IOP Publishing Phys. Med. Biol. 65 (2020) 015005 (9pp) https://doi.org/10.1088/1361-6560/ab5427

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

RECEIVED
28 June 2019

REVISED
18 October 2019

ACCEPTED FOR PUBLICATION
4 November 2019

PUBLISHED
10 January 2020

Automatic classification of dental artifact status for efficient image veracity checks: effects of image resolution and convolutional neural network depth

Mattea L Welch[1,5,7], Chris McIntosh[4,5,7], Tom G Purdie[2,4,5,7], Leonard Wee[6], Alberto Traverso[6], Andre Dekker[6], Benjamin Haibe-Kains[1,7,8,9] and David A Jaffray[1,2,3,4,5,7]

1 Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada
2 Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada
3 IBBME, University of Toronto, Toronto, Ontario, Canada
4 Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto, Ontario, Canada
5 The Techna Institute for the Advancement of Technology for Health, Toronto, Ontario, Canada
6 Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, The Netherlands
7 Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada
8 Ontario Institute of Cancer Research, Toronto, Ontario, Canada
9 Vector Institute, Toronto, Ontario, Canada

E-mail: mattea.welch@rmp.uhn.ca

Keywords:automation, deep learning, dental artifacts, CT imaging, quality classification

Supplementary material for this article is available online

Abstract

Enabling automated pipelines, image analysis and big data methodology in cancer clinics requires thorough understanding of the data. Automated quality assurance steps could improve the efficiency and robustness of these methods by verifying possible data biases. In particular, in head and neck (H&N) computed-tomography (CT) images, dental artifacts (DA) obscure visualization of structures and the accuracy of Hounsfield units; a challenge for image analysis tasks, including radiomics, where poor image quality can lead to systemic biases. In this work we analyze the performance of three-dimensional convolutional neural networks (CNN) trained to classify DA statuses. 1538 patient images were scored by a single observer as DA positive or negative. Stratified five-fold cross validation was performed to train and test CNNs using various isotropic resampling grids (64^3, 128^3 and 256^3), with CNN depths designed to produce 32^3, 16^3, and 8^3 machine generated features. These parameters were selected to determine if more computationally efficient CNNs could be utilized to achieve the same performance. The area under the precision recall curve (PR-AUC) was used to assess CNN performance. The highest PR-AUC (0.92 ± 0.03) was achieved with a CNN depth $= 5$, resampling grid $= 256$. The CNN performance with 256^3 resampling grid size is not significantly better than 64^3 and 128^3 after 20 epochs, which had PR-AUC $= 0.89 \pm 0.03$ (p-value $= 0.28$) and 0.91 ± 0.02 (p-value $= 0.93$) at depths of 3 and 4, respectively. Our experiments demonstrate the potential to automate specific quality assurance tasks required for unbiased and robust automated pipeline and image analysis research. Additionally, we determined that there is an opportunity to simplify CNNs with smaller resampling grids to make the process more amenable to very large datasets that will be available in the future.

Introduction

Good clinical practice and scientific developments in oncology are enabled by high quality data (Bray and Parkin 2009). Imaging data in particular has potential for large variations in quality, which has led to the development of standardized site specific imaging guidelines (Olliff *et al* 2014, Lewis-Jones *et al* 2016). The National Cancer Institute (NCI) has also recognized the importance of data quality, with multiple projects defining quality

assurance protocols for national and clinical screening trials (Moore *et al* 2005, Cagnon *et al* 2006). Traditional data quality requirements and detection methods are sufficient for most research questions and controlled clinical trials; however, the number of data quality 'rules' increases as we move towards Big Data methodology and data-driven science in Radiology (Raghupathi and Raghupathi 2014, Kansagra *et al* 2016). Detecting and fixing the issues also becomes more challenging, and in some cases may not be feasible or appropriate.

Digitization of patient health information and images provides opportunities for integration of big data and automation into all aspects of patient care. In particular, the area of automated information generation through quantitative analysis of medical images for detection and prognostic modeling has seen immense interest in recent years. This field is being referred to as radiomics (Gillies *et al* 2015, Lambin *et al* 2017) and leverages past pattern recognition and computer vision research (Hall *et al* 1971, Harlow and Eisenbeis 1973) to develop prognostic and predictive models based on image intensity values. However, radiomic features extracted from images without an understanding of data quality may incorrectly assign causality to features and signatures, rendering results unusable (Welch and Jaffray 2017). To safeguard against this, it was suggested by our group in a previous publication that greater understanding of image quality prior to utilization in radiomic pipelines is needed (Welch *et al* 2019). In that work we discovered that the performance of a well cited radiomic signature did not depend on the image intensity values and only required the patient's tumour contour. Data quality curation would safeguard these methods, thereby representing a fundamental step towards reliable and reproducible results.

