

External validation and transfer learning of convolutional neural networks for computed tomography dental artifact classification

Citation for published version (APA):

Welch, M. L., McIntosh, C., Traverso, A., Wee, L., Purdie, T. G., Dekker, A., Haibe-Kains, B., & Jaffray, D. A. (2020). External validation and transfer learning of convolutional neural networks for computed tomography dental artifact classification. *Physics in Medicine and Biology*, 65(3), Article 035017. <https://doi.org/10.1088/1361-6560/ab63ba>

Document status and date:

Published: 01/02/2020

DOI:

[10.1088/1361-6560/ab63ba](https://doi.org/10.1088/1361-6560/ab63ba)

Document Version:

Publisher's PDF, also known as Version of record

Document license:

Taverne

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

PAPER

External validation and transfer learning of convolutional neural networks for computed tomography dental artifact classification

To cite this article: Matteo L Welch *et al* 2020 *Phys. Med. Biol.* **65** 035017

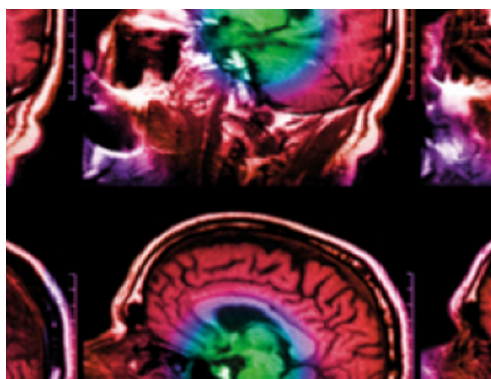
View the [article online](#) for updates and enhancements.

You may also like

- [Identification of maize disease based on transfer learning](#)
Xiaolin Sun and Jiangshu Wei
- [Obstructive sleep apnea prediction from electrocardiogram scalograms and spectrograms using convolutional neural networks](#)
Huseyin Nasifoglu and Osman Eroglu
- [Atrial fibrillation identification based on a deep transfer learning approach](#)
Ali Ghaffari and Nasimalsadat Madani

Recent citations

- [Machine Learning Algorithm Validation](#)
Farhad Maleki *et al*



IPEM | IOP

Series in Physics and Engineering in Medicine and Biology

Your publishing choice in medical physics,
biomedical engineering and related subjects.

Start exploring the collection—download the
first chapter of every title for free.



PAPERS

External validation and transfer learning of convolutional neural networks for computed tomography dental artifact classification

RECEIVED
5 November 2019REVISED
5 December 2019ACCEPTED FOR PUBLICATION
18 December 2019PUBLISHED
5 February 2020Mattea L Welch^{1,6,9,10}, Chris McIntosh^{1,4,6,8,9}, Alberto Traverso⁵, Leonard Wee⁵, Tom G Purdie^{2,4,6,9},
Andre Dekker⁶, Benjamin Haibe-Kains^{1,6,7,8} and David A Jaffray^{1,2,3,4,6,9}¹ Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada² Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada³ IBBME, University of Toronto, Toronto, Ontario, Canada⁴ Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto, Ontario, Canada⁵ Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, The Netherlands⁶ Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada⁷ Ontario Institute of Cancer Research, Toronto, Ontario, Canada⁸ Vector Institute, Toronto, Ontario, Canada⁹ The Techna Institute for the Advancement of Technology for Health, Toronto, Ontario, Canada¹⁰ Author to whom any correspondence should be addressed.E-mail: mattea.welch@rmp.uhn.ca**Keywords:** computed tomography, dental artifacts, quality assurance, deep learning, external validationSupplementary material for this article is available [online](#)**Abstract**

Quality assurance of data prior to use in automated pipelines and image analysis would assist in safeguarding against biases and incorrect interpretation of results. Automation of quality assurance steps would further improve robustness and efficiency of these methods, motivating widespread adoption of techniques. Previous work by our group demonstrated the ability of convolutional neural networks (CNN) to efficiently classify head and neck (H&N) computed-tomography (CT) images for the presence of dental artifacts (DA) that obscure visualization of structures and the accuracy of Hounsfield units. In this work we demonstrate the generalizability of our previous methodology by validating CNNs on six external datasets, and the potential benefits of transfer learning with fine-tuning on CNN performance. 2112 H&N CT images from seven institutions were scored as DA positive or negative. 1538 images from a single institution were used to train three CNNs with resampling grid sizes of 64^3 , 128^3 and 256^3 . The remaining six external datasets were used in five-fold cross-validation with a data split of 20% training/fine-tuning and 80% validation. The three pre-trained models were each validated using the five-folds of the six external datasets. The pre-trained models also underwent transfer learning with fine-tuning using the 20% training/fine-tuning data, and validated using the corresponding validation datasets. The highest micro-averaged AUC for our pre-trained models across all external datasets occurred with a resampling grid of 256^3 ($\text{AUC} = 0.91 \pm 0.01$). Transfer learning with fine-tuning improved generalizability when utilizing a resampling grid of 256^3 to a micro-averaged AUC of 0.92 ± 0.01 . Despite these promising results, transfer learning did not improve AUC when utilizing small resampling grids or small datasets. Our work demonstrates the potential of our previously developed automated quality assurance methods to generalize to external datasets. Additionally, we showed that transfer learning with fine-tuning using small portions of external datasets can be used to fine-tune models for improved performance when large variations in images are present.

Introduction

Increased computing power has provided opportunities for integration of automation into all aspects of cancer care. Methods are being researched and developed within radiation oncology to automate routine cognitive processes, thereby freeing up valuable time for more complex tasks; namely, methods for automated segmentation (Clark *et al* 1998, Davis *et al* 2005, Dou *et al* 2017) and machine-based treatment planning (Purdie *et al* 2011, Hansen *et al* 2016, McIntosh *et al* 2017) are of great interest and have had success in clinical integration (McIntosh *et al* 2017, Bodensteiner 2018). Additionally, automated information generation, whereby automated pipelines are utilized to extract and generate new information not immediately human minable from data (e.g. radiomics), are also undergoing a surge of interest for detection and prognostic modeling (Esteva *et al* 2017, Welch *et al* 2019a). However, despite the promise of automation, it is challenged by the nuances of medical data; including variations in data quality that can lead to unforeseen biases in results (Welch *et al* 2019a). Therefore, data curation with respect to its quality would represent a fundamental step towards reliable and reproducible results.

