SUPPLEMENTARY INFORMATION

Automated detection of dental artifacts for large-scale radiomic analysis in radiation oncology

Data

0.1 RADCURE Dataset Details

Supplementary table 1 provides a summary of acquisition parameters for all 3,211 patient images in the RADCURE dataset. The minimum, maximum, and median values of slice thickness, pixel spacing, X-Ray tube current, and number of slices across all images are given.

0.2 Annotator Agreement

Although the three DA magnitude classes described above define the classes distinctly, they are still highly qualitative and open to inter-annotator interpretation. In order to study any inter-annotator variability in class labelling, 482 images were annotated twice by different annotators, who we will refer to as annotators A, B and C. Annotator A labelled all 482 of these images, while annotator B labelled 381 of the images a second time, and the annotator C labelled the remaining 101 images a second time. The agreement between the different annotators was then studied. It was found that the annotators agreed on the overall magnitude of the artifact for 83% of the patients. In 78% of cases where the annotators disagreed on the overall DA label, they disagreed on whether it was strong or weak. In only 22% of cases did the annotators disagree about whether the artifact existed or not (i.e. annotator A labelled the image “none”, while annotator B labelled it “weak” or “strong”). In 67% of cases where the annotators disagreed about whether there was an artifact at all, the image was labelled as “weak” by one of the annotators. In addition, the most common kind of disagreement between annotators (65 of the 83 disagreements) occurred when annotator A labelled an image as strong, while either B or C labelled the same image as weak. This suggests that annotator A more readily labels images as strong which other annotators may have classified as weak. These classifications are summarized in Supplementary Table 2

It was also found that the annotators agreed on the z-index location of the DA or “mouth slice” in most image volumes. In particular, the annotators agreed on the exact same slice index for 46% of image volumes. Their location labels were within 5 slices of each other in 82% of cases, within 10 slices in 95% of cases, within 15 slices in 98% of cases, and within 20 slices in 99% of cases.

0.3 Confounding Factors in PyRadiomic Feature Analysis

0.3.1 Measuring the GTV-DA Distance

We made efforts to account for various confounding factors in our analysis of the correlation between PyRadiomic features and GTV-DA distance. One major factor is the way in which the distance between the DA and GTV is measured. In particular, DA streaks may only affect a subset of the pixels in the GTV. Representing the location of the GTV using its centre of mass may not capture the fact that pixels toward the edge of the tumour, closer to the DA are affected more strongly and therefore more strongly correlated with DA-GTV distance. In order to account for this, we computed correlations between GTV-DA distance and radiomic features, using the GTV pixel closest to the DA slice to compute this distance. As a sanity check, we also computed the correlation between these two distance metrics.

We found that the two distance metrics (centre of mass and nearest GTV pixel) were highly correlated, with a Pearson correlation coefficient of 0.93 and a Spearman rank correlation of 0.91.

0.3.2 Confounding PyRadiomic features

We also attempted to correct for radiomic features which are known to be correlated with many other radiomic features. In particular, we computed the partial Spearman correlation between GTV-DA distance and each radiomic feature, controlling for GTV volume. We found that many features still had high correlation with DA-GTV distance and that the features with the highest partial correlation were the same features that were directly-correlated (all using the “lbp-3D-k ” filter).

0.3.3 Confounding Clinical Features

Finally, we investigated any clinical features which may be correlated with DA-GTV distance. A χ^2 test was performed in order to investigate if categorical variables such as sex, smoking status, primary disease site, and stage had different distributions between different DA classes. These test results are summarised in Supplementary Table 3. Smoking status showed a high degree of stratification by DA group (P value $= 1.90 \times 10^{-8}$).

We also performed statistical tests to compare the distributions of two continuous clinical variables between DA groups. The distributions of age between DA classes (figure 8, left) were compared using a one-way ANOVA (P value $= 1.96 \times 10^{-9}$) and its non-parametric form, the Kruskal-Wallis H-test (P value $= 3.44 \times 10^{-11}$). We found more significant differences in the distributions of smoking rates (reported in number of cigarette packs per year) between DA classes (figure 8, right). The ANOVA (P value $= 7.446 \times 10^{-32}$) and the Kruskal-Wallis H-test (P value $= 8.41 \times 10^{-27}$) showed significantly different smoking distributions between DA classes.

Supplementary Methods

0.4 Human DA annotation in RADCURE

The dataset used for this study consists of 3211 head and neck cancer CT image volumes collected from 2005-07-26 to 2017-08-17 at the University Health Network (UHN) in Toronto, Canada (REB approval #17-5871). This dataset is referred to as RADCURE. Each patient's CT volume contained a median of 181 slices, with each slice consisting of 512 $\times$ 512 pixels. The median slice thickness varied between 2.0 and 3.0 mm for different patients and had a median of 2.0 mm (see Supplementary Table 1 for a comprehensive list of imaging settings).

We developed Artifact Labelling Tool for Artifact Reduction (ALTAR), an open-source web-application enabling the review of large sets of images and the annotation of the magnitude and location of the dental artifacts. The application clipped all images to be between -1000 and 1000 Hounsfield Units (HU), thereby facilitating viewing of the grayscale images in a dynamic range conducive to DA identification. Manual annotations classified each of the patients as either having "strong" artifacts, "weak" artifacts, or no visible artifacts for the entire image stack. A patient labelled as "strong" had to have at least one slice containing artifact streaks which obscured significant portions of the patient's body and which were easily visible outside the profile of the body. A patient labelled as "weak" had to have at least one slice with easily identifiable metal artifact streaks, but which did not fully obscure sections of the image and which were not plainly visible outside the patient's body. Finally, a patient labelled as "none" had no identifiable dental artifacts. This custom three-class scale was intended to provide our analysis with more information than a simple binary grouping, while acting as a simplification of 5 or 6-class scales used in previous studies [1, 2] which could be used by non-radiologists. The z-index of the axial slice containing the strongest artifacts was also annotated for each patient in the strong or weak DA class. For patients with no DA, an axial slice index in the mouth was labelled. Other metal artifact streaks caused by catheters, metal implants, etc. were not taken into consideration for our study. These non-dental artifacts may be less common, as one study found that of 1300 patients, 131 (10%) had artifacts below the head and neck region [3].