A common contributor to poor data quality in head and neck (H&N) computed tomography (CT) images are dental artifacts (DA). Although clinical metal artifact reduction (MAR) techniques (Diehn *et al* 2017) and new deep learning methods for metal artifact suppression (Zhang and Yu 2018, Huang *et al* 2018, Hu *et al* 2019) exist, salvaging incorrect HU data in H&N image can be a challenging task. This is due to new artifacts being introduced into the image (Block *et al* 2016), and uncertainty as to whether the new resulting voxel information is representative of the actual patient phenotype that was masked by the artifact initially. Therefore, most radiomics studies choose to remove patient image volumes impacted by DAs; a method which was recently proven by Wei *et al* to improve prediction performance of a radiomic signature (Wei *et al* 2019). Alternatively, some researchers choose to remove image slices from the overall imaging volume that contain visible Das (Ger *et al* 2018, Elhalawani *et al* 2018); granted, justification would be required to explain the implications of slice removal on shape and texture features. In both of these cases user intervention is required to decide whether the patient is appropriate for inclusion in the study. This is currently a feasible, yet time consuming task in the recent proof-of-concept radiomic studies. However, as more retrospective data becomes available, and more sites begin to share their data, a method of automated patient classification would increase efficiency and reduce subjectivity of these important radiomic pipeline steps.

Until now research into DAs has focused on their reduction. However, recent radiomics publications have shown an interest in their classification, permitting operators to decide whether the patient's image is appropriate for inclusion. In Wei *et al* (2019) features extracted from regions of interest were capable of classifying DA+ patients with an AUC of 0.89. Oh *et al* (2019) developed a method for classification of DA+ slices that performs with a prediction rate of 97.10% and 74.10% for DA+ and DA− image slices, respectively. Both of the mentioned classification methodologies required the definition of a region of interest and the extraction of features, as well as operator interaction. Convolutional neural networks (CNN) (Krizhevsky *et al* 2012) are an alternative machine driven method that can provide an automated classification, and have shown promise for identifying motion artifacts in various magnetic resonance imaging (MRI) types (Kelly *et al* 2016, Graham *et al* 2018). Although we are interested in a different data quality issue, these results motivate the usage of similar techniques for the classification of DA artifacts in H&N CT imaging.

An automated process designed to flag images as 'dental artifact positive' would increase efficiency and subjectivity of these tedious, but important, tasks. After classification users could then decide if an image should be removed or included in a specific study based on artifact magnitude or interference of the artifact with a region-of-interest. This work aims to demonstrate that a CNN can be trained and validated using H&N data to classify CT images as those with and without DAs, while exploring the impact of CNN depth and image resolution on prediction performance.

Methods

Dataset and dental artifact classification

Training and testing of the model utilized 1538 H&N planning CTs from the Princess Margaret Cancer Centre. Image details can be found in table 1. We converted the images from DICOM to nearly raw raster data (nrrd) to enable processing by the CNN using Python and the SimpleITK library (Yaniv *et al* 2017). This conversion removes meta-data related to patient, institution and scanner, but retains meta-data regarding resolution, size, center and directionality of the image. nrrd format was chosen for this study, but any data format that can be

Table 1. Dataset details outlining the imaging year range, tube voltage peak, scanner types, and median number of slices, thickness and resolution for all images.

					Scanner Type		
Median slice thickness and range (mm)	Median slice num. and range	Median pixel size and range (mm)	Imaging year range	Tube voltage peak (kVp)	Toshiba	Philips	GE medical
2 (1–4)	182 (130–330)	0.98 (0.61–2.00)	2010–2019	120	1223	257	58

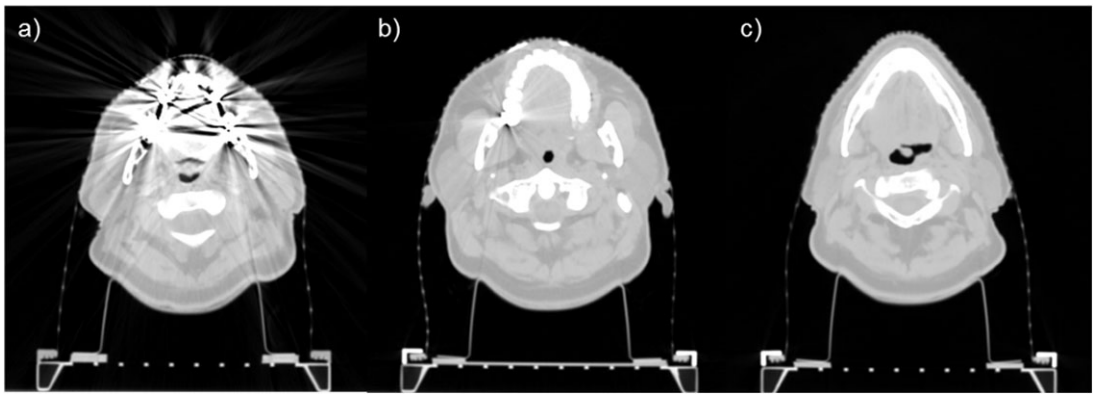

Figure 1. Three patient images showing examples of DA+ and DA− image slices. (a) Example image slice from CT volume with prominent artifacts classified as DA+, (b) example image slice from CT volume with less artifact interference classified as DA+, and (c) example image slice from volume with no artifacts classified as DA−. Window width and level were set to 1346 and −325, respectively.

loaded as a SimpleITK object is compatible with this methodology. This process was done automatically for all CT volumes. A single observer with 8 years of medical imaging experience then scored the DA status of each patient's converted CT volume as DA present/positive (DA+, status = 1) or DA absent/negative (DA−, status = 0); the magnitude of the DA was not considered. For example both (a) and (b) in figure 1 were scored with a status of 1, despite having different impacts on HU across the entire image.