Dental artifacts (DA) in head and neck (H&N) computed tomography (CT) images have been identified as a challenge for automating routine cognitive processes and information generation (Block *et al* 2017, Ger *et al* 2018, Welch *et al* 2019b). Like many other artifacts, they have the potential to introduce a-priori biases that affect the results of automated image analysis tools. DAs result in poor structure visualization and Hounsfield unit calculation, which can affect contouring (Hansen *et al* 2017), treatment planning (Mail *et al* 2013), and quantification of images (Leijenaar *et al* 2016, Block *et al* 2017). By identifying CT images impacted by these artifacts we can safeguard automated methods against certain biases. However, proper curation for these types of artifacts is a labour intensive task usually involving manual classification of DA positive (DA+) and DA negative (DA-) images; a seemingly simple tasks for the dataset sizes currently being utilized, but a laborious task for large retrospective datasets that will become available in the future as the big data paradigm integrates into clinics.

We have previously studied a method for automated classification of DA status in H&N CT images using convolutional neural networks (CNN) (Welch *et al* 2019a). Using a dataset of 1538 images we achieved a precision recall area under the curve of 0.92 ± 0.03 . Furthermore, we explored the impact of various resampling grid sizes and CNN depths, discovering that more computationally efficient CNNs could be utilized for increased speed with insignificant loss of classification performance. These results were obtained using data from a single institution, and therefore provided evidence about the performance of our CNN on a dataset with consistent imaging practises. However, validation comes in many forms and internal validation can lead to overfitting and suboptimal performance with unstable results depending on the size of the training dataset (Feinstein 1996). External validation is another test of validity and is designed to test generalizability of a model to a different, but related, source of data. It is also the most convincing test of validity that proves a modeling method has not simply memorized irrelevant noise associated with the training images.

In addition, CNNs provide an opportunity to ‘fine-tune’ models for better performance on external datasets. Transfer learning is the process of tuning a previously trained CNN’s learned features to a new dataset using a smaller number of training images (Pan and Yang 2010, Kelly *et al* 2016). It assumes that a pre-trained network contains information about generic features that can be generalized to a new dataset, and does not require related categories to be present. For example, a dataset containing over 15 million images with crowd-sourced annotations from 22 000 categories, including 120 dog breeds (Krizhevsky *et al* 2017) has trained models used for transfer learning with medical data. Despite the seeming irrelevance of the training images from these datasets to medical data, models have successfully been utilized to classify cellular morphological changes from high-content microscopy images (Kensert *et al* 2019), and skin cancer lesions at a similar accuracy to dermatologists (Esteva *et al* 2017). Usage of these models and training datasets are often limited to 2D RGB images compatible with the initial training set, which can be challenging for medical images since many of them are three-dimensional (3D) grey-scale images. This warrants the development of a pre-trained model compatible with 3D medical data that can be fine-tuned for improved generalizability performance.

Automated processes for quality assurance of medical images would increase objectivity and efficiency of these important tasks, thereby reducing potential biases in conclusions and results. In this work we focus on the classification of DA+ /DA- H&N CT images, where we aim to determine the generalizability of pre-trained CNNs on external datasets, and the impact of resampling grid sizes on classification performance. Additionally, the feasibility and performance of transfer learning with fine-tuning with a variety of external datasets containing different event distributions, imaging practices and dataset sizes is explored.

Methods

Datasets and dental artifact classification

Seven H&N CT datasets were used in this work. A single dataset comprised of 1538 H&N CTs from the Princess Margaret Cancer Centre was used for training of the initial model, hence forth referred to as the ‘pre-trained

model'. Six external datasets from different institutions, totaling 574 CTs, were used for external validation and fine-tuning transfer learning with the pre-trained model: (1) H&N1 with 156 planning CTs (Aerts *et al* 2014); (2) H&N2 with planning CTs for 129 patients (Aerts *et al* 2014); (3) HGJ with planning CTs for 90 patients; (4) HMR with planning CTs for 41 patients; (5) CHUM with planning or diagnostic CT for 59 patients; (6) CHUS with CT images from diagnostic PET/CTs for 98 patients. Datasets HGJ, HMR, CHUM and CHUS are from a publically available Cancer Imaging Archive (TCIA) dataset (Valli res *et al* 2017). Image details are found in table 1. All images were collected using 16 bit allocation, preventing truncated HU values caused by 12 bits. Additionally, to the authors' knowledge, MAR methods were not utilized during imaging.

We converted all DICOM CT image volumes to nearly raw raster data (nrrd) formats automatically using Python and the SimpleITK library (Yaniv *et al* 2018); however, any imaging format compatible with SimpleITK imaging loading is appropriate. A single observer with eight years of medical imaging experience scored the DA status of each patient's converted nrrd CT volume. DA status options were DA positive if a dental artifact existed (DA+, status = 1) or DA negative if a dental artifact did not exist (DA−, status = 0). As in our previous publication (Welch *et al* 2019a), magnitude of the DA artifact was not considered. A patient was considered DA+ if DA streaking existed on any slice of the image volume (figure 1).

Data preprocessing

All CT volumes were processed prior to utilization using a multistep procedure. Details for each of the steps are outlined in Welch *et al* (2019a), and include: (1) interpolation of voxels to 1 mm³ using the SimpleITK linear resampling image filter; (2) data augmentation using random cropping of 10% and left-right flipping of 60% of the training or fine-tuning data; (3) padding of CT volumes to a uniform size to maintain the aspect ratio of the volume during resizing; (4) resizing of CT image volumes to determine the impact of various resampling grids on CNN generalizability and transfer learning with fine-tuning performance. For our work, resampling grids sizes of 256³, 128³ and 64³ voxels were analyzed for performance. Examples of image slices at the different resampling grids can be found in figure 2. Smaller resampling grids result in images with less detail, while larger resampling grids retain more of the detail found in the original image.

