above, cycle GANs achieved a similar performance as CGAN and EDN trained on simulated data, slightly outperforming them in some experiments. The difference in performance might be explained by the differences in the architectures used: the generator in CGAN and the encoder-decoder is a 5-layer network while there are 9 ResNet blocks [57] (27 convolutional layers) in the cycle GAN's generators. Related articles have hypothesized [22] that neural networks for 'low level' domain adaptation – such as denoising – should be kept shallow, since texture transfer in 'low level' domain adaptation is not significant. However, the results in our study seem to show that very deep neural network can also achieve good performance in some 'low level' domain adaptation tasks.

Our training cohort population is smaller than the testing cohort population for two main reasons. First, we considered the size of our training sets (ranging from 3144 frames to 4260 frames) was sufficient based on 2D cycle GAN training set examples in the literature, that range between several hundreds to a few thousands [27][59]. Secondly, CCC of radiomics is sensitive to the number of subjects used. Moving more subjects (images) from testing datasets to training datasets would decrease the reliability of radiomics features' CCC calculations.

As shown in Figure 6-6, the cycle GAN trained on simulated data (Figure 6-6 (d)) seems to have a better denoising performance in some cases in terms of tissue enhancement and intensity smoothing on homogeneous regions compared with the model trained on real data (Figure 6-6 (e)). Of course, it might just be that among the tens of thousands of CT images in the experiment, this is one where the cycle GAN trained on simulated data fared better than its counterpart trained on real data. In addition, it might be that the data distribution (as well as the noise) in the simulated training data is more homogeneous than the real data and this might lead to more appealing visual results, but statistical metrics point at a consistently superior performance by the model trained on real data.

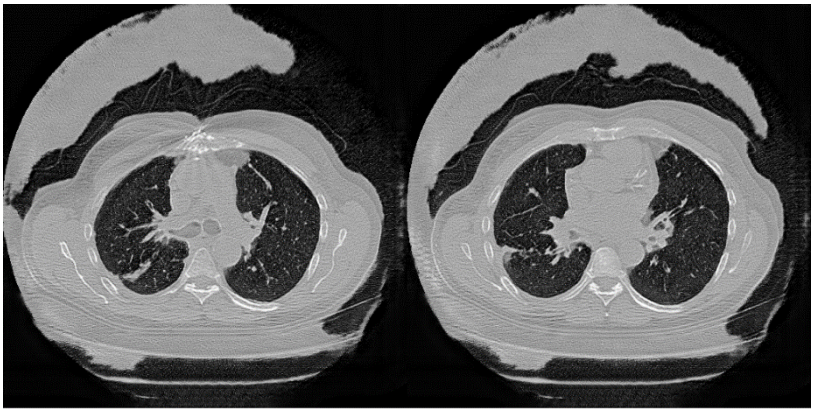
One potential limitation of our study is the low AUCs achieved by the models for pre-treatment survival prediction for lung cancer based on radiomic features. However, these are in line with results reported elsewhere. For example, Isensee et al. [60] reported an accuracy of 52.6% based on the BraTS 2017 dataset [61] for brain tumors using radiomics; Choi et al. [62] reported an integrated AUC (iAUC) of 0.620 [95% CI: 0.501–0.756] using the TCGA/TCIA dataset and random survival forest to derive a prediction

model; Finally, Bae et al. [63] reported an iAUC of 0.590 [95% CI: 0.502, 0.689] for overall survival prediction in glioblastoma using MRI radiomic features.

Our study suffered from a few other limitations. First, there were important differences between the populations in different training datasets (LIDC-IDRI and TCGA-LUAD). For example, patients in TCGA-LUAD were thinner than patients in LIDC-IDRI, as shown in Supplementary Figure 4. Hence, the cycle GAN trained on these datasets learnt to not only denoise the images, but make the patients thinner as illustrated in Figure 6-6 (e-1). Fortunately, the ROIs of this study are located in the lung and the volume of patients' lung in two domains are similar. Therefore, there was no significant size shift in the ROIs. Second, due to the differences of the CT bed in LIDC-IDRI and TCGA-LUAD, the cycle GAN also transforms bottom part of the image as shown in Figure 6-6 (e-1). Third, the cycle GAN trained on real data performed relatively poorly on simulated noisy images in terms of improving the reproducibility of radiomic features. One of the potential reasons is the domain distribution gap between real data and simulated data. The variations of scanners, patient cohorts, reconstruction algorithms in real training dataset may reduce the network's denoising performance in the simulated dataset. [53] Moreover, we believe that the good performance in real data is more important than the performance in simulated data, since it is more representative of real applications. Fourth, one of the assumptions of our slice-paired training strategy is that the first slice of a low dose CT scan will have higher similarity with the first slice of a high dose CT scan, is not automatically true. The similarity of the first slice of a CT scan depends on a lot of factors such as the patient position, section of the body

scanned etc. These factors were ignored in this paper. Fifth, as we mentioned in section 2.4, no early stopping of training was adopted in this study. However, as shown in Figure 6-2, Figure 6-5 and Supplementary Figure 4 that we cannot witness the improvements of the model's performance during training. This may mean that the generator of the cycle GAN does not learn the real data distribution, since the loss function fluctuated in all training steps (50000 steps in our case). Therefore, early stopping techniques and AutoML-based hyperparameter selection [64] seem like promising topics for further research. Sixth, in this study, the trained models were only tested in two applications: improving radiomics reproducibility in same-day repeat low dose CTs and radiomics performance survival prediction. More experiments to better understand the relationship between denoising and radiomics performance are needed. Seventh, radiation dose is not the only source of lack of reproducibility of radiomics' features, and in some cases, it might not be the most relevant. [12] Therefore, a denoising model might not solve all the reproducibility issues of radiomic features and other measures will need to be put in place to address other sources of inconsistency (slice thickness, reconstruction parameters, contrast enhancement, etc.). Our GANs were trained on datasets collected from different scanners with different scanning parameters, and images were reconstructed using different software and kernels. This might lead to more robust models but at the same time we cannot guarantee that our trained GAN does not introduce new inconsistencies to radiomics. Moreover, there are some deep learning-based methods for extracting radiomics features, usually referred to as 'deep radiomics' [65][66][67]. However, there are limited studies focusing on extracting features from low dose CT and, to the best of our knowledge, no study focused on improving deep radiomics

reproducibility or performance in low dose CT. Studies focusing on assessing deep radiomics' reproducibility and performance in low dose CT would be of interest. Eighth, we did not compare the performance of the cycle GAN with non-AI commercial low dose CT reconstruction algorithms, such as model-based iterative reconstruction (MBIR) [68]). Such a comparison would be of interest, but we could not perform it in our study due to the absence of sinograms (which are required to use reconstruction algorithms) in the datasets used. Finally, due to the absence of a structure similarity term in our cycle GAN's cost function, some images develop distortions in microstructures. Therefore, further adjustments on cost function and network architecture should be assessed in the future.



(a) (b)

Figure 6-9. Examples of failed cycle GAN training.

Conclusions

In this study, we investigate the potential of denoising low dose CT using cycle GANs to improve the reproducibility of radiomics features and the performance of radiomics based prediction models. We trained two cycle

GANs: using paired simulated low dose CTs and unpaired real low and high CT images. To accelerate convergence, we introduced a slice-paired training strategy.

The results of our experiments show that a cycle GAN trained to denoise low dose CT scans from unpaired low and high dose CT scans can improve the reproducibility of radiomic features in simulated low dose CTs and same-day repeat low dose CTs. In addition, we showed that radiomics based pre-treatment survival prediction models trained on low dose CT scans denoised with said cycle GAN can achieve better performance. The improvement in reproducibility and prediction model performance are comparable to those achieved with CGANs and encoder decoder networks trained on simulated paired data. Cycle GANs have better potential because they do not need paired data, but they are burdened by the volatility of the treatment process, which limits their applicability. More research is needed to make cycle GAN training more robust, for them to be able to be trained on a more diverse dataset.