Data preprocessing

Prior to training or testing of our CNN, image volumes were processed using a multistep procedure: (1) CT volumes were interpolated to iso-tropic voxel sizes of 1 mm^3 using the SimpleITK linear resampling image filter to reduce variability within the images, thereby improving processing by the CNN; (2) 10% of the training CT volumes were randomly selected for cropping and 60% for flipping. Images selected for cropping were reduced three quarters of their original size in all directions, while maintaining the image center. This cropping size was selected to ensure that cropping of the dental artifacts was highly unlikely. Images selected for flipping were flipped using Python's NumPy (Oliphant 2006) function 'flip' in the 'left/right' direction to create a mirror image. This data augmentation introduces a form of uncertainty into the training data to improve generalizability of the model; (3) CT volumes were padded to a uniform size to maintain the aspect ratio of the volume during resizing that occurs in step 4. Python's NumPy function 'pad' was used. The largest dimension of the 3D image array (width, height or slice number) was found, and all other dimensions were padded using zeros; (4) CT image volumes were resized to determine the impact of various resampling grids on CNN performance, and to generate a uniform volume more conducive with CNN training. Resizing was performed using the open-source scikit-image library (van der Walt *et al* 2014), which preserves the image's HU distribution. For our work, resampling grids of 256, 128 and 64 voxels were analyzed for performance. Examples of image slices at these different resampling grids can be found in figure 2.

Model training

For this study we used the open-source python library, PyTorch (Paszke *et al* 2017), to train our three-dimensional CNN. A virtual machine from VMware, Inc. with 10 Intel Xeon CPU E5-2690 processors and a NVIDIA Tesla K40m GPU was used for training and testing.

Stratified five-fold cross validation was used for training and testing. Batches of 14 images randomly selected from the training data without replacement were fed into the CNN. Batch normalization and rectified linear unit functioning (ReLU) were present on all convolutional layers and max pooling was used on all convolutional layers except the final one (figure 3) (Nielsen 2015, LeCun *et al* 2015); average pooling was used on the outputs of the final convolutional layer, followed by a fully connected layer and softmax classification. A convolutional

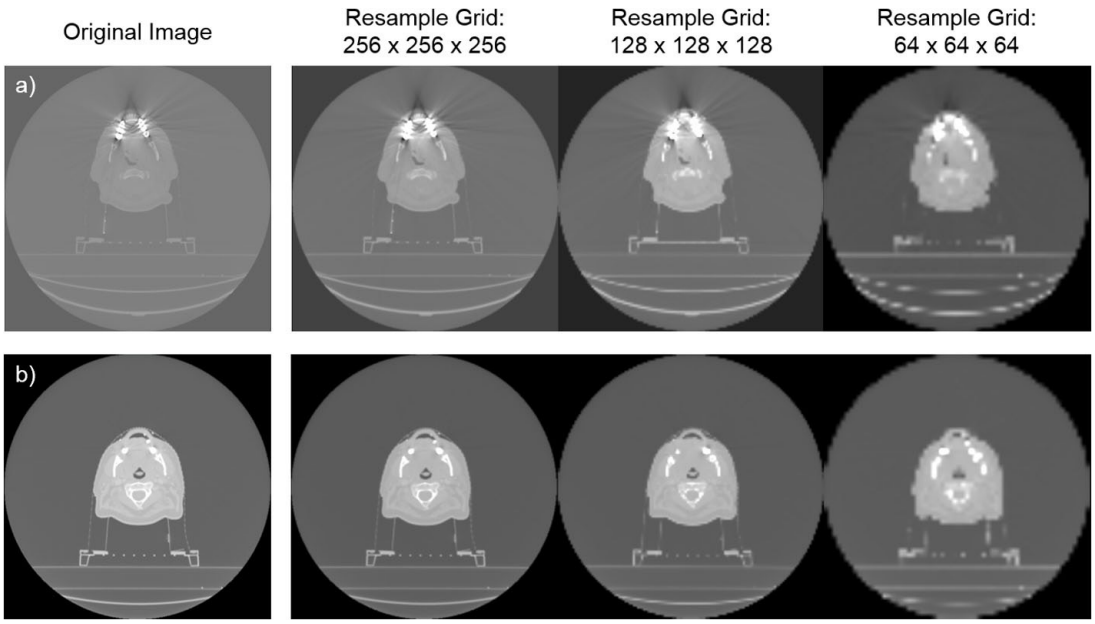

Figure 2. Example of image slices at various resampling grids of interest. (a) shows a DA with a large magnitude, (b) shows a DA with a small magnitude.