Training of pre-trained model

We used the open-source python library, PyTorch (Shaikh 2018) for this study, and a VMware, Inc. virtual machine with 10 Intel Xeon CPU E5-2690 processors and a NVIDIA Tesla K40m GPU. Batches of seven images, randomly selected from our training dataset, were fed into our 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. Convolutional kernels with a size of 5 and padding of 2 were used on the first convolutional layer, all subsequent layers used a kernel size of 3 with a padding of 1. Uneven class distributions were accounted for by using weighted optimization.

Three CNNs were pre-trained based on our previous results (Welch *et al* 2019b). CNN depths of 3, 4, and 5 were chosen for resampling grid sizes of 64³, 128³ and 256³, respectively. 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 author knowledge regarding model convergence from unpublished studies.

Transfer learning with fine-tuning

For each of the six external datasets, the impact of transfer learning with fine-tuning on DA classification was tested using five-fold cross validation. Each of the datasets was divided into five-folds where $k - 1$ folds (80% of the data) were used for validation and the k th fold (the remaining 20% of the data) was used for transfer learning with fine-tuning. This split size was selected to simulate an external sites potential usage of the pre-trained CNN; it is assumed that the external users would classify a smaller portion of their data for training with no knowledge of the DA status distribution, see supplementary material figures 1 and 2 (stacks.iop.org/PMB/65/035017/mmedia) for the distributions of DA+ and DA− in each subsampled dataset.

All five transfer learning folds from each of the six external datasets were used to fine-tune the three pre-trained models (with resampling grids of 64³, 128³ and 256³). Fine-tuning occurred for 20 epochs on all CNN and fully-connected layers. The weights of the optimizer were updated to reflect the distribution of DA+ and DA− patients found in the dataset fold.

External validation of pre-trained and transfer learned models

All five validation subsets ($k - 1$ folds, 80% of dataset images) from the six external datasets were used to validate the three pre-trained models and their corresponding fine-tuned models. Each image volume from the validation subsets was fed through a CNN to obtain the model's soft-max DA status classification. A model's performance was evaluated every five epochs on both fine-tuning and validation datasets.

Table 1. Dataset details outlining the location, public/private status, image type, median number of slices, thickness, resolution, tube voltage peak, scanner manufacturer and bits for the seven datasets.

	PM	H&N2	MAASTRO	HGJ	HMR	CHUM	CHUS
Location	Princess Margaret Cancer Centre, Toronto, Canada	VUmc Cancer Clinic, Amsterdam Netherlands	MAASTRO Clinic, Maastricht, Netherlands	Hôpital Général Juif, Montréal, Canada	Hôpital Maisonneuve-Rosemont, Montréal, Canada	Centre Hospitalier de l'Université de Montréal, Montréal, Canada	Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, Canada
Public/private	Private	Private	Public (XNAT)	Public (TCIA)	Public (TCIA)	Public (TCIA)	Public (TCIA)
Image type	Planning CT	Planning CT	Planning CT	Planning CT	Planning CT	Planning and Diagnostic CT	Planning CT
Num. patients	1538	129	156	91	41	59	98
Median slice thickness and range (mm)	2 (1–4)	2.5 (2.5–5)	3 (1.5–3)	2.5 (2.5–3)	3 (1.5–3)	1.5 (1.5–3.27)	3 (2–3)
Median slice num. and range	182 (130–330)	110 (52–173)	134 (103–307)	156 (90–222)	127 (111–266)	237 (90–348)	141 (109–236)
Median pixel size and range (mm)	0.98 (0.61–2.00)	0.78 (0.56–0.98)	0.98 (0.98–1.10)	1.06 (0.88–1.27)	1.12 (0.61–1.26)	0.98 (0.98–1.17)	1.17 (0.68–1.17)
Bits	16	16	16	16	16	16	16
Manu.	GE medical	58	37	0	91	35	18
	Philips	257	0	1	0	6	41
	Toshiba	1223	0	0	0	0	0
	Siemens	0	7	99	0	0	0
	CMS, Inc.	0	0	56	0	0	0
	Unknown	0	21	0	0	0	0
Tube	120	1538	7	156	91	10	59
voltage	140	0	36	0	0	31	0
peak	Unknown	0	22	0	0	0	0
(kVp)							

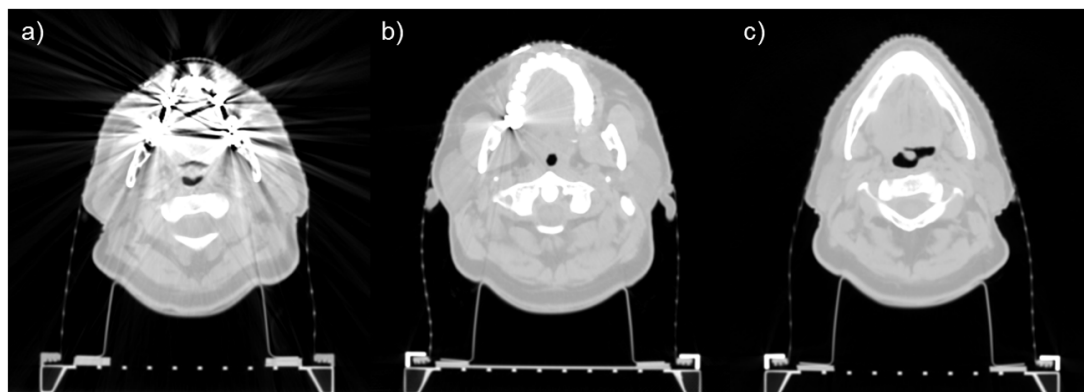


Figure 1. Example axial slices from three PM patient imaging volumes. This figure is a reproduction of a figure found in Welch *et al* (2019a) (a) shows a large magnitude DA, (b) shows a small magnitude DA, and (c) shows an image with no DA.