Appendix

Support materials of this Chapter can be found in this link.

References:

[1]La Thangue, Nicholas B., and David J. Kerr. "Predictive biomarkers: a paradigm shift towards personalized cancer medicine." *Nature reviews Clinical oncology* 8.10 (2011): 587-596.

[2]Aerts, Hugo JWL, et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." *Nature communications* 5.1 (2014): 1-9.

[3]Desseroit, Marie-Charlotte, et al. "Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non–small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort." *Journal of Nuclear Medicine* 58.3 (2017): 406-411.

[4]Bogowicz, Marta, et al. "Stability of radiomic features in CT perfusion maps." *Physics in Medicine & Biology* 61.24 (2016): 8736.

[5]Tixier, Florent, et al. "Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET." *Journal of Nuclear Medicine* 53.5 (2012): 693-700.

[6]Li, Hui, et al. "Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set." *NPJ breast cancer* 2.1 (2016): 1-10.

[7]Leandrou, Stephanos, et al. "Quantitative MRI brain studies in mild cognitive impairment and Alzheimer's disease: a methodological review." *IEEE reviews in biomedical engineering* 11 (2018): 97-111.

[8]Chaddad, Ahmad, Christian Desrosiers, and Matthew Toews. "Multi-scale radiomic analysis of sub-cortical regions in MRI related to autism, gender and age." *Scientific reports* 7.1 (2017): 1-17.

[9]Bodalal, Zuhir, et al. "Radiogenomics: bridging imaging and genomics." *Abdominal radiology* 44.6 (2019): 1960-1984.

[10]Berenguer, Roberto, et al. "Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters." *Radiology* 288.2 (2018): 407-415.

[11]Welch, Mattea L., et al. "Vulnerabilities of radiomic signature development: the need for safeguards." *Radiotherapy and Oncology* 130 (2019): 2-9.

[12]Meyer, Mathias, et al. "Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings." *Radiology* 293.3 (2019): 583-591.

[13]Chen, Junhua et al. "Generative Models Improve Radiomics Reproducibility in Low Dose Cts: A Simulation Study." *Physics in Medicine & Biology* 66.16 (2021): 165002.

[14]Sagheer, Sameera V. Mohd, and Sudhish N. George. "A review on medical image denoising algorithms." *Biomedical signal processing and control* 61 (2020): 102036.

[15]Rudin, Leonid I., Stanley Osher, and Emad Fatemi. "Nonlinear total variation based noise removal algorithms." *Physica D: nonlinear phenomena* 60.1-4 (1992): 259-268.

[16]Chan, Raymond H., Chung-Wa Ho, and Mila Nikolova. "Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization." *IEEE Transactions on image processing* 14.10 (2005): 1479-1485.

[17]Chen, Yunjin, and Thomas Pock. "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image

restoration." *IEEE transactions on pattern analysis and machine intelligence* 39.6 (2016): 1256-1272.

[18]Zhang, Kai, et al. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising." *IEEE transactions on image processing* 26.7 (2017): 3142-3155.

[19]Zhang, Kai, et al. "Learning deep CNN denoiser prior for image restoration." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[20]Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012): 1097-1105.

[21]Lefkimmiatis, Stamatios. "Non-local color image denoising with convolutional neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[22]Chen, Hu, et al. "Low-dose CT with a residual encoder-decoder convolutional neural network." *IEEE transactions on medical imaging* 36.12 (2017): 2524-2535.

[23]Yang, Qingsong, et al. "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss." *IEEE transactions on medical imaging* 37.6 (2018): 1348-1357.

[24]Goyal, Bhawna, et al. "Image denoising review: From classical to state-of-the-art approaches." *Information fusion* 55 (2020): 220-244.

[25]Li, Yajun, et al. "Normalization of multicenter CT radiomics by a generative adversarial network method." *Physics in Medicine & Biology* 66.5 (2021): 055030.

[26]Chen, Junhua, et al. " Generative Models Improve Radiomics Performance in Different Tasks and Different Datasets: An Experimental Study" *arXiv preprint arXiv: 2109.02252* (2021).

[27]Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.

[28]Yang, Wei, et al. "Improving low-dose CT image using residual convolutional network." *IEEE access* 5 (2017): 24698-24705.

[29]Chen, Hu, et al. "Low-dose CT with a residual encoder-decoder convolutional neural network." *IEEE transactions on medical imaging* 36.12 (2017): 2524-2535.

[30]Yang, Qingsong, et al. "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss." *IEEE transactions on medical imaging* 37.6 (2018): 1348-1357.

[31]Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." *arXiv preprint arXiv:1411.1784* (2014).

[32]Manning, Christopher, and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[33]Olkin, Ingram, and Friedrich Pukelsheim. "The distance between two random vectors with given dispersion matrices." *Linear Algebra and its Applications* 48 (1982): 257-263.

[34]Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." *European conference on computer vision*. Springer, Cham, 2016.

[35]McCollough, C.H., et al. (2020). Low Dose CT Image and Projection Data [Data set]. *The Cancer Imaging Archive*. https://doi.org/10.7937/9npb-2637.

[36]Zhao, Binsheng, et al. "Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non–small cell lung cancer." *Radiology* 252.1 (2009): 263-272.

[37]Kang, Eunhee, et al. "Cycle-consistent adversarial denoising network for multiphase coronary CT angiography." *Medical physics* 46.2 (2019): 550-562.

[38]Yang, Serin, Eung Yeop Kim, and Jong Chul Ye. "Continuous Conversion of CT Kernel using Switchable CycleGAN with AdaIN." *IEEE Transactions on Medical Imaging* (2021).

[39]Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[40]Lawrence, I., and Kuei Lin. "A concordance correlation coefficient to evaluate reproducibility." *Biometrics* (1989): 255-268.

[41]Chen, Junhua, et al. "Lung Cancer Diagnosis Using Deep Attention Based on Multiple Instance Learning and Radiomics." *arXiv preprint arXiv:2104.14655* (2021).

[42]Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.

[43]Yang, Heran, et al. "Unsupervised MR-to-CT Synthesis Using Structure-Constrained CycleGAN." *IEEE transactions on medical imaging* 39.12 (2020): 4249-4261.

[44]Zhovannik, Ivan, et al. "Learning from scanners: Bias reduction and

feature correction in radiomics." *Clinical and translational radiation oncology* 19 (2019): 33-38.

[45]Aerts, H. J. W. L., Wee, L., Rios Velazquez, E., Leijenaar, R. T. H., Parmar, C., Grossmann, P., … Lambin, P. (2019). Data From NSCLC-Radiomics [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI

[46]Bakr, Shaimaa, et al. "A radiogenomic dataset of non-small cell lung cancer." *Scientific data* 5.1 (2018): 1-9.

[47]Armato III, Samuel G., et al. "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 (2011): 915-931.

[48]Albertina, B., Watson, M., Holback, C., Jarosz, R., Kirk, S., Lee, Y., … Lemmerman, J. (2016). Radiology Data from The Cancer Genome Atlas Lung Adenocarcinoma [TCGA-LUAD] collection. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2016.JGNIHEP5

[49]Huang, Jin, and Charles X. Ling. "Using AUC and accuracy in evaluating learning algorithms." *IEEE Transactions on knowledge and Data Engineering* 17.3 (2005): 299-310.

[50]Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[51]Van Griethuysen, Joost JM, et al. "Computational radiomics system to decode the radiographic phenotype." *Cancer research* 77.21 (2017): e104-e107.