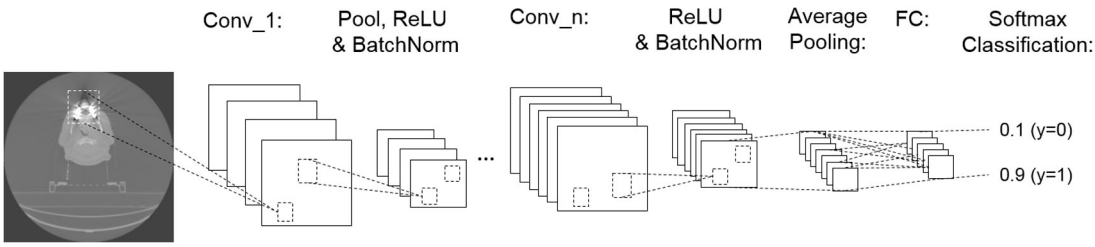

Figure 3. Example schematic of the DA status CNN. For simplicity, the training is shown here with a 2D image. Batches of 14 images were input into the network. The output of the convolutional layers underwent BatchNormalization, rectified linear unit (ReLU) and pooling. The final layer of the CNN only underwent BatchNormalization, ReLU and average pooling. A single fully connected layer was followed by a softmax classification which returned probabilities that a given image was DA+ or DA−. The ellipses (…) indicates the additional convolutional layers that are added as a function of the image resolution.

kernel size of 5 with a padding of 2 was used on the first convolutional layer, all subsequent layers used a kernel size of 3 with a padding of 1. Weighted optimization was used to account for uneven class distribution. CNNs were trained for the various resampling grids (256^3, 128^3 and 64^3) and depths (1, 2, 3, 4, and 5) of interest. Three depths were analyzed for each resampling grid size, whereby the final machine generated features fed into the fully connected layer were of size 32^3, 16^3 and 8^3, respectively. The kernel used in the final CNN layers was of size 3, therefore 8^3 was chosen as the smallest machine generated feature size to ensure adequate sampling of the convolutional output layers by the kernel. Details of input and output sizes used for the different depths and resampling grids are found in table 2. Training was performed for 20 epochs based on knowledge regarding model convergence gained through unpublished studies completed by the authors. Model training for each resampling grid and depth was repeated five separate times using different splits of the data.

Model evaluation

Models were evaluated every 5 epochs for performance on both training and hold-out test datasets. During evaluation of a CNN, each image volume from a dataset was fed through the CNN to obtain the model's softmax prediction of DA status. The Area Under the Precision Recall curve (PR-AUC) was calculated for the training and testing datasets across all five-folds. The PR curve summarizes the precision and recall of a given predictive model. Precision describes the ratio of the number of true positives divided by the sum of true positives and false positives, while recall describes the ratio of the number of true positives divided by the sum of true positives and false negatives of a predictive model. PR curves are more sensitive to class imbalances and therefore provide a better metric for our study which is heavily imbalanced towards DA+ images. The PR-AUC was calculated using Python's Sci-kit learn library (Pedregosa *et al* 2011).

Table 2. Details of the number of convolutional layers, sizes of convolutional layers input (IN) and outputs (OUT), and size of the fully connected layer are given in this table relative to the resampling grids used. The number of layers used was a function of the resampling grid size, and the experimental design that stated the size of the fully connected layer would be 8^3, 16^3 and 32^3.

		Conv_1		Conv_2		Conv_3		Conv_4		Conv_5		Fully connected layer feature size
Resampling grid	Depth	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	
256^3	3	1	4	4	8	8	16	N/A	N/A	N/A	N/A	32^3
	4	1	4	4	8	8	16	16	32	N/A	N/A	16^3
	5	1	4	4	8	8	16	16	32	32	64	8^3
128^3	2	1	4	4	8	N/A	N/A	N/A	N/A	N/A	N/A	32^3
	3	1	4	4	8	8	16	N/A	N/A	N/A	N/A	16^3
	4	1	4	4	8	8	16	16	32	N/A	N/A	8^3
64^3	1	1	4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	32^3
	2	1	4	4	8	N/A	N/A	N/A	N/A	N/A	N/A	16^3
	3	1	4	4	8	8	16	N/A	N/A	N/A	N/A	8^3

Furthermore, the performance of the CNNs was compared to a more simplistic baseline method. This method generated regions of interest (ROI) for each image in the dataset by thresholding above 2000 HU, capturing voxels with a density greater than bone. The values were then sorted based upon volume of the resulting ROI and used as a predictor for the presence of DA to calculate a PR-AUC.

Results

Manual classification of dental artifact status by a single observer resulted in 1092 DA+ and 446 DA− classifications, 71% and 29%, respectively. The simplistic baseline classification using volume of the high density ROI had a PR-AUC of 0.73.

A resampling grid of 256^3 with CNN depth of 5 resulted in the highest overall PR-AUC of 0.92 ± 0.03 calculated across all five-folds of the test datasets, as shown in table 3 and figure 4(c)). The average precision and recall for this CNN across the five-folds of test datasets were 0.96 ± 0.03 and 0.90 ± 0.05, respectively.

The highest average PR-AUC values for resampling grids 64^3 (PR-AUC $= 0.89 \pm 0.03$) and 128^3 (PR-AUC $= 0.91 \pm 0.02$) occurred after 20 epochs at depths of 3 and 4, respectively (table 3 and figures 4(a) and (b)). We cannot conclude from these results that a CNN trained and tested with a resampling grid size of 256^3 performs significantly better than with a resampling grid of 64 (p-value $= 0.28$) or 128 (p-value $= 0.93$), indicating that our CNN performance is reasonably consistent regardless of resampling grid size. PR-AUC values for individual folds of all resampling grids and CNN depths can be found in the supplementary material (stacks.iop.org/PMB/65/015005/mmedia).

Approximate training time per epoch with a resampling grid size of 256^3 and depth of 5 is 70 min, 128^3 and 4 is 40 min, and 64^3 and 3 is 30 min. Approximate CNN prediction times with a resampling grid sizes of 256^3 and depth of 5 is 4 s, 128^3 and 4 is 2.2 s, and 64^3 and 3 is 1.5 s.