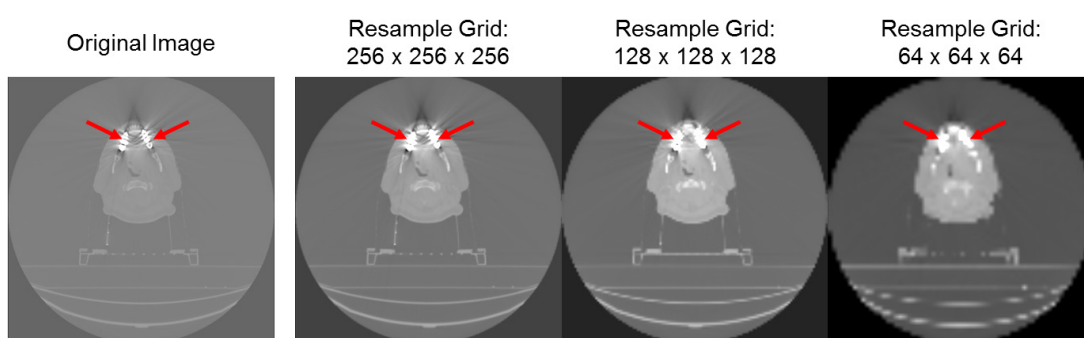


Figure 2. Example of image slices at various resampling grids of interest. Dental artifact is indicated by red arrows.

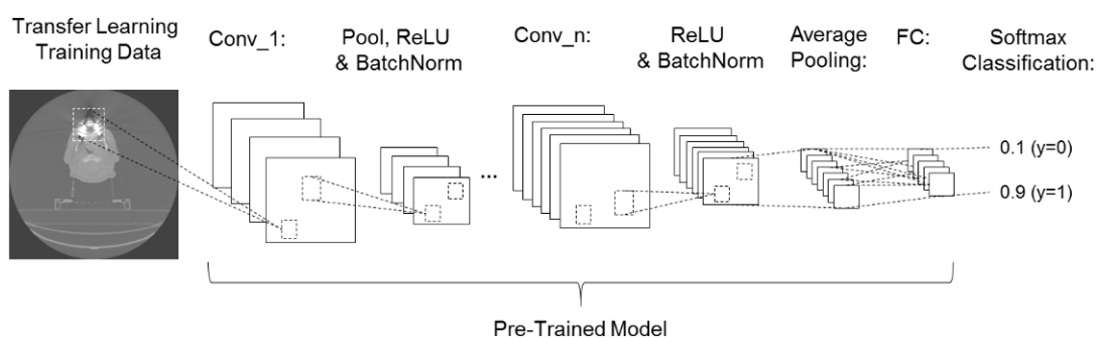


Figure 3. Example schematic of the DA status CNN. For simplicity, the training is shown here with a 2D image. Batches of 14 images were input into the network. The output of the convolutional layers underwent BatchNormalization, rectified linear unit (ReLU) and pooling. The final layer of the CNN only 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. Filter dimensions are found in table 2.

Table 2. Details of the number of convolutional layers, sizes of convolutional layers input and outputs, and size of the fully connected layer are given in this table relative to the resampling grids used. All depths resulted in a fully connected layer feature size of 8^3 .

					Conv_1		Conv_2		Conv_3		Conv_4		Conv_5		Fully connected
			Input	Output	Input	Output	Input	Output	Input	Output	Input	Output	Input	Output	layer feature size
Resampling Grid	256	Depth	5	1	4	4	8	8	16	16	32	32	64	$8 \times 8 \times 8$	
	128		4	1	4	4	8	8	16	16	32	N/A	N/A	$8 \times 8 \times 8$	
	64		3	1	4	4	8	8	16	N/A	N/A	N/A	N/A	$8 \times 8 \times 8$	

Table 3. Distributions of DA+ and DA− images in each dataset.

	Positive dental artifact (DA+)	Negative dental artifact (DA−)
Pre-trained model training data		
PM	1092 (71%)	446 (29%)
External data: transfer learning/fine-tuning and validation		
H&N2	70 (54%)	59 (46%)
MAASTRO	73 (47%)	83 (53%)
HGJ	56 (62%)	35 (38%)
HMR	10 (24%)	31 (76%)
CHUM	32 (54%)	27 (46%)
CHUS	44 (45%)	54 (55%)

Since we were working under the assumption during fine tuning that the distribution of DA statuses was unknown we used the area under the receiver operating characteristic curve (AUC) to evaluate performance of a CNN instead of the precision recall curve. The calculated AUC could then be used as a representative value of the true positive rate versus the false positive rate. The AUC was calculated using Python's Sci-kit learn library (Pedregosa *et al* 2012), and the average and standard deviation (STDev) across the five iterations of testing sets is also reported.

Results

The prevalence of DA+ and DA− images in each dataset, as scored by a single observer with 8 years of medical imaging experience, are found in table 3 below.

The pre-trained models had micro-averaged AUCs of 0.88 ± 0.01 , 0.89 ± 0.02 , and 0.91 ± 0.01 for resampling grid sizes of 64^3 , 128^3 and 256^3 , respectively, across all external datasets and data splits. After 20 epochs, the transfer learning with fine-tuning models trained for 64^3 , 128^3 and 256^3 resampling grid sizes had micro-average AUCs of 0.89 ± 0.01 , 0.90 ± 0.01 , and 0.92 ± 0.01 , respectively, across all external dataset and data splits. Average AUCs and STDev for each of the individual external datasets over the five-folds using the pre-trained and fine-tuned models after 20 epochs are found in table 4.

Average AUCs and STDevs for each dataset validated on the pre-trained models, and fine-tuned models with resampling grids of 64^3 , 128^3 and 256^3 after 5, 10, 15 and 20 epochs are found in figure 4. AUCs for each of the five external data splits are in supplementary material tables 1–3.