[52]Traverso, Alberto, et al. "Repeatability and reproducibility of radiomic features: a systematic review." *International Journal of Radiation Oncology* Biology* Physics* 102.4 (2018): 1143-1158.

[53]Bashyam, Vishnu M., et al. "Deep Generative Medical Image Harmonization for Improving Cross - Site Generalization in Deep Learning Predictors." *Journal of Magnetic Resonance Imaging* 55.3 (2022): 908-916.

[54]Cawley, Gavin C., and Nicola LC Talbot. "On over-fitting in model selection and subsequent selection bias in performance evaluation." *The Journal of Machine Learning Research* 11 (2010): 2079-2107.

[55]Lee, Junghyun, Jawook Gu, and Jong Chul Ye. "Unsupervised CT Metal Artifact Learning using Attention-guided β-CycleGAN." *IEEE Transactions on Medical Imaging* 40.12 (2021): 3932-3944.

[56]Da-Ano, R., D. Visvikis, and M. Hatt. "Harmonization strategies for multicenter radiomics investigations." *Physics in Medicine & Biology* 65.24 (2020): 24TR02.

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

[58]Liu, Quanzhong, et al. "Feature selection for support vector machines with RBF kernel." *Artificial Intelligence Review* 36.2 (2011): 99-115.

[59]Liu, Wei, et al. "End-to-end single image fog removal using enhanced cycle consistent adversarial networks." *IEEE Transactions on Image Processing* 29 (2020): 7819-7833.

[60]Isensee, Fabian, et al. "Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge." *International MICCAI Brainlesion Workshop*. Springer, Cham, 2017.

[61]Menze, Bjoern H., et al. "The multimodal brain tumor image segmentation benchmark (BRATS)." *IEEE transactions on medical imaging* 34.10 (2014): 1993-2024.

[62]Choi, Yoon Seong, et al. "Machine learning and radiomic phenotyping of lower grade gliomas: improving survival prediction." *European radiology* (2020): 1-9.

[63]Bae, Sohi, et al. "Radiomic MRI phenotyping of glioblastoma: improving survival prediction." *Radiology* 289.3 (2018): 797-806.

[64]He, Xin, Kaiyong Zhao, and Xiaowen Chu. "AutoML: A Survey of the State-of-the-Art." *Knowledge-Based Systems* 212 (2021): 106622.

[65]Zhao, Binsheng. "Understanding sources of variation to improve the reproducibility of radiomics." *Frontiers in oncology* 11 (2021): 826.

[66]Afshar, Parnian, et al. "From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities." *IEEE Signal Processing Magazine* 36.4 (2019): 132-160.

[67]Lao, Jiangwei, et al. "A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme." *Scientific reports* 7.1 (2017): 1-8.

[68]Ziegler, A., Th Köhler, and R. Proksa. "Noise and resolution in images reconstructed with FBP and OSC algorithms for CT." *Medical physics* 34.2 (2007): 585-598.

Chapter 7

General Discussion and Conclusions

In this thesis, we studied the use of generative models to improve the reproducibility and performance of radiomics in low dose CTs. More specifically, we discussed the benefits of shortcuts in encoder-decoder networks for CT denoising, with implications on the design of such generative models (**Chapter 2**). Subsequently, we investigated the improvement of low dose CT radiomics reproducibility using generative models trained on paired simulated data (**Chapter 3**). To investigate the effect of generative models in improving low dose CTs radiomics performance more comprehensively, we applied radiomics into a new application – deep attention-based multiple instance learning (MIL) for lung cancer diagnosis (**Chapter 4**). Next, we used pre-trained denoising generative models to validate its effect in improving low dose CT radiomics performance on multiple applications - pre-treatment survival prediction and lung cancer diagnosis (**Chapter 5**). Finally, we investigated the possibility of training a low dose CT denoiser from unpaired real training data to improve low dose CT radiomics reproducibility and performance (**Chapter 6**).

In this chapter, we will discuss (1) the remaining challenges and possible solutions of using generative models as a pre-processing tool to improve low dose CT radiomics performance (2) other possible solutions to improve low dose CT radiomics performance (3) overarching conclusions of this body of work.

Barriers for the application of generative models into low dose CTs radiomics

The challenges and barriers that prevent the widespread application of generative models into low dose CTs radiomics can be divided into two categories, technical barriers and barriers related to implementation. These barriers will be discussed in the following sections.

Technical barriers

The mentioned generative models have some common shortcomings preventing their widespread adoption in real low dose CTs radiomics and related applications. For example, **model collapse** (an issue arising in GAN training where a generator model is only capable of generating a small subset of outcomes, see [2] for more details) **during generative models training** will blur outputs [3][4], decrease the performance of CT denoising and prevent the improvement of low dose CT radiomics performance. In addition, it is difficult to **find the best architecture and hyperparameter settings for generative models** to better enhance image quality, improve radiomics performance and more quickly push generative models into real low dose CT radiomics practices.

Besides the common technical barriers for all generative models, there are significant differences in major technical barriers and corresponding solutions for applying generative models trained on paired and unpaired data, which will be discussed separately.

As we discussed in Chapter 3, **one of the big barriers for training most generative models is the shortage of open access training datasets**, especially for generative models based on supervised training. Although

some datasets have been made publicly available (such as the dataset from The American Association of Physicists in Medicine (AAPM) [1]), there are limited datasets for training generative models in a supervised manner. The mentioned dataset [1] consisted of simulated data where the simulated low dose CTs had a noise level that corresponded to CT exposure at 50 Milliampere-seconds (mAs). However, radiation exposure of low dose CT is generally much lower than 50 mAs in clinical practice (see for example, the distribution in the RIDER dataset, as shown in Table 6-1). Moreover, simulated data was generated by inserting noise with a certain distribution into CT sinograms. [5][6] However, the patterns of noise in real world low dose CTs are generally more complex. Therefore, models trained on the mentioned simulation dataset might not be able to reach a good performance in real low dose CT datasets.

Compared with paired data trained generative models, major challenges for their unsupervised counterparts are **selecting the most suitable training data from the huge volume of CT data pooling for successful generative model training and enhancing models' image synthesis ability**. Typical architectures for unsupervised generative models are cycle GAN and style GAN. [7]

As mentioned in the original cycle GAN article [8], the training is more likely to be successful and stable when focusing on pairs of visually similar images. **Multiple variations** -- data collection settings, image reconstruction kernels and differences across patient cohorts -- **will change the texture of CT image significantly, some of which are important for successful cycle GAN training.** Based on the study of this thesis, some parameters of CT scans - table height and field of view of CT - have a strong impact on cycle GAN training. Differences in table heights can lead to a

strong variation for the position of the body in the transversal plane and images from this plane are widely used in CT related image-to-image translation.[9][10] Hence, image pairs should be considered as having low visual similarity when the variance in table height is large. In other words, a high variance in the table height in training datasets will lead to failures in cycle GAN training. Additionally, the field of view is similar to zooming in and out of the region of interest (ROI). [12] Unfortunately, cycle GANs have a poor performance in zoom related image translation. [13][14]

Compared with paired trained generative models, **denoising performance of cycle GANs is questionable due to their imperfect ability to keep semantic information from noisy images to denoised images**. The original cycle GAN article [8] mentioned the "upper bound" of performance should be paired data trained counterpart (Pix2pix, [9]) if the same training data is used, which our experiments in Chapter 6 seem to support. Therefore, another possible barrier for applying unsupervised generative models into real practice is the poor image synthesis ability of existing models.

The black box nature of deep learning is a well-known topic for debate between researchers that has been going on for decades. [16] A black box means that we do not know how all the individual neurons work together to arrive at the final output of the network. However, interpretability is extremely important in medical image analysis [17]. Unfortunately, images and related radiomics features from denoised by opaque, black box networks may lead to features that are not fully trusted by clinicians and decisions based on these features may be unconvincing. Moreover, clinicians might be led to make an incorrect decision if denoising networks introduce external errors into images.