Discussion

Data quality is integral for the future of automated pipelines and processes. Integration of a diverse set of disease and host factors, including imaging data, is expanding the volume, variety, velocity and veracity of measurement data. As the big data paradigm approaches clinical cancer management, utilization of efficient quality assurance methods become of paramount importance. In this work we trained CNNs to predict DA status of H&N patients from the Princess Margaret Cancer Institute. A CNN with depth 5 trained and validated on images resampled with a grid of 256^3 had the highest PR-AUC of 0.92 ± 0.03. PR-AUC values for resampling grids of 64^3 with depth 3 (PR-AUC $= 0.89 \pm 0.03$, p-value $= 0.28$) and 128^3 with depth 4 (PR-AUC $= 0.91 \pm 0.02$ p-value $= 0.93$) were not significantly different than our best performing CNN. These models demonstrate the potential to increase the efficiency of data quality checks in radiation oncology, thereby improving automated pipelines and processes important to patient prognosis and treatment.

Our CNNs were capable of classifying DA statuses from single institution images effectively, even when utilizing small resampling grid sizes. The ability to use a simplified CNN with smaller resampling grid sizes speeds up training and predictions, making it more amenable to very large datasets that will be available in the future. It is also important to note that although training time may be lengthy it often occurs as an offline task. For these types of models, which are designed for automation of routine tasks, prediction time is most important. For our

Table 3. Average training and testing PR-AUC values for all resampling grids and CNN depths. Averages and standard deviations (STDEV) are calculated across the five-folds of model training and testing.

		PR-AUC and STDEV at 20 epochs	
Resampling grid	Depth	Testing	Training
256^3	3	0.80 ± 0.09	0.99 ± 0.01
	4	0.90 ± 0.03	0.99 ± 0.01
	5	0.92 ± 0.03	1.00 ± 0.01
128^3	2	0.81 ± 0.06	0.99 ± 0.02
	3	0.88 ± 0.05	0.99 ± 0.01
	4	0.91 ± 0.02	0.99 ± 0.01
64^3	1	0.74 ± 0.06	0.98 ± 0.01
	2	0.84 ± 0.04	0.97 ± 0.01
	3	0.89 ± 0.03	0.97 ± 0.01

CNNs, prediction times with a resampling grid size of 256^3 and depth of 5 is 4 s, 128^3 and 4 is 2.2 s, and 64^3 and 3 is 1.5 s; when extrapolated to large datasets, for example 5000 images, this results in predictions times of approximately 5.6 h, 3.1 h, and 2.1 h, respectively. A user may therefore decide that the increase in speed gained by using a smaller resampling grid outweighs the non-significant increase in PR-AUC seen at higher resampling grids. However, it should be noted that these results are for classification of DA status. In more complex tasks, such as disease prognostication with images, higher resampling grids may be required to retain important disease information embedded in the image that could be lost through resampling.

Furthermore, our most simplistic CNN (resampling grid 64^3 and depth of 1) achieved a PR-AUC (0.74 ± 0.06) comparable to our baseline high density ROI volume metric (0.73). A check was performed on a number of images falsely classified by the baseline high density ROI volume metric. It was found that many of the patient images that were incorrectly flagged as DA+ (false positives) contained other apparatuses in the field of view (i.e. pacemakers and trachea tubes). There was also a trend for the false negative classifications to occur in patient images with a smaller magnitude DA.

It was also observed in our results that as the depth of the CNN increased, so did the average PR-AUC. This was consistent across all epochs and resampling grids. This result indicates that in this classification problem the depth of the CNN may be more important to performance than resampling size. Additionally, going from machine generated feature sizes of 32^3 to 16^3 generated the greatest gains in PC-AUC, versus going from 16^3 to 8^3; this demonstrates that feature values quantifying finer details of the image are needed for this classification problem.

Despite the promising performance of our CNN, it is not without limitations. One such limitation is the utilization of data from a single institution and classification of a single type of artifact. Although our CNN accurately predicts DA status on Princess Margaret Cancer Centre patients, and will be useful for future big data projects within the institution, it may also be biased towards imaging practices from our institution (e.g. couch, imaging apparatuses, slice thicknesses, reconstruction methods, etc). Additionally, we have not evaluated our CNN for its classification performance on different types of metal artifacts (e.g. pacemakers and joint replacements) that degrade images in a similar way to DAs. If it is found that our CNN is not generalizable to external dataset or different metal artifacts there are opportunities to implement transfer learning (LeCun *et al* 2015, Kensert *et al* 2019) or increase data augmentation (e.g. rotation, warping) (Mikolajczyk and Grochowski 2018) as a way to fine tune the models and/or increase their generalizability and reduce overfitting. We plan to leverage these preliminary results to motivate increased sharing of data by other groups to make these future studies possible.

Future work may also choose to focus on different neural network topologies, as there are many other ones that could be selected for this type of image quality classification. These range from completely state of the art methods such as dense nets (Huang *et al* 2017) or residual nets (He *et al* 2016), to different data augmentation methods (Mikolajczyk and Grochowski 2018) or activation functions (Klambauer *et al* 2017). However, for this study we were focused on feasibility and determining whether resampling grid sizes and/or CNN depths would affect performance. Different topologies may provide improvements in performance, but gains are minimal (He *et al* 2016, Huang *et al* 2017). A thorough investigation of all state of the art methods is beyond the scope of this manuscript, but would be an interesting avenue to pursue in future work.