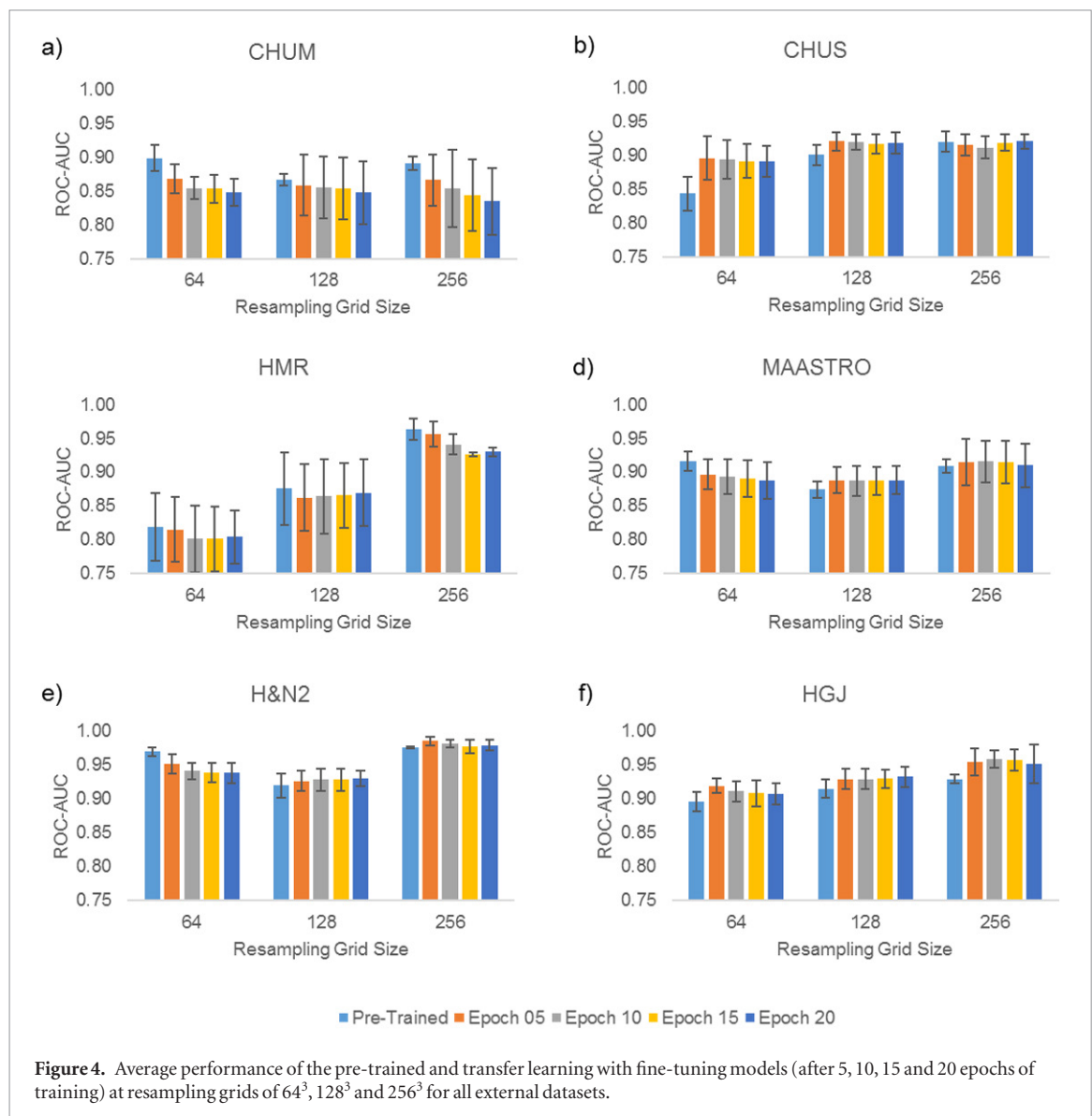
Discussion

The volume, variety, velocity and veracity of measurement data in cancer care is expanding to capture and integrate a diverse set of disease and host factors. Automated pipelines and processes are being developed to explore this data, but their future success depends on an understanding of the nuances of medical data and its quality. For this reason, efficient methods designed to safeguard big data methodology against potential data biases, while allowing users to have control over the eventual usage of the data, becomes vitally important. In this work we externally validated pre-trained CNNs designed to classify DA status in H&N CTs on six external datasets, exploring the impact of resampling grid sizes and transfer learning with fine-tuning on classification performance. Generalizability increased with resampling grid sizes when using the pre-trained models, with the highest micro-averaged AUC across all external datasets occurring with a resampling grid of 256^3 ($\text{AUC} = 0.91 \pm 0.01$); transfer learning with fine-tuning further improved generalizability when utilizing a resampling grid of 256^3 , to a micro-averaged AUC of 0.92 ± 0.01 . Our results demonstrate not only the potential to automate data quality checks, but also the benefits and pitfalls of fine-tuning models for usage with datasets external to training data.

This work builds upon our previous publication (Welch *et al* 2019b) and aims to determine whether prior acquired knowledge from pre-trained CNNs could be leveraged for usage in new unique datasets. External validation demonstrates performance of a model in a different patient population and puts in perspective preliminary single site studies being performed across the literature, while acknowledging how those results might change for other centers (Altman and Royston 2000, Liu *et al* 2019). Our selected external datasets utilized data from different institutions that contained variations in CT acquisition (e.g. slice number and thickness, pixel size, peak tube voltage, scanner manufacturer, etc) and other potential nuanced differences. Despite the pre-processing of our images to make them uniform in voxel size and compatible with usage in the CNNs, these variations can still

Table 4. AUCs for all external datasets and resampling grid sizes. AUCs are shown for validation of the pre-trained model, as well as the fine-tuned models after 20 epochs of training.

		Resampling grid size		
		64 ³	128 ³	256 ³
CHUM	Pre-transfer learning	0.90 ± 0.02	0.87 ± 0.01	0.89 ± 0.01
	20 epochs	0.85 ± 0.02	0.85 ± 0.05	0.83 ± 0.05
CHUS	Pre-transfer learning	0.84 ± 0.03	0.90 ± 0.02	0.92 ± 0.02
	20 epochs	0.89 ± 0.02	0.92 ± 0.02	0.92 ± 0.01
HMR	Pre-transfer learning	0.82 ± 0.05	0.88 ± 0.05	0.96 ± 0.02
	20 epochs	0.80 ± 0.04	0.87 ± 0.05	0.93 ± 0.01
MAASTRO	Pre-transfer learning	0.92 ± 0.01	0.87 ± 0.01	0.91 ± 0.01
	20 epochs	0.89 ± 0.03	0.89 ± 0.02	0.91 ± 0.03
H&N2	Pre-transfer learning	0.97 ± 0.02	0.92 ± 0.02	0.98 ± 0.01
	20 epochs	0.94 ± 0.01	0.93 ± 0.01	0.98 ± 0.01
HGJ	Pre-transfer learning	0.90 ± 0.01	0.92 ± 0.01	0.93 ± 0.01
	20 epochs	0.91 ± 0.02	0.93 ± 0.02	0.95 ± 0.03