Barriers related to implementation

Radiomics related applications are mainly used by medical physicists and clinical researchers, [18][19][20] whose **computer coding ability is limited**. However, as a new technology within artificial intelligence, such software engineering skills are needed for building, training, testing, and applying generative models into clinical practice. The high coding effort necessary for implementing generative models prevent researchers from applying them in low dose CTs radiomics widely. [21]

On the other hand, massive GPU computational resources are needed for training, validating, and testing of generative models. However, these **resources are hard for medical physicists and clinical researchers to access** due to their high cost. [22]

Potential solutions for the technical barriers

As mentioned above, model collapse is a typical problem encountered when training generative models, as reported multiple times in the literature. [23][24] Early stopping, an optimization technique used to stop training before overfitting happens without compromising on model accuracy, **seems an effective method** to avoid model collapse. Moreover, early stopping can save computational resource during training. [25][26] **Replacing cost function of networks from L2 loss to Wasserstein loss** seems to be another possible solution to alleviate model collapse as reported in recent studies. [27][28]

Neural architecture search and automatic hyperparameter setting is a hot research topic in machine learning and is referred to as **AutoML**. [29][30] Some studies have focused on using AutoML methods to automatically

build network and set hyperparameter in generative models designing. [31][32] We think AutoML has potential to find the best settings of the network for denoising low dose CT and improving radiomics performance. However, no studies have been published on this topic, which makes it a promising topic for further research.

The major shortcoming preventing supervised trained generative models (encoder-decoder network and CGAN) into low dose CTs radiomics is the absence of suitable training datasets. **One of solutions for this issue is training networks based on simulated data** as shown in Chapter 3. However, this kind of model may not be able to achieve good performance in real datasets due to the multiple variations (such as vendor scanners, reconstruction kernels) and the difference between simulated noise and real noise. Another potential solution for the absence of training datasets is collecting the paired dataset, which might be costly or even unfeasible.

Variation of table height is an important parameter for successful cycle GAN training, because it is a significant source of variation in a sinogram. Three possible methods may decrease the impact of this variation. First, **using the sinogram of CT images** to let the network adjust for the effect of table height on the images. Second, **limiting the spread of table height in the training data** by **data selection**. In our study (Chapter 6), we limited the table height of CT scans to be between 160mm and 170mm. Third, taking CT table as marker to **register CT scans** and moving the CT table into the same position for all scans. [33] This CT table moving based method seems a more suitable method to normalize table height of CT scans. **Field of view** is another parameter introducing variation that affects training of cycle GANs; we believe that one of effective methods to decrease the impact of

this variation is **data selection** before model training. In general, variance of mentioned factors should be limited in training datasets by selecting or image pre-processing to prevent the training of cycle GANs being affected.

Improving cycle GAN image synthesis performance in the whole CT slide is a difficult task. As mentioned earlier, the upper bound of the performance cycle GANs are expected to achieve is that of the supervised learning counterparts. However, in radiomics, only the part of the image inside the mask (ROI) is relevant. Therefore, generative models for radiomics related studies focus on the ROI. Masks should be regarded as part of inputs for model training and mask-related terms should be added into the cost function when training the networks in order to push models to focus on the ROI. This kind of studies are called **segmentation guided image synthesis**, [34][35] which belong to a more general topic - task driven image synthesis. [36] Segmentation guided image synthesis may can be included into generative models for improving radiomics performance in the future.

Explainable artificial intelligence (XAI) is artificial intelligence (AI) whose decisions or predictions can be understood by humans. [37] **XAI** is a hot research topic in recently years that has focused on how to **decode the black box of AI** for humans. Multiple techniques have been introduced to increase the interpretability of results from discriminative models. For example, class activation mapping (CAM) has been proposed to show a heatmap the relevance of each pixel on the model's decision making during image classification, [38] More details about the state-of-the-art XAI for discriminative models can be found in [39]. On the other hand, some studies

are beginning to **make image synthesis of generative models more controllable** and **make the results more explainable for humans**. For examples, StyleGAN introduced controllable factors into generative models at different scale during image synthesis; [40] Disentangled Representation Learning (DRL) aims to establish the mapping between image attributes and latent feature space to make image semantic edition in latent space more controllable. [41] However, studies of controllable edition in medical image analysis are limited. Generative models still are black boxes for medical imaging and networks cannot provide a confidence for the output. More studies about increasing the interpretability of generative models in medical imaging are needed in the future.

Potential solutions for implementation barriers

Regarding the **limited coding ability of medical physicists** and clinical related researchers, **open science** is a good tendency to remove the coding barriers and alleviate this question. [42] More and more top-tier computer science conferences and journals have recommended authors to publish their source code when submitting their manuscripts. Therefore, most of the source code for the famous networks are publicly available in code share sharing platforms such as GitHub [43] and Gitlab. [44] Contributions of computer science researchers are thus removing part of the barrier. To contribute tho this approach, all software codes of this thesis have been made public. The links to the source code repositories codes can be found in the corresponding chapters.

In addition to the source code of generative models, pre-trained generative models are hard for non-experts to access. This poses another barrier to implement generative networks into clinical practice. Training generative

models is computationally expensive and time-consuming, and it is generally difficult for care professionals to access high-performance computational resources. Compared with source code, there are fewer researchers that make their pre-trained models publicly available. Potential reasons for this phenomenon are the limited space of code sharing platform for each project and the volume of pre-trained models, which is much larger than that of the source code. Researchers need to upload their models into cloud platforms (such as Google Drive) if they want to share their models with others. The additional effort and fees will decrease the willingness of researchers to share their models. **Code sharing platforms such as GitHub should lift the volume limitation of each project while journals and conferences should recommend researchers to publish their models.** This may alleviate the difficulty of accessing pre-trained models. Important pre-trained models of this thesis have been uploaded to cloud platforms (Google Drive) and links to the models are published in the source code repositories of the corresponding projects.

More solutions to improve low dose CT Radiomics

This thesis only considered CT denoising in image post-processing. However, **low dose CT images quality can be enhanced not only in post-processing domain but also in the sinogram domain**. Some commercial low dose CT reconstruction solutions, such as Model-based iterative reconstruction (MBIR) algorithm, enhance low dose CT image quality by removing noise from sinogram data. [45] Moreover, some deep learning methods focused on improving low dose CT based on sinogram data, [46][47] or deep neural networks were used to perform CT image

reconstruction and denoising at the same time. [48] Although sinogram data is difficult to access for most researchers, investigating the possibility of improving low dose CT radiomics performance based on sinogram data is an interesting research topic for future studies.

As shown in Figure 1-1 and Chapter 1, we think **radiation dose is the biggest source of variation that decreases radiomics reproducibility in low dose CT. However, it is not the only source of variation that decreases the reproducibility and performance of radiomics**. Different scanner vendors, different reconstruction kernels, and different contrast enhancement techniques can result in significant changes in the texture and decrease the reproducibility of features in low dose CT. As mentioned in Chapter 1, the effect of some of these variations on the repeatability of features in normal dose CT has been reported in literature. However, these studies did not specifically focus on the effect of these variations in low dose CTs. Denoising networks trained as part of this thesis will not solve the reproducibility issues of radiomic features caused by these variations. A comprehensive image normalization method for low dose CTs may be useful to improve radiomics performance. [49][50] In some published articles, this is called radiomics harmonization. [52][53] Cycle GANs are a suitable basic architecture for this task.