The utility of our CNNs can be seen in the promising field of radiomics, which relies on the quantification of intensity based imaging features in a region of interest (ROI) for prognostic and predictive model develop-

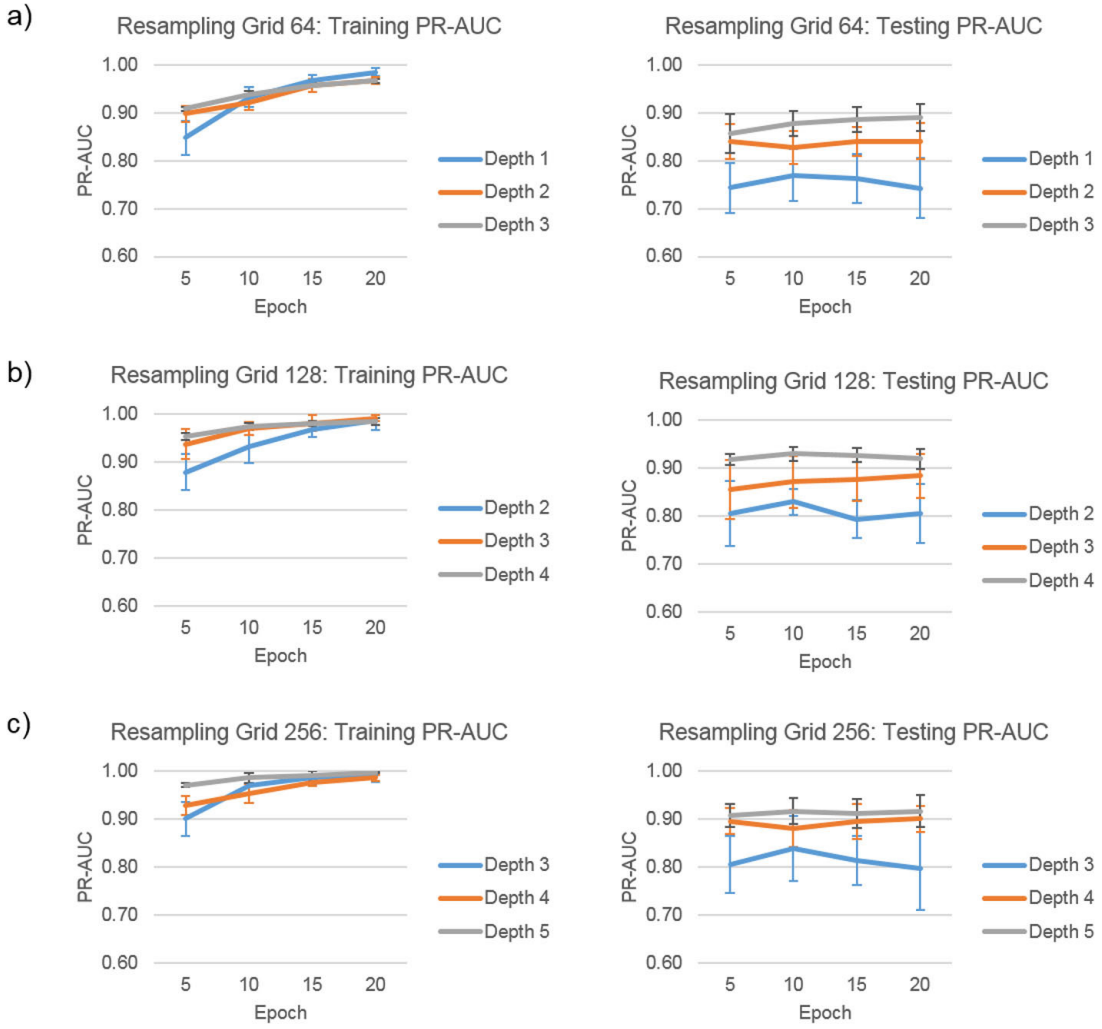

Figure 4. Average PR-AUC values across epoch 5, 10, 15 and 20 with error bars representing STDEV of PR-AUCs across the five-folds of model training and testing. (a) Resampling grid size of 64 with CNN depths of 1, 2 and 3; (b) resampling grid size of 128 with CNN depths of 2, 3, and 4; and (c) resampling grid size of 256 with CNN depths of 3, 4, and 5.

ment. Therefore, incorrect HU and segmentations could lead to misinterpretation of results and suboptimal model accuracies. Additionally, robust radiomic-based models require large datasets, and an automated method of quality assurance would reduce subjectivity of quality scoring, while changing a time consuming manual task to a passive one. Although it is not possible to study the impact DA's have on patient features due to the inability to obtain images with and without DAs, groups like Block *et al* (2016), Leijenaar *et al* (2016) and Elhalawani *et al* (2018) have all stated the importance of considering these artifacts. It is also worth noting that the CNNs we present in this paper do not specify whether the DA is present inside or outside a ROI; however, radiomic features are designed to probe imaging biomarkers not visible to an observer, and therefore DA streaking may impede feature quantification even if it does not visibly enter the ROI. For this reason, addition of this model to a radiomics pipeline would require the user to decide whether the image is appropriate for model training and testing based on the research question, clinical application and any other pertinent information. Alternatively, it may be possible to use DA status as a feature in radiomics modeling.