have an impact on image quantification (Meyer *et al* 2019, Traverso *et al* 2019), and could have similar impacts on the ability of the CNNs to reproduce results in external datasets. In this work, we achieved micro-averaged AUCs with our pre-trained CNNs on the six external datasets of 0.88 ± 0.01 , 0.89 ± 0.02 , and 0.91 ± 0.01 , across the five-folds, for resampling grid sizes of 64³, 128³, and 256³, respectively. Furthermore, when portions of the exter-

nal datasets were used to fine-tune the pre-trained models, micro-averaged AUCs increased after 20 epochs to 0.89 ± 0.01 , ($p = 2.06e^{-1}$), 0.90 ± 0.01 ($p = 1.20e^{-1}$), and 0.92 ± 0.01 ($p = 4.37e^{-2}$), for resampling grid sizes of 64^3 , 128^3 and 256^3 , respectively. The data used for fine-tuning of the models was selected to be representative of an external site's usage. Specifically, we assumed that an external user would manually classify a small random selection (20%) of their entire dataset with no knowledge or consideration of DA event distributions. This further shows the versatility and robustness of our method and results.

Comparison of our results to the literature show improved performance over other methods, as well as a need for these approaches. In Wei *et al* (2019) they demonstrated the importance of DA consideration in radiomics studies by showing that the removal of DA+ patients from analysis improved predictions. However, in their work they only achieved a test AUC of 0.89 when recognizing DA+ patient image volumes, which does not outperform our CNNs when utilizing our pre-trained model with resampling grid sizes of 256^3 , or transfer learning with fine-tuning and resampling grid sizes of 128^3 and 256^3 . Additionally, their methodology requires definition of ROIs within the image and extraction of features prior to model application, a more hands-on approach when compared to our methods. Oh *et al* (2019) developed a method for classification of DA+ image slices that performs with a prediction rate of 97.10% and 74.10% for DA+ and DA− image slices, respectively. If it is desired by the user to retain image slices without visible artifacts in their analysis, our CNN could be used in conjunction with this type of method for initial flagging prior to the more computationally expensive methods presented by Oh *et al*. Granted, justification would be required to explain the implications of slice removal on shape and texture features.

Transfer learning with fine-tuning had the greatest benefit to AUC performance in the CHUS dataset when using a resampling grid size of 64^3 (figure 4(b)). Upon further investigation of the CHUS data it was discovered that immobilization bite blocks were used (figure 5) for this cohort, and present in $55.2\% \pm 5.1\%$ of all patients in the test sets. Bite blocks were not present in the pre-trained model training data, and were exclusive to the CHUS dataset. The increase in AUC for the CHUS data from 0.84 ± 0.03 to 0.89 ± 0.02 after transfer learning with fine-tuning therefore indicates that features specific to bite blocks may have been learned and resulted in significant performance improvements compared to our pre-trained model. To test this theory we removed the patients with bite blocks from the dataset and found that the performance of the pre-trained model on the new CHUS dataset remained unchanged at 0.88 ± 0.02 after 20 epochs of fine-tuning. These results demonstrate how transfer learning with large datasets containing image anomalies can be leveraged to improve model performance in external data that differs from the original training data. However, it should be noted that this was only observed at small resampling grid sizes and that images that retained more detail (i.e. larger resampling grid sizes), performed equally as well in the CHUS dataset with and without fine-tuning.

Despite the promising performance of transfer learning with fine-tuning that we observed with CHUS at small resampling grid sizes, we found it did not contribute to significant improvements in performance with the other datasets at small or large resampling grids (figure 4). More importantly, transfer learning with fine-tuning was not effective when the overall dataset, and therefore the number of training images available, was small. This observation is irrespective of the resampling grid size and is best observed in our HMR ($n = 41$) and CHUM ($n = 59$) datasets. In these datasets it can be seen that the AUC decreases or remains stable after transfer learning with fine-tuning (figures 4(a) and (c)), and has large variations across the five-folds of transfer learning with fine-tuning and validation. This indicates that transfer learning with fine-tuning may not be able to observe enough data variation to increase AUC, and thus warrant its application.

Additionally, when large variations in imaging are present in a small dataset the challenges are compounded. This is demonstrated in the CHUM dataset, where 16 of the 59 image volumes (27%) were diagnostic CTs instead of planning CTs. The difference in these images can be seen in figure 6. Transfer learning with fine-tuning in the CHUM dataset reduced AUCs from 0.90 ± 0.02 , 0.87 ± 0.01 , and 0.89 ± 0.01 by 0.05 ± 0.03 , 0.02 ± 0.04 , and 0.06 ± 0.05 , for resampling grid sizes of 64^3 , 128^3 and 256^3 , respectively. However, when the diagnostic images were removed from the dataset, and fine-tuning and validation were repeated, the post-transfer learning with fine-tuning AUCs increased by 0.004 ± 0.044 for a resampling grid sizes of 128^3 , and decreased by 0.004 ± 0.034 and 0.018 ± 0.044 for a resampling grid sizes of 64^3 and 256^3 , respectively. These results indicate that fine-tuning with a small dataset, containing images with large differences, makes the CNN more susceptible to overfitting; this differs from what we saw with the CHUS dataset that, although it contained large differences in the images, was large enough to provide a sufficient number of examples to the CNN. Based on these results we suggest that for smaller datasets with large imaging variations, manual classification or classification with our pre-trained model are preferable to transfer learning with fine-tuning.