Reproducibility is not the only characteristic that has been challenged in low dose CT radiomics, information representation ability has also been challenged. [53][54] In the deep learning era, [55] deep features have achieved impressive performance in natural image analysis and shown its strong information representation ability. [56][57] In light of the good results of deep features in natural image analysis, some studies have

combined deep features with radiomics features to enhance radiomics signature performance. These kinds of methods are referred to as **deep radiomics**. [58][59] Although deep radiomics has been reported to achieve a good performance, few studies focused on low dose CTs. Moreover, interpretability of included deep features is lower than radiomics features and the generalizability of pure deep features-based signatures' performance remains unclear. Therefore, more research on the performance of deep radiomics on low dose CT compared with classical radiomics is still needed.

The radiomics features mentioned so far have been extracted from single modality data – CT scans. However, multi-modality medical examinations are being used increasingly in clinical practice. For example, PET/CT scans are regarded as a powerful tool for tumor staging; [60] CT combined with gene data are used to decode tumor phenotypes and may provide clinical-decision support for doctors (known as radiogenomics); [61] and both CT and MRI are used for radiotherapy planning. [62] Low dose CTs are widely used in these examinations and **combining low dose CT radiomics features with features** (radiomics features and/or deep features) **from other modalities** can improve signature performance.

Conclusions

This thesis focused on using generative models to improve radiomics performance in low dose CTs.

We first investigated the benefits of shortcuts in encoder-decoder network for low dose CT denoising. Conclusions of this chapter provided some guidelines for generative models designing. (Chapter 2) Then we tested the effect of paired simulated data trained CGAN and encoder-decoder network to improve radiomics reproducibility in low dose CT. Results of this chapter showed paired simulated data trained generative models have a good performance in improving low dose CT radiomics reproducibility. (Chapter 3) Thirdly, radiomics was applied into a new application -radiomics and deep attention multiple instance learning based lung cancer diagnosis - for evaluating the improvement of generative models for radiomics performance. Results of this chapter shown that our proposed method achieved a better performance in lung cancer diagnosis with higher interpretability. (Chapter 4) Fourthly, paired simulation data trained generative models were applied to multiple datasets and radiomics based applications for evaluating these models' effect to improved radiomics performance in low dose CT. Results of this chapter showed tha pre-trained models can improve low dose CT radiomics performance. (Chapter 5) Finally, an unpaired real low-high dose CT data trained generative model - Cycle GAN - was adopted to investigate the possibility of achieving a similar performance from paired simulated data trained counterparts. Results of this chapter showed that unpaired data trained generative models can improve radiomics reproducibility and performance. (Chapter 6)

Results from different chapters showed that generative models can improve radiomics reproducibility and performance in low dose CT.

Reference:

[1]McCollough, Cynthia H., et al. "Low‐dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge." *Medical physics* 44.10 (2017): e339-e352.

[2]Goodfellow, Ian. "Nips 2016 tutorial: Generative adversarial networks." *arXiv preprint arXiv:1701.00160* (2016).

[3]Lala, Sayeri, et al. "Evaluation of mode collapse in generative adversarial networks." *High Performance Extreme Computing* (2018).

[4]Srivastava, Akash, et al. "Veegan: Reducing mode collapse in gans using implicit variational learning." *Advances in neural information processing systems* 30 (2017).

[5]Zeng, Dong, et al. "A simple low-dose x-ray CT simulation from high-dose scan." *IEEE transactions on nuclear science* 62.5 (2015): 2226-2233.

[6]Bankier, Alexander A., et al. "Air trapping: comparison of standard-dose and simulated low-dose thin-section CT techniques." *Radiology* 242.3 (2007): 898-906.

[7]Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[8]Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.

[9]Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[10]Gu, Jawook, and Jong Chul Ye. "AdaIN-based tunable CycleGAN for efficient unsupervised low-dose CT denoising." *IEEE Transactions on Computational Imaging* 7 (2021): 73-85.

[11]You, Chenyu, et al. "CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE)." *IEEE transactions on medical imaging* 39.1 (2019): 188-203.

[12]Chityala, R., et al. "Region of interest (ROI) computed tomography (CT): comparison with full field of view (FFOV) and truncated CT for a human head phantom." *Medical Imaging 2005: Physics of Medical Imaging*. Vol. 5745. SPIE, 2005.

[13]Zhu, Jin, Guang Yang, and Pietro Lio. "How can we make GAN perform better in single medical image super-resolution? A lesion focused multi-scale approach." *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019.

[14]Puiu, Andrei, et al. "Generative Adversarial CT Volume Extrapolation for Robust Small-to-Large Field of View Registration." *Applied Sciences* 12.6 (2022): 2944.

[15]He, Xin, Kaiyong Zhao, and Xiaowen Chu. "AutoML: A Survey of the State-of-the-Art." *Knowledge-Based Systems* 212 (2021): 106622.

[16]Castelvecchi, Davide. "Can we open the black box of AI?." *Nature News* 538.7623 (2016): 20.

[17]Vellido, Alfredo. "The importance of interpretability and visualization in machine learning for applications in medicine and health care." *Neural computing and applications* 32.24 (2020): 18069-18083.

[18]Peeken, Jan C., et al. "Radiomics in radiooncology–challenging the medical physicist." *Physica medica* 48 (2018): 27-36.

[19]Sollini, Martina, et al. "Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics." *European journal of nuclear medicine and molecular imaging* 46.13 (2019): 2656-2672.

[20]Shaikh, Faiq A., et al. "Technical challenges in the clinical application of radiomics." *JCO Clinical Cancer Informatics* 1 (2017): 1-8.

[21]Shen, Chenyang, et al. "An introduction to deep learning in medical physics: advantages, potential, and challenges." *Physics in Medicine & Biology* 65.5 (2020): 05TR01.

[22]Liu, Mengchen, et al. "Analyzing the training processes of deep generative models." *IEEE transactions on visualization and computer graphics* 24.1 (2017): 77-87.

[23]Dieng, Adji B., et al. "Avoiding latent variable collapse with generative skip models." *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.

[24]Bau, David, et al. "Seeing what a gan cannot generate." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[25]Zur, Richard M., et al. "Noise injection for training artificial neural networks: A comparison with weight decay and early stopping." *Medical physics* 36.10 (2009): 4810-4818.

[26]Li, Mingchen, Mahdi Soltanolkotabi, and Samet Oymak. "Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks." *International conference on artificial intelligence and statistics*. PMLR, 2020.

[27]Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks." *International conference on machine learning*. PMLR, 2017.

[28]Chen, Yingying, and Xinwen Hou. "An Improvement based on Wasserstein GAN for Alleviating Mode Collapsing." *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020.

[29]Elsken, Thomas, Jan Hendrik Metzen, and Frank Hutter. "Neural architecture search: A survey." *The Journal of Machine Learning Research* 20.1 (2019): 1997-2017.

[30]Liu, Chenxi, et al. "Progressive neural architecture search." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[31]Gong, Xinyu, et al. "Autogan: Neural architecture search for generative adversarial networks." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[32]Such, Felipe Petroski, et al. "Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data." *International Conference on Machine Learning*. PMLR, 2020.

[33]Yang, Bining, et al. "Deep learning improves image quality and radiomics reproducibility for high-speed four-dimensional computed tomography reconstruction." *Radiotherapy and Oncology* (2022).

[34]Zhou, Yang, et al. "3D Segmentation Guided Style-based Generative Adversarial Networks for PET Synthesis." *IEEE Transactions on Medical Imaging* (2022).

[35]Zhang, Yue, et al. "Unsupervised X-ray image segmentation with task driven generative adversarial networks." *Medical image analysis* 62 (2020): 101664.

[36]Chaitanya, Krishna, et al. "Semi-supervised and task-driven data augmentation." *International conference on information processing in medical imaging*. Springer, Cham, 2019.