Additionally, automated radiation treatment (RT) planning software is being developed and implemented clinically that could benefit from methods similar to ours. These automated planning pipelines are designed to improve efficiency, standardization and quality of treatments for radiation therapy, and have seen commercial and clinical success in sites such as H&N (Mcintosh *et al*, Bodensteiner 2018). However, some plans and sites may be inherently more difficult to plan due to the presence of metal artifacts; this can be seen in the planning of prostate plans containing hip replacements that require special consideration and plan characteristics such as increased numbers of treatment beams (Dirkx *et al* 2013). Therefore, a quality assurance step designed to flag patients requiring further clinician involvement due to the presence of metal artifacts (e.g. dental artifacts, hip replacements, stents) would reduce the chance of erroneous treatment planning, while still increasing efficiency for patients without.

Commercialized metal artifact reduction software has also been developed by multiple groups (Li *et al* 2015, Huang *et al* 2015). These methods commonly reduce the impact of DAs by interpolating voxel HUs within the image sinogram (Abdoli *et al* 2010), but research is also being done to salvage image signal using deep learning methods (Zhang and Yu 2018, Huang *et al* 2018, Hu *et al* 2019). Although these methods are designed to reduce the qualitative impact of the artifact, it is possible for new artifacts to be generated in the image that mitigate its benefits in fields such as radiomics (Block *et al* 2016). In the future, a classification model, such as the one presented in this paper, could determine whether these methods are effectively removing the DA prior to utilizing the data for other purposes. This type of test would also be beneficial for companies to perform on their methods to demonstrate the effectiveness of their techniques.

Furthermore, the utilization of a single observer for scoring of DA status permitted consistency amongst the classification of DAs since each manual observer has a specific sensitivity and specificity for what is considered a DA. However, it is possible that misclassifications occurred due to fatigue or inappropriate window leveling, thereby reducing the reproducibility of the DA status classification. These potential misclassifications would not only affect validation, but also the training of the model. Multi-observer classification is often preferred to single observer classification in order to obtain more reliable ground truth labels. Future work may be able to reduce potential spurious misclassifications and increase classification reproducibility by using multi-observer scoring with standardized window-leveling and well defined classification rules.

This work demonstrates the potential for efficient image quality assurance methods. Our findings demonstrate the usage of accepted methods, and data from a single institution for the generation of a DA sorting model that could be used in automated pipelines and image analysis protocols. These automated methods are capable of completing routine tasks and freeing up clinician time for more important duties, but only if done correctly. The current state of machine learning in cancer care still requires standardized, high quality data because there is not enough openly available data to generate models robust to all potential variants. Therefore, development of quality assurance protocols and models are essential to the progress of automated methods in clinical cancer care.

Conclusion

Our work demonstrates the potential to automate specific quality assurance steps through model development, making an important and time consuming task passive. Our best performing CNN classified H&N CT images from a single institution based on the presence of DAs with an AUC of 0.92 ± 0.03, and the studied CNN resampling grid sizes were found to impact the AUC non-significantly; indicating that smaller resampling grid sizes could be used effectively if increased speed is required. Future work will explore the generalizability of our model to external datasets.

Acknowledgments

The authors thank Scott Bratman, Mike Sharpe, Shao Hui Huang, Brian O'Sullivan and Biu Chan for their assistance in obtaining and curating the utilized datasets. The work was supported by the Natural Sciences and Engineering Research Council, the Strategic Training in Transdisciplinary Radiation Science for the 21st Century Program, the Canadian Institutes for Health Research, the Ontario Institute for Cancer Research, and the Terry Fox Research Institute.

Conflicts of interest

The authors have no conflicts of interest to report.

References

Abdoli M, Ay M R, Ahmadian A, Dierckx R A J O and Zaidi H 2010 Reduction of dental filling metallic artifacts in CT-based attenuation correction of PET data using weighted virtual sinograms optimized by a genetic algorithm *Med. Phys.* **37** 6166–77

Block A M *et al* 2016 Radiomics in head and neck radiation therapy: impact of metal artifact reduction *Int. J. Radiat. Oncol. Biol. Phys.* **99** E640

Bodensteiner D 2018 RayStation: external beam treatment planning system *Med. Dosim.* **43** 168–76

Bray F and Parkin D M 2009 Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness *Eur. J. Cancer* **45** 747–55

Cagnon C, Cody D, McNitt-Gray M, Seibert J, Judy P and Aberle D 2006 Description and implementation of a quality control program in an imaging-based clinical trial *Acad. Radiol.* **13** 1431–41

Diehn F E *et al* 2017 CT dental artifact: comparison of an iterative metal artifact reduction technique with weighted filtered back-projection *Acta Radiologica Open* **6** 1–8

Dirkx M, Voet P, Breedveld S and Heijmen B 2013 Automated multicriterial plan generation for prostate cancer patients with metal hip prostheses: comparison of planning strategies *Med. Phys.* **40** 380

