Original paper

Generative models improve radiomics performance in different tasks and different datasets: An experimental study

Junhua Chen *, Inigo Bermejo, Andre Dekker, Leonard Wee

Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre +, Maastricht 6229 ET, Netherlands

A R T I C L E   I N F O
Keywords:
Radiomics
Generative models
Image denoising
Comparative study

A B S T R A C T

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.

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 [8–10]. Radiomics has also garnered attention in the field of radiotherapy, where it is known as dosimomics [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 [15–18]. 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 [27–29]. 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 Fig. 1 summarizes our study methodology.

Denosing models’ development

Based on [37], we selected two generative models – encoder-decoder...
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 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-CDSS/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 Fig. 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 Fig. 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 8 GB 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 Fig. 2.

### 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 1 and Fig. 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].

### 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 2 and Fig. 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.
**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 Fig. 3 and Fig. 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 Fig. 4 (c) than Fig. 3. 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.
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 Fig. 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

---

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

| Training length | Without Denoising | 25 Epochs | 50 Epochs | 75 Epochs | 100 Epochs |
|-----------------|-------------------|-----------|-----------|-----------|------------|
| Encoder-decoder |                   |           |           |           |            |
| AUC             | 0.842 ± 0.071     | 0.883 ± 0.077 | 0.844 ± 0.067 | 0.823 ± 0.067 | 0.866 ± 0.070 |
| p-value*        | < 0.01            | 0.86      | 0.07      | 0.02      |            |
| CGAN            |                   |           |           |           |            |
| AUC             | 0.894 ± 0.056     | 0.863 ± 0.066 | 0.837 ± 0.085 | 0.866 ± 0.056 |
| 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.

---

Fig. 3. Experimental results (AUC) of survival prediction task.

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

Conclusion

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.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

JC is supported by a China Scholarship Council scholarship (201906540036). The remaining authors acknowledge funding support from the following: STRaTegy (STW 14930), BIONIC (NWO 629.002.205), TRAIN (NWO 629.002.212), CARRIER (NWO 628.011.212) and a personal research grant by The Hanarth Funds Foundation for LW.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jemp.2022.04.008.

References

[1] Avanzo M, Porzio M, Lorenzo L, Milan L, Sghedoni R, Russo G, et al. Artificial intelligence applications in medical imaging: a review of the medical physics research in Italy. Physica Med 2021;83:221–41.
[2] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5(1).
[3] Dessent M-C, Tixier F, Weber WA, Siegel BA, Rest CCL, Viwikis D, 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. J Nucl Med 2017;58(3):406–11. https://doi.org/10.2907/jnumed.118.067792.
[4] Bogowicz M, Rießerter O, Bundeuh B, Veit-Haibach P, Hällner M, Studer G, et al. Stability of radiomic features in CT perfusion maps. Phys Biol 2016;13(24):8736–49.
[5] Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Viwikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18 F-FDG PET. J Nucl Med 2012;53(5):693–700.
[6] Zhang B, Tian J, Dong Di, Gu D, Dong Y, Zhang Lu, et al. Radiomics features of multivariate MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. Clin Cancer Res 2017;23(15):4259–69.
[7] Placidi, Lorenzo, Eliana Giacosa, Cristina Garibaldi, Davide Maestri, Raffaella Massafr et al., A multicentre evaluation of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. Clin Cancer Res 2017;23(15):4259–69.
[8] Comes, Maria Colomba, Davide Maestri, Raffaella Massafra et al., A multicentre evaluation of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. Clin Cancer Res 2017;23(15):4259–69.
[9] Corcos, Maria Colomba, Daniele La Forgia, Vittorio Didonna, Annarita Fanizzi, Davide Maestri, Raffaella Massafra et al., A multicentre evaluation of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. Clin Cancer Res 2017;23(15):4259–69.
[10] comes, Maria Colomba, Daniele La Forgia, Vittorio Didonna, Annarita Fanizzi, Francesco Giotta, Agnese Latorre, Evgueni Martinelli et al., Early prediction of neoadjuvant chemotherapy response by exploiting a transfer learning approach on breast DCE-MRIs. Sci Rep 2021;11(1).
[11] La Forgia, Daniele, Angela Vestito, Maurilia Lasciutari, Maria Colomba Comes, Sergio Diotaiuti, Francesco Giotta, Agnese Latorre et al., Response predictive 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.
[12] Mussolini, Stephen V., Joseph DeFranco, and Richard Schlaefer. ‘The ALARA principle in the context of a radiological or nuclear emergency.’ Health Phys. 94 (2) (2008): 109–111. https://doi.org/10.1097/01.HIP.0000285801.87304.3f.
[13] Bil, Wenyu Linda, Ahmed Hosny, Matthew B. Schabath, Maryellen L. Giger, Nicolai J. Birkbak, Alireza Mehratas, 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.
[14] Gillis RJ, Schabath MB. Radiomics improves cancer screening and early detection. Cancer Epidemiol Biomarkers Prev 2020;29(12):2556–67.
[15] Choi W, Oh JH, Riyahy S, Liu C-J, Jiang F, Chen W, et al. Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer. Med Phys 2018;45(4):1537–49.
[16] Homayounieh F, Yan P, Digumarthy SR, Kruger U, Wang Ge, Kalra MK. Prediction of coronary calcification and stenosis: role of radiomics from Low-Dose CT. Acad Radiol 2021;28(7):972–9. https://doi.org/10.1016/j.acra.2020.09.021.

Table 3

Accuracy and recall for lung cancer diagnosis.

| Metrics                        | Encoder-decoder network | 0 Epochs | 25 Epochs | 50 Epochs | 75 Epochs | 100 Epochs |
|-------------------------------|-------------------------|----------|-----------|-----------|-----------|------------|
| Accuracy                      | 0.807 ± 0.069           | 0.818 ± 0.077 | 0.792 ± 0.061 | 0.75 ± 0.067 | 0.796 ± 0.067 |
| p-value                        | –                      | 0.70      | 0.10      | < 0.01    | 0.26      |
| Recall                        | 0.870 ± 0.061           | 0.829 ± 0.096 | 0.831 ± 0.089 | 0.810 ± 0.097 | 0.848 ± 0.072 |
| p-value                        | –                      | < 0.01    | < 0.01    | < 0.01    | 0.02      |
| CGAN                          | –                      | 0.780 ± 0.068 | 0.779 ± 0.085 | 0.776 ± 0.074 | 0.798 ± 0.064 |
| Accuracy                      | –                      | 0.01      | 0.01      | < 0.01    | 0.12      |
| p-value                        | –                      | 0.802 ± 0.091 | 0.774 ± 0.104 | 0.827 ± 0.079 | 0.811 ± 0.079 |
| p-value                        | –                      | < 0.01    | < 0.01    | < 0.01    | < 0.01    |
| Encoder-decoder network versus CGAN (p-values) | – | 0.04      | 0.04      | 0.21      | 0.01      | 0.18        | 0.67 | < 0.01 |

* Compared with results from original radiomics.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jemp.2022.04.008.