[37]Vilone, Giulia, and Luca Longo. "Notions of explainability and evaluation approaches for explainable artificial intelligence." *Information Fusion* 76 (2021): 89-106.

[38]Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[39]Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information fusion* 58 (2020): 82-115.

[40]Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[41]Chartsias, Agisilaos, et al. "Disentangled representation learning in cardiac image analysis." *Medical image analysis* 58 (2019): 101535.

[42]Foster, Erin D., and Ariel Deardorff. "Open science framework (OSF)." *Journal of the Medical Library Association: JMLA* 105.2 (2017): 203.

[43]Dabbish, Laura, et al. "Social coding in GitHub: transparency and collaboration in an open software repository." *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 2012.

[44]Choudhury, Prithwiraj, et al. "GitLab: work where you want, when you want." *Journal of Organization Design* 9.1 (2020): 1-17.

[45]Katsura, Masaki, et al. "Model-based iterative reconstruction technique for radiation dose reduction in chest CT: comparison with the adaptive statistical iterative reconstruction technique." *European radiology* 22.8 (2012): 1613-1623.

[46]Ghani, Muhammad Usman, and W. Clem Karl. "CNN based sinogram denoising for low-dose CT." *Mathematics in Imaging*. Optical Society of America, 2018.

[47]Kim, Kwanyoung, Shakarim Soltanayev, and Se Young Chun. "Unsupervised training of denoisers for low-dose CT reconstruction without full-dose ground truth." *IEEE Journal of Selected Topics in Signal Processing* 14.6 (2020): 1112-1125.

[48]Shan, Hongming, et al. "Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction." *Nature Machine Intelligence* 1.6 (2019): 269-276.

[49]Li, Yajun, et al. "Normalization of multicenter CT radiomics by a generative adversarial network method." *Physics in Medicine & Biology* 66.5 (2021): 055030.

[50]Wei, Leihao, Yannan Lin, and William Hsu. "Using a generative adversarial network for ct normalization and its impact on radiomic features." *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020.

[51]Crombé, Amandine, et al. "Intensity harmonization techniques influence radiomics features and radiomics-based predictions in sarcoma patients." *Scientific reports* 10.1 (2020): 1-13.

[52]Selim, Md, et al. "Ct image harmonization for enhancing radiomics studies." *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021.

[53]Hosny, Ahmed, Hugo J. Aerts, and Raymond H. Mak. "Handcrafted versus deep learning radiomics for prediction of cancer therapy response." *The Lancet Digital Health* 1.3 (2019): e106-e107.

[54]Kim, Sangwook, et al. "Deep-Radiomics-Based Approach to the Diagnosis of Osteoporosis Using Hip Radiographs." *Radiology: Artificial Intelligence* (2022): e210212.

[55]LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436-444.

[56]Babenko, Artem, and Victor Lempitsky. "Aggregating local deep features for image retrieval." *Proceedings of the IEEE international conference on computer vision*. 2015.

[57]Li, Guanbin, and Yizhou Yu. "Visual saliency based on multiscale deep features." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[58]Lao, Jiangwei, et al. "A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme." *Scientific reports* 7.1 (2017): 1-8.

[59]Zheng, Xueyi, et al. "Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer." *Nature communications* 11.1 (2020): 1-9.

[60]Chen, Yen-Kung, et al. "Application of PET and PET/CT imaging for cancer screening." *Anticancer research* 24.6 (2004): 4103-4108.

[61]Mazurowski, Maciej A. "Radiogenomics: what it is and why it is important." *Journal of the American College of Radiology* 12.8 (2015): 862-866.

[62]Nishioka, Takeshi, et al. "Image fusion between 18FDG-PET and MRI/CT for radiotherapy planning of oropharyngeal and nasopharyngeal carcinomas." *International Journal of Radiation Oncology* Biology* Physics* 53.4 (2002): 1051-1057.

[63]Lovinfosse, Pierre, et al. "FDG PET/CT radiomics for predicting the outcome of locally advanced rectal cancer." *European journal of nuclear medicine and molecular imaging* 45.3 (2018): 365-375.

Chapter 8

Appendices

Appendix I Summary

English summary

Along with the increasing demand of low dose CT in clinical practices, low dose CT radiomics has shown its potential to provide clinical decision support in oncology. As a trade-off of low radiation exposure in low dose CT imaging, higher noise is present in these images. Noise in low dose CT decreases the texture information of image, and the reproducibility and performance of CT radiomics. One potential solution worth exploring for improving the reproducibility and performance of radiomics based on low dose CT is denoising the images before extracting radiomic features. As the state of art method for low dose CT denoising, generative models have been widely used in denoising practices. This thesis investigated the possibility of using generative models to enhance the image quality of low dose CTs and improve radiomics reproducibility and performance.

In the first research chapter **(Chapter 2)** of this thesis, we investigate the benefits of shortcuts in encoder-decoder network for CT denoising. An encoder-decoder network (EDN) is an important architecture for the generator in generative models and this chapter provides some guidelines to help us design generative models. Results showed that over half of the shortcuts are necessary for CT denoising. However, the network should keep sparse connection between the encoder and decoder. Moreover, deeper shortcuts have a higher priority to be removed in favor of keeping sparse connections.

Paired training datasets are needed for training most generative models. However, collecting these kinds of datasets is difficult and time-consuming. To investigate the effect of generative models in improving low dose CT radiomics reproducibility, **(Chapter 3)** two included generative models – Conditional Generative Adversarial Network (CGAN) and END - were trained on paired simulation low-high dose CT images. The trained models are applied to simulated noisy CT images and real low dose CT images. Results showed that denoising using EDN and CGANs can improve the reproducibility of radiomic features from noisy CTs (including simulated data and real low dose CTs).

To test the improvement of enhanced low dose CT radiomics in real applications more comprehensively, low dose CT radiomics was applied for a new application. **(Chapter 4)** The objective of this application is to develop a lung cancer classification model at the subject (patient) level from multiple examined nodules, without the need to have specific expert findings reported at the level of each individual nodule. Lung cancer classification was regarded as a multiple instances learning problem, CT radiomics were used as biomarkers to extract information from each nodule and deep attention-based MIL is used as the classification algorithm at the patient level. Results showed that the proposed method can achieve the best performance in lung cancer classification compared with other MIL methods and that the introduced attention mechanism can increase the interpretability of results.

To comprehensively investigate the improvements of generative models for CT radiomics performance in real applications, pre-trained generative models are applied into multiple real low dose CT datasets without fine-

tuning. **(Chapter 5)** Improved radiomics features were applied into multiple radiomics related applications – tumor pre-treatment survival prediction and deep attention-based MIL based lung cancer diagnosis. The results showed that generative models can improve low dose CT radiomics performance.

To investigate the possibility of using unpaired real low-high dose CT image to establish a denoiser and using thus trained denoiser to enhance radiomics reproducibility and performance, a Cycle GAN was adopted as the testing model in this chapter. **(Chapter 6)** The Cycle GAN was trained based on paired simulated datasets (for comparison study with EDN and CGAN) and unpaired real datasets. The trained models were applied to simulated noisy CT images and real low dose CT images to test the improvement of radiomics reproducibility and performance. The results showed that Cycle GANs trained on both simulated and real data can improve radiomics reproducibility and performance in low dose CT and achieve similar results compared to CGANs and EDNs

Finally, the discussion section of this thesis **(Chapter 7)** summarized the barriers that prevent generative models to be applied apply for real low dose CT radiomics and propose the possible solutions for these barriers. Moreover, this discussion section mentioned other possible methods to improve low dose CT radiomics performance.