The performance of our CNN on the CHUS and CHUM datasets indicate an opportunity for more sophisticated data augmentation to be included in future work. Data augmentation introduces variations and uncertainty into the training data to generate a more robust and generalizable CNN (Mikolajczyk and Grochowski 2018). In our work we included image cropping and flipping; however, image rotation, translation and noise introduction are other methods that may be explored in the future. Future work in this area could also explore

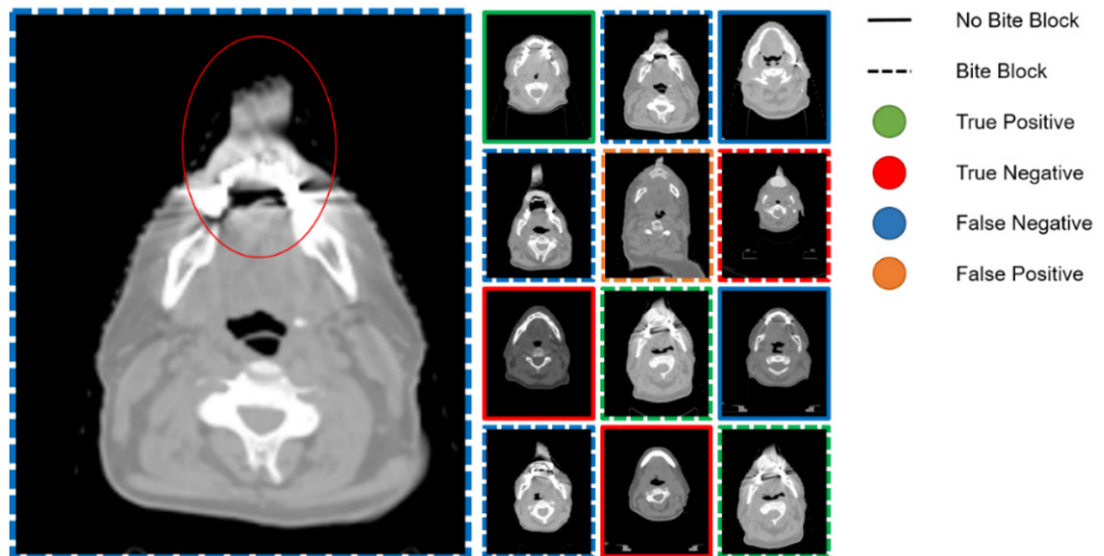


Figure 5. Example images from the CHUS dataset showing correct and incorrect classifications for patients with and without immobilization bite blocks. Images are in native resolution and have been cropped for easier visualization. A solid line represents patients without a bite block and a dashed line represents patients with bite blocks. Green, red blue and orange represent true positive, true negative, false negative and false positive predictions of DA status.

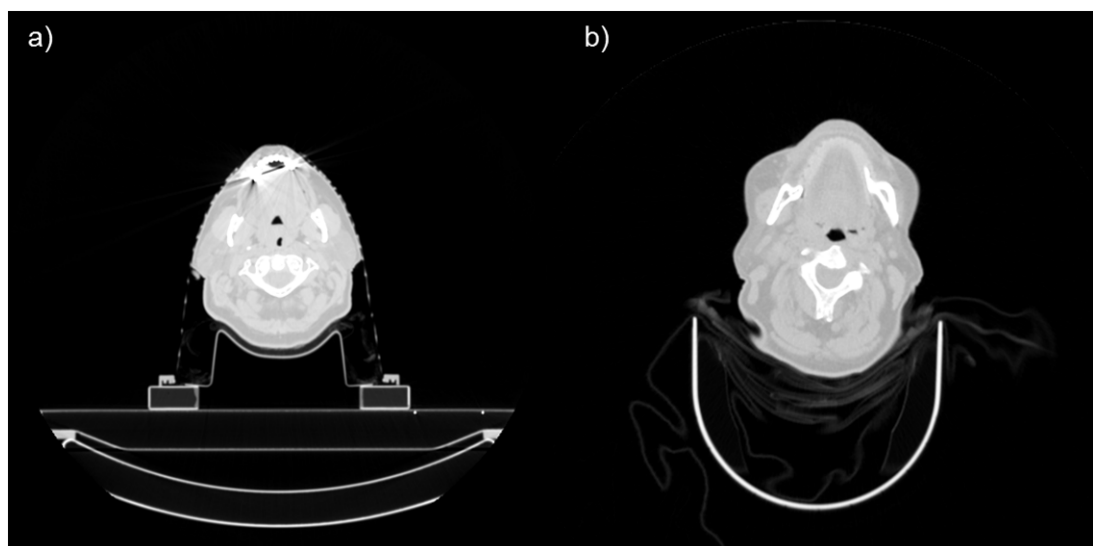


Figure 6. Example images from the CHUM dataset. (a) An axial slice of a planning image volume, (b) an axial slice of a diagnostic image volume. Both planning and diagnostic image volumes were present in the CHUM dataset. The differences in the images (couch, patient positioning, etc) provided challenges for fine-tuning our CNN that were compounded by the overall size of the dataset.

the impact of including synthetic data in training; an idea which has been explored for classification of motion artifacts in magnetic resonance imaging (Graham *et al* 2018). The inclusion of these data alterations may be able to provide enough examples during training that the challenges observed in CHUS and CHUM are mitigated.

Three main limitations exist in our work, including utilization of a single observer for ground-truth DA classification, binary classification of DAs and classification of a single type of artifact. Single observer classification allowed for consistent classification of DAs, since sensitivity and specificity of what is considered a DA is observer specific. However, inappropriate window leveling or fatigue may have resulted in misclassifications. Utilization of binary DA classifications further exacerbates this limitation since the streaking required for a patient to be considered DA+ by our observer was minimal and a second observer may disagree with the classifications. These discrepancies in DA status would impact both training and validation performance of our CNNs. Future work may be able to obtain more reliable ground truth labels, and reduce potential misclassifications due to user error, by using multi-observer classifications. Classification of DAs based on their magnitude may also improve results; however, in fields such as radiomics, even small artifacts should be considered potentially harmful since the goal

of the field is to quantify features not readily visible. Additionally, these models have been trained for DA classification in H&N CT images. However, other artifacts can be present in images that also require consideration. For example, metal artifacts pose similar problems in pelvic CTs due to hip replacements, and chest CTs from stents and pacemakers. Future validation of our model may demonstrate that these types of artifacts are classifiable with our pre-trained model, or learnable in a similar transfer learning approach.