Elhalawani H *et al* 2018 Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients *Sci. Rep.* **8** 1524
Ger R B *et al* 2018 Practical guidelines for handling head and neck computed tomography artifacts for quantitative image analysis *Comput. Med. Imaging Graph.* **69** 134–9
Gillies R J, Kinahan P E and Hricak H 2015 Radiomics: images are more than pictures, they are data *Radiology* **278** 151169
Graham M S, Drobnjak I and Zhang H 2018 A supervised learning approach for diffusion MRI quality control with minimal training data *NeuroImage* **178** 668–76
Hall E L *et al* 1971 A survey of preprocessing and feature extraction techniques for radiographic images *IEEE Trans. Comput.* **C–20** 1032–44
Harlow C A and Eisenbeis S A 1973 The analysis of radiographic images *IEEE Trans. Comput.* **C–22** 678–89
He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* vol 2016-Decem pp 770–8
Hu Z *et al* 2019 Artifact correction in low-dose dental CT imaging using Wasserstein generative adversarial networks *Med. Phys.* **46** 1686–96
Huang G, Liu Z, Van Der Maaten L and Weinberger K Q 2017 Densely connected convolutional networks *Proc.—30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017* vol 2017-Janua pp 2261–9
Huang S J *et al* 2015 An evaluation of three commercially available metal artifact reduction methods for CT imaging *Phys. Med. Biol.* **60** 1047–67
Huang X *et al* 2018 Metal artifact reduction on cervical CT images by deep residual learning *BioMed. Eng.* **17** 175
Kansagra A P *et al* 2016 Big data and the future of radiology informatics *Acad. Radiol.* **23** 30–42
Kelly C, Pietsch M, Counsell S and Tournier J 2016 Transfer learning and convolutional neural net fusion for motion artefact detection *Proc. Intl. Soc. Mag. Reson. Med.* **3523** 1–2
Kensert A, Harrison P J and Spjuth O 2019 Transfer learning with deep convolutional neural networks for classifying cellular morphological changes *SLAS Discov.* **24** 466–75
Klambauer G, Unterthiner T, Mayr A and Hochreiter S 2017 Self-normalizing neural networks *Adv. Neural Inf. Process. Syst.* **2017-Decem** 972–81 (arXiv:1706.02515v5)
Krizhevsky A, Sulskever Ii and Hinton G E 2012 ImageNet classification with deep convolutional neural networks *Adv. Neural Inf. Process. Syst.* **60** 84–90
Lambin P *et al* 2017 Radiomics: the bridge between medical imaging and personalized medicine *Nat. Rev. Clin. Oncol.* 1–20
LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44
Leijenaar R T H *et al* 2016 Radioimcs in OPSCC: a novel quantitative imaging biomarker for HPV status? *ESTRO 35* p S196
Lewis-Jones H, Colley S and Gibson D 2016 Imaging in head and neck cancer: United Kingdom national multidisciplinary guidelines *J. Laryngol. Otol.* **130** S66–7
Li H *et al* 2015 Clinical evaluation of a commercial orthopedic metal artifact reduction tool for CT simulations in radiation therapy *Med. Phys.* **39** 7507–17
Mcintosh C, Welch M and Mcniven A Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method *Phys. Med. Biol.* **62** 5926–44
Mikolajczyk A and Grochowski M 2018 Data augmentation for improving deep learning in image classification problem *2018 Int. Interdiscip. PhD Work. IIPhDW 2018* pp 117–22
Moore S M *et al* 2005 Image quality assurance in the prostate, lung, colorectal, and ovarian cancer screening trial network of the national lung screening trial *J. Digit. Imaging* **18** 242–50
Nielsen M A 2015 *Neural Networks and Deep Learning* (San Francisco: Determination Press)
Oh J H, Pouryahya M, Iyer A, Apte A P, Tannenbaum A and Deasy J O 2019 *Kernel Wasserstein Distance* pp 1–10 (arXiv:1905.09314)
Oliphant T E 2006 *A guide to NumPy* (USA: Trelgol Publishing)
Olliff G, Richards J, Connor P, Wong S, Beale W L and Madani T 2014 Recommendations for cross-sectional imaging in cancer management *Headn and Neck Cancers* 2nd edn (London: The Royal College of Radiologists)
Pedregosa F, Weiss R and Brucher M 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30 (arXiv:1201.0490)
Raghupathi W and Raghupathi V 2014 Big data analytics in healthcare: promise and potential *Heal. Inf. Sci. Syst.* **2**
Paszke A *et al* 2017 Automatic differentiation in PyTorch *31st Conf.on Neural Information Processing Systems (NIPS 2017)* pp 1–4
van der Walt S *et al* 2014 Scikit-image: image processing in Python *PeerJ* **2** e453
Wei L *et al* 2019 Automatic recognition and analysis of metal streak artifacts in head and neck computed tomography for radiomics modeling *Phys. Imaging Radiat. Oncol.* **10** 49–54
Welch M L and Jaffray D A 2017 Radiomics: the new world or another road to El Dorado? *JNCI Natl Cancer Int.* **109** 7–8
Welch M L *et al* 2019 Vulnerabilities of radiomic signature development: the need for safeguards *Radiother. Oncol.* **130** 2–9
Yaniv Z, Lowekamp B C, Johnson H and Beare R 2017 SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research *J. Digit. Imaging* **31** 290–303
Zhang Y and Yu H 2018 Convolutional neural network based metal artifact reduction in x-ray computed tomography *IEEE Trans. Med. Imaging* **37** 1370–81