Nederlandse samenvatting

Samen met de toenemende vraag naar lage dosis CT in de klinische praktijk, hebben radiomics gebaseerd op lage dosis CT potentieel om klinische besluitvormingsondersteuning te bieden in de oncologie. De lage blootstelling aan straling bij CT-beeldvorming met lage dosis betekent wel dat er meer ruis aanwezig is in deze afbeeldingen. Ruis in lage dosis CT vermindert de textuurinformatie van het beeld en de reproduceerbaarheid en prestaties van CT-radiomics. Een mogelijke oplossing die het onderzoeken waard is voor het verbeteren van de reproduceerbaarheid en prestaties van radiomics op basis van een CT met een lage dosis, is het verwijderen van ruis voordat de radiomic-kenmerken worden geëxtraheerd. De meest geavanceerde methode voor ruisverwijdering in CTs met een lage dosis, zijn generatieve modellen die op grote schaal gebruikt worden. Dit proefschrift onderzocht de mogelijkheid om generatieve modellen te gebruiken om de beeldkwaliteit van lage dosis CT's te verbeteren en de reproduceerbaarheid en prestaties van radiomics in lage dosis CT's te verbeteren.

In het eerste hoofdstuk (**Hoofdstuk 2**) van dit proefschrift onderzoeken we de voordelen van snelkoppelingen in het encoder-decodernetwerk voor CT-ruisonderdrukking. Een encoder-decoder netwerk (EDN) is een belangrijke architectuur voor het generator deel van generatieve modellen. Dit hoofdstuk kan enkele richtlijnen geven om ons te helpen bij het ontwerpen van generatieve modellen. Resultaten toonden aan dat meer dan de helft van de snelkoppelingen nodig zijn voor CT-ruisonderdrukking, maar het netwerk moet de schaarsere verbindingen tussen encoder en decoder behouden.

Gepaarde datasets zijn nodig voor het trainen van de meeste generatieve modellen, maar het verzamelen van dit soort datasets is duur en tijdrovend. Om het effect van generatieve modellen op het verbeteren van de reproduceerbaarheid van lage dosis CT radiomics te onderzoeken, (**Hoofdstuk 3**) zijn twee generatieve modellen - Conditional Generative Adversarial Network (CGAN) en END - getraind op gesimuleerde lage-hoge dosis CT beelden. Getrainde modellen worden toegepast op gesimuleerde CT-beelden met ruis en CT-beelden gemaakt met een daadwerkelijk lage dosis. De resultaten toonden aan dat ruisonderdrukking met behulp van EDN en CGAN's kan worden gebruikt om de reproduceerbaarheid van radiomische kenmerken van CT's met ruis (inclusief gesimuleerde gegevens en CT’s gemaakt met een lage dosis) te verbeteren.

Om de verbetering van ruis-onderdrukte lage dosis CT-radiomics in echte toepassingen te testen, werd meer uitgebreide, lage dosis CT-radiomics toegepast in een nieuwe toepassing. (**Hoofdstuk 4**) Het doel van deze toepassing is het ontwikkelen van een classificatiemodel voor longkanker op het niveau van de patiënt uit meerdere onderzochte nodules, zonder dat specifieke bevindingen van deskundigen op het niveau van elke afzonderlijk nodule moeten worden gerapporteerd. Classificatie van longkanker wordt beschouwd als een ”multi-instance learning (MIL)”, CT-radiomics kan worden gebruikt als biomarkers om informatie uit elke nodule te extraheren en diepgaande op aandacht gebaseerde MIL wordt gebruikt als het classificatie-algoritme op patiëntniveau. Resultaten toonden aan dat de voorgestelde methode de beste prestaties kan leveren bij de classificatie van longkanker in vergelijking met andere MIL-methoden en dat het

geïntroduceerde aandachtsmechanisme de interpreteerbaarheid van de resultaten kan vergroten.

Om de verbeteringen van generatieve modellen voor CT-radiomics prestaties in echte toepassingen uitgebreid te onderzoeken, werden vooraf getrainde generatieve modellen toegepast in meerdere lage dosis CT-datasets zonder fine-tuning. (**Hoofdstuk 5**) Verbeterde radiomic-functies werd toegepast in meerdere radiomic-gerelateerde toepassingen - overlevingsvoorspelling voor de behandeling van tumoren en diepgaande, op aandacht gebaseerde en op MIL gebaseerde longkankerdiagnoses. De resultaten toonden aan dat generatieve modellen de prestaties van lage dosis CT-radiomics kunnen verbeteren.

Om de mogelijkheid te onderzoeken van het gebruik van ongepaarde lage-hoge dosis CT-beelden om ruisonderdrukking vast te stellen en het gebruik van getrainde ruisonderdrukking om de reproduceerbaarheid en prestaties van radiomics te verbeteren, werd Cycle GAN in dit hoofdstuk als testmodel onderzocht. (**Hoofdstuk 6**) Het Cycle GAN model werd getraind op basis van gepaarde gesimuleerde datasets (voor vergelijkingsonderzoek met EDN en CGAN) en ongepaarde echte datasets. De getrainde modellen werden toegepast op gesimuleerde CT-beelden met ruis en CT-beelden met een lage dosis om de verbetering van de reproduceerbaarheid en prestaties van radiomics te testen. De resultaten toonden aan dat Cycle GANs die zijn getraind op zowel gesimuleerde als echte gegevens de reproduceerbaarheid en prestaties van radiomics in lage dosis CT kunnen verbeteren en vergelijkbare resultaten kunnen bereiken als CGAN’s en EDN's.

Tot slot, geeft de discussiesectie van dit proefschrift (**Hoofdstuk 7**) een samenvatting van de belemmeringen die bij generatieve modellen kunnen worden tegengekomen als zij worden toegepast in lage dosis CT-radiomics en stelt de mogelijke oplossingen voor deze belemmeringen voor. Bovendien geeft de discussiesectie de andere mogelijke methoden om de prestaties van lage dosis CT-radiomics te verbeteren.

Appendix II Impact Paragraph

CT radiomics has the potential to provide clinical decision support in oncology due to the wide us of CT scanning in clinical practive. [1] Due to the long-term risk posed by ionizing radiation exposure, low dose CTs have become more popular (according to the As Low As Reasonably Achievable (ALARA) principle [2]) in clinical practice, especially for screening and monitoring of populations at risk. Radiomics from low dose CT might be an effective tool for quicker and more reliable screening. [3] However, as a consequence of the low radiation exposure in low dose CT imaging, noise in such images is more pronounced and this noise decreases the reliability and performance of radiomics. Improving the reproducibility of radiomics and its performance in clinical applications from low dose CTs is therefore a timely and potentially impactful research topic.

One potential solution worth exploring for improving the reproducibility and performance of radiomics based on low dose CT is denoising the images before extracting radiomic features. As the state of art low dose CT denoising method, generative models are used as denoisers in this thesis to improve low dose CT radiomics reproducibility and performance. These studies may bring certain scientific and social impacts.

Scientific impacts

1. All our studies are published in international peer-reviewed journals (such as: Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, Physics in Medicine & Biology, Medical Physics, Medica Physica, Journal of Applied Clinical Medical Physics).

2. All our studies are available as open access publications.

3. **Chapter 2** investigates the beneficials of shortcuts in encoder-decoder network for CT denoising, results provided guidelines for network designing in denoising task.

4. **Chapter 3** is the first effort to improve the reproducibility of radiomic features calculated on low dose CT scans by applying generative models.

5. **Chapter 4** introduces a new lung cancer diagnosis method; this method achieves a good performance in classification with a higher interpretability.

6. **Chapter 5** is the first effort to improve the performance of radiomics-based models from features extracted from low dose CT scans.

7. **Chapter 6** assess the potential of using cycle GAN to denoise low dose CTs and improve radiomics reproducibility and performance.