By verifying the ability of a pre-trained model to classify DAs in external datasets, and the ability to fine-tune the model when large imaging variations are present, we have demonstrated the potential use of these techniques as a method of open-access data curation. Often, to obtain the number of events required to generate a robust conclusion, researchers will utilize open-access data (Herrick *et al* 2012, Clark *et al* 2013), generating large datasets where manually curating each image volume for DAs becomes a cumbersome task. By flagging images that may be of concern to a user in a passive manner, using models such as the presented CNNs, we are able to increase efficiency, while still allowing the user control of whether the images should be included in their pipeline. If it is decided that DAs are a concern, regardless of their magnitude or location, the image volume can be removed from analysis using the pre-classified data (Leijenaar *et al* 2016), the affected slices can be removed from the image volume (Elhalawani *et al* 2018), or metal artifact reduction (MAR) techniques could be applied (Zhang and Yu 2018). However, caution is warranted for usage of MAR methods in radiomics studies since new artifacts can be generated with these methods (Block *et al* 2017). Areas of automated research (e.g. contouring, RT treatment planning and quantification of imaging features), all rely on machine learning, putting results at risk of misinterpretation, and warranting the need for data quality checks. Having a method that is capable of data quality checks on large quantities of images, regardless of the data source, will be of great benefit as big data methods and large retrospective datasets become more common in cancer care.

Conclusion

Efficient and objective methods of medical data quality assurance are needed for effective utilization of big data methods in cancer care. Without appropriate consideration of the nuances and potential biases present in medical data we put our research at risk of false, misunderstood, or biased conclusions. Our work demonstrates the potential of our previously developed automated quality assurance methods to generalize to external datasets, and the ability of transfer learning to fine-tune our models for improved performance when large variations in imaging data are present. Future work will explore the generalizability of our model to metal artifacts external to H&N image volumes, while utilizing multi-class, multi-observer artifact classifications.

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

- Aerts H J W L *et al* 2014 Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach *Nat. Commun.* **5** 4006
- Altman D G and Royston P 2000 What do we mean by validating a prognostic model? *Stat. Med.* **19** 453–73
- Block A M *et al* 2017 Radiomics in head and neck radiation therapy: impact of metal artifact reduction *Int. J. Radiat. Oncol.* **99** E640
- Bodensteiner D 2018 RayStation: External beam treatment planning system *Med. Dosim.* **43** 168–76
- Clark K *et al* 2013 The cancer imaging archive (TCIA): maintaining and operating a public information repository *J. Digit. Imaging* **26** 1045–57
- Clark M C, Hall L O, Goldgof D B, Velthuizen R, Murtagh F R and Silbiger M S 1998 Automatic tumor segmentation using knowledge-based techniques *IEEE Trans. Med. Imaging* **17** 187–201
- Davis B C, Foskey M, Rosenman J, Goyal L, Chang S and Joshi S 2005 Automatic segmentation of intra-treatment CT images for adaptive radiation therapy of the prostate *Med. Image Comput. Comput. Assist. Interv.* **8** 442–50
- Dou Q *et al* 2017 3D deeply supervised network for automated segmentation of volumetric medical images *Med. Image Anal.* **41** 40–54
- 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
- Esteva A *et al* 2017 Dermatologist-level classification of skin cancer with deep neural networks *Nature* **542** 115–8
- Feinstein A R 1996 8.5. validation of results *Multivariable Analysis: an Introduction* (New Haven, CT: Yale University Press) pp 184–7

- 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
- 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
- Hansen C R *et al* 2016 Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans *Clin. Transl. Radiat. Oncol.* **1** 2–8
- Hansen C R *et al* 2017 Contouring and dose calculation in head and neck cancer radiotherapy after reduction of metal artifacts in CT images *Acta Oncol.* **56** 874–8
- Herrick R, Horton W, Olsen T, McKay M, Archie K A and Marcus D S 2012 XNAT central: open sourcing imaging research data *NeuroImage* **17** 1310–4
- 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
- Krizhevsky A, Sutskever I and Hinton G E 2017 ImageNet classification with deep convolutional neural networks *Commun. ACM* **60** 84–90
- LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44
- Leijenaar R T H *et al* 2016 Radiomics in OPSCC: a novel quantitative imaging biomarker for HPV status? *ESTRO 35* p S196
- Liu X *et al* 2019 A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis *Lancet Digit. Heal.* **1** e271–97
- Mail N *et al* 2013 The impacts of dental filling materials on RapidArc treatment planning and dose delivery: challenges and solution *Med. Phys.* **40** 081714
- McIntosh C, Welch M, McNiven A, Jaffray D A and Purdie T G 2017 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
- Meyer M *et al* 2019 Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings *Radiology* **293** 583–91
- Mikolajczyk A and Grochowski M 2018 Data augmentation for improving deep learning in image classification problem 2018 *Int. Interdisciplinary PhD Workshop* pp 117–22
- Nielsen M A 2015 *Neural Networks and Deep Learning* (Determination Press)
- Oh J H, Pouryahya M, Iyer A, Apte A P, Tannenbaum A and Deasy J O 2019 Kernel wasserstein distance (arXiv:1905.09314)
- Pan S J and Yang Q 2010 A survey on transfer learning *IEEE Trans. Knowl. Data Eng.* **22** 1345–59
- Pedregosa F *et al* 2012 Scikit-learn: machine learning in python *J. Mach. Learn. Res.* **12** 2825–30
- Purdie T G, Dinniwell R E, Letourneau D, Hill C and Sharpe M B 2011 Automated planning of tangential breast intensity-modulated radiotherapy using heuristic optimization *Int. J. Radiat. Oncol.* **81** 575–83
- Shaikh F 2018 An introduction to PyTorch—a simple yet powerful deep learning library (www.analyticsvidhya.com/blog/2018/02/pytorch-tutorial/)
- Traverso A *et al* 2019 Sensitivity of radiomic features to inter-observer variability and image pre-processing in apparent diffusion coefficient (ADC) maps of cervix cancer patients *Radiother. Oncol.* (<https://doi.org/10.1016/j.radonc.2019.08.008>)
- Vallières M *et al* 2017 Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer *Sci. Rep.* **7** 10117
- 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 *et al* 2019a Automatic classification of dental artifact status for efficient image veracity checks: effects of image resolution and convolutional neural network depth *Phys. Med. Biol.* **65** 015005
- Welch M L *et al* 2019b Vulnerabilities of radiomic signature development: The need for safeguards *Radiother. Oncol.* **130** 2–9
- Yaniv Z, Lowekamp B C, Johnson H J and Beare R 2018 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