Social impacts

1. Codes and important pre-trained generative models of this thesis are available for public use as open source; we hope that these can help medical physicists and other care professionals to remove barriers for applying generative models for low dose CT radiomics.

2. Our proposed lung cancer diagnosis solution can improve the detection and management of early lung cancer, and we hope our method can reduce the mortality of lung cancer for patients.

3. Improving low dose CT radiomics performance may reduce ionizing radiation exposure for patients and therefore reduce the number of cancers and other diseases caused by this exposure.

References:

[1] Ibrahim, Abdalla, et al. "Radiomics analysis for clinical decision support in nuclear medicine." *Seminars in nuclear medicine*. Vol. 49. No. 5. WB Saunders, 2019.

[2] Musolino, Stephen V., Joseph DeFranco, and Richard Schlueck. "The ALARA principle in the context of a radiological or nuclear emergency." *Health physics* 94.2 (2008): 109-111.

[3] Homayounieh, Fatemeh, et al. "Prediction of coronary calcification and stenosis: role of radiomics from Low-Dose CT." *Academic Radiology* 28.7 (2021): 972-979.

Appendix III Acknowledgements

I received lots of help and support in the past several years. Without these, this thesis would not be possible and I would like to express my gratitude. I would like to thank for the financial support from the China Scholarship Council (File No. 201906540036) during the last 3 years.

I would like to thank the assessment committee of my thesis -Prof. Joachim Wildberger, Prof. Bram van Ginneken, Dr. Wouter van Elmpt, Prof. Nico van den Berg. Thank you very much for spending time checking my thesis and for your comments on my thesis!

I would like to extend my gratitude to my supervisor Prof. dr Andre Dekker for giving me the opportunity to work and study here. Thank you for your help to revise my manuscripts and my thesis, your valuable comments significantly improve my manuscripts and thesis.

I would like to express my sincere gratitude to my co-supervisor Dr. Inigo Bermejo. We worked closely for all my jobs during last 3 years, my poor English introduced heavily editing work for you, you provided priceless help for me to improve my manuscripts, language and research ability. Your kindly advises corrected my attitude about doing research and made me more confident for complex questions. Weekly meetings with you let me know there is another one to share my pressure and happiness with.

I would like to extend my gratitude to my co-supervisor Dr. Leonard Wee. You let me know what a gentleman should look like, you are always so friendly to everyone you have met and so patient to every questions you have

encountered. You provided the opportunity for me to visit University of Southern Denmark with financial support from the Yerun research mobility grant and provided precious comments to improve my manuscripts.

I would like to thank my colleagues in the lab: Dr. Alberto Traverso, Dr. Zhenwei Shi, Dr. Ivan Zhovannik, Ms Haiyan Zeng, Mr. Chong Zhang, Mr. Shenlun Chen, Mr. Zhixiang Wang, Mr Zhen Zhang, etc. Thank you for your help during last three years.

I would like to thank my friends in Maastricht: Mr. Shuhe Zhang, Ms. Min Wu, Ms. Shunxin Jin, Mr. Letao Li, Ms. Jinnan Huang. It is my pleasure to meet you here, hope you are happy in the future.

I'm deeply indebted to my family! 感谢我父母给予我无私的不求回报的爱，血浓于水，我相信亲子关系是这个世界上最牢不可破的纽带，谢谢你们，愿你们身体健康。感谢我的女友在我即将完成学业时来到我的生活中，希望我们能早日成为家人。

最后，我谢谢自己，谢谢自己过去 20 多年如一日的坚守，让自己保持进步，尽全力远离愚蠢。

Appendix IV Curriculum Vitae

Junhua Chen was born on July 5th, 1994 in Shaoxing City of Zhejiang Province, China. He finished his high school education in Shaoxing Shengzhou No.1 Middle School and was admitted to Wenzhou Medical University in Wenzhou City of Zhejiang Province in 2012 to study Biomedical Engineering. After obtaining his bachelor's degree in 2016, he was enrolled in a 3-year Master program of Biomedical Engineering in Beijing University of Technology in Beijing. He worked on improving coronary artery centerline extraction algorithm based on tracking in X-ray coronary angiography and was awarded the National Scholarship for Graduate Students in 2018.

Right after acquiring a master's degree in 2019, he was awarded a scholarship from the China Scholarship Council to fund his PhD training in Maastricht University, under the supervision of Prof. dr. Andre Dekker, Dr. Inigo Bermejo and Dr. Leonard Wee. During his PhD training he worked on improving low dose CT radiomics reproducibility and performance by using generative models.

Appendix V List of Publications

Publications in This Thesis

Chen, J., Zhang, C., Wee, L., Dekker, A., & Bermejo, I. (2022). Are all shortcuts in encoder–decoder networks beneficial for CT denoising?. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 1-8. DOI: 10.1080/21681163.2022.2044908

Chen, J., Zhang, C., Traverso, A., Zhovannik, I., Dekker, A., Wee, L., & Bermejo, I. (2021). Generative models improve radiomics reproducibility in low dose CTs: a simulation study. *Physics in Medicine & Biology*, 66(16), 165002. DOI: 10.1088/1361-6560/ac16c0

Chen, J., Zeng, H., Zhang, C., Shi, Z., Dekker, A., Wee, L., & Bermejo, I. (2022). Lung cancer diagnosis using deep attention ‐ based multiple instance learning and radiomics. *Medical Physics*, 49(5), 3134-3143. DOI: 10.1002/mp.15539

Chen, J., Bermejo, I., Dekker, A., & Wee, L. (2022). Generative models improve radiomics performance in different tasks and different datasets: An experimental study. *Physica Medica*, 98, 11-17. DOI: 10.1016/j.ejmp.2022.04.008

Chen, J., Wee, L., Dekker, A., & Bermejo, I. (2022) Improving Reproducibility and Performance of Radiomics in Low Dose CT using Cycle GANs. *Journal of Applied Clinical Medical Physics*, 23(10), e13739. DOI: 10.1002/acm2.13739

Other Publications

Chen, J., Wee, L., Dekker, A., & Bermejo, I. (2023) Using 3D deep features from CT scans for cancer prognosis based on a video classification model: A multi-dataset feasibility study. *Medical Physics*, 1-14. DOI: 10.1002/mp.16430

Chen, J.#, Chen, S.#, Wee, L., Dekker, A., & Bermejo, I. (2023) Deep Learning Based Unpaired Image-to-Image Translation Applications for Medical Physics: A Systematic Review. *Physics in Medicine & Biology*, 68(5), 05TR01. DOI: 10.1088/1361-6560/acba74 (#co-first author)

Chen, J., Ke, D., Wang, Z., & Liu, Y. (2018). A high splicing accuracy solution to reconstruction of cross-cut shredded text document problem. *Multimedia Tools and Applications*, *77*(15), 19281-19300. DOI: 10.1007/s11042-017-5389-z

Chen, J., Yang, Y., Tian, M., Qi, X., & Liu, Y. (2018). Coronary Artery Centerline Tracking and Coronary Artery Tree Segmenting in X-ray Angiographic Images. *Journal of Medical Imaging and Health Informatics*, *8*(6), 1226-1232. DOI: 10.1166/jmihi.2018.2470

Chen, J., Tian, M., Qi, X., Wang, W., & Liu, Y. (2019). A solution to reconstruct cross-cut shredded text documents based on constrained seed K-means algorithm and ant colony algorithm. *Expert Systems with Applications*, *127*, 35-46. DOI: 10.1016/j.eswa.2019.02.039

Chen, J., Qi, X., Wang, W., Li, B., & Liu, Y. (2020). Real-time location of surgical incisions in cataract phacoemulsification. *Multimedia Tools and Applications*, *79*(41), 30311-30327. DOI: 10.1007/s11042-020-09560-8