The images in our study were labelled by four different researchers, each labelling a subset of the data. The process was supervised by a researcher with 10 years of experience (details about the annotation accuracy and the analysis are reported in Supplementary Information). A subset containing 482 3D patient scans was randomly selected to be labelled by two different annotators in order to study the agreement between human annotators. The Matthews correlation coefficient between double-labelled images was calculated in order to be compared to automated classifiers.

0.5 Radiomic Feature Analysis

The relationship between quantitative imaging features and the existence and location of dental artifacts was studied. Radiomic features were extracted using the default settings of the open-source Python package, PyRadiomics (version 2.1.2) [4, 5]. 2490 of the 3211 patient image volumes contained a gross tumour volume (GTV) mask contoured by a radiation oncologist. The PyRadiomics package was then used to extract 1547 radiomic features from each of the 2490 patients with a GTV. We rescaled all features such that each feature had zero mean and unit variance across all patients. This reduced the inter-feature variation due to different units, allowing us to compare only the distributions of data between features.

The axial (z-axis) slice index of the dental artifact, or the patient's mouth for those without DAs, was manually labelled for each patient as described above. The z-index of the DA was defined to be the slice with the strongest visible artifact streaks, or the most central slice in cases where no slice was obviously brightest. In cases where no DA was present, the z-index of the mouth was labelled. All images and GTV masks were resampled to $1\times1\times1$ mm voxel spacing and the physical z-distance between the DA centre and GTV centre was calculated. The location of the GTV pixel nearest to the DA slice was also extracted for further analysis.

The difference in the distribution of radiomic feature values between strong DA and no DA images with respect to DA-GTV distances was evaluated. The image volumes were grouped into 40mm DA-GTV distance bins and a Wilcoxon rank-sum test between image features from volumes with strong DAs and image volumes with no DAs was performed. We used the Bonferroni correction for adjusting the nominal P-values of each correlation for multiple testing taking the number of tests to be the number of radiomic features (1547). A Bonferroni-corrected P-value < 0.05 was considered significant. The relationship between DA-GTV z-distance and the radiomic features was investigated using all 2490 patient volumes containing GTVs. For each feature, the partial Spearman correlation (adjusted for tumor volume) between the feature values and DA-GTV distances was computed independently for each DA magnitude (1039 strong, 751 weak, 877 none). This was repeated for both measures of DA-GTV distance (GTV centre of mass and GTV pixel nearest to the DA).

Finally, in order to validate the effect of removing "bad" images on radiomic features, we removed all patients where the centre slice of the DA overlapped with any pixel in the GTV. We performed the same Wilcoxon rank sum test between radiomic features from strong-DA and no-DA images from this smaller group of 1006 patients (529 strong, 477 no DA). In order to verify if any change in the number of significant P-values was simply due to inflated P-values as a result of a smaller sample size, we repeated the analysis 1000 times, each time taking a random sample of 1006 patients from the full dataset.

0.6 Detailed Description of the Sinogram-Based Detection Algorithm

All images were cropped by taking the first 350 pixels in the x-axis to remove the majority of the imaging table and fixture accessories which occupied the last 100-200 pixels of the x-axis. The image intensity range was then clipped to between -1000 and 1000 HU and divided by 1000 HU, normalizing all images to a range of (-1, 1). Next, the head was segmented in each two-dimensional slice of the patient's CT volume. This was done using the Otsu threshold [6] which minimizes the intra-class variance, σ_m, defined below:

$$\sigma_m^2(t) = \omega_0(t)\sigma_0^2(t) + \omega_1(t)\sigma_1^2(t)$$

where $\omega_0(t)$ and $\omega_1(t)$ are the probabilities of the two intensity classes separated by the threshold t and $\sigma_0^2(t)$ and $\sigma_1^2(t)$ are the variances of each class. The Otsu threshold was computed using the Python scikit-image package (version 0.16.2) [7]. The resulting mask of the patient's body and the head is extended outward by applying a Gaussian filter to the mask, followed by thresholding the filtered mask at a value of 0.01. This threshold value was chosen based on the width of the Gaussian filter kernel, to extend the edges of the body. Finally, each normalized image slice was multiplied by the mask, creating a set of two-dimensional images containing only the original pixels in the background. These images were then thresholded at a value of 0.04.

Each two-dimensional slice was then transformed into a sinogram by calculating 180 parallel projections using the Radon transform in Python's scikit-image. Finally, a sinogram is made by adding up projections from different angles. The mean pixel intensity in a central region of each slice's sinogram is then computed (Figure 2).

The thresholding, filtering, and sinogram steps resulted in a list of mean sinogram intensities for each axial CT slice in a given patient. We used the peak detection algorithm from the Python scipy package (version 1.4.1) to detect slices with intensities much higher than the mean for that patient. A peak is detected if it has a higher value than its neighbours and if it has an intensity more than four standard deviations above the mean (Figure 2B). The algorithm classifies the patient as DA positive if any peaks in the z-axis sinogram intensity curve are found. If no peaks are found, the patient is classified as DA negative. This method additionally outputs a prediction of the z-index of the DA when a patient CT volume is classified as DA positive.

0.7 Detailed description of the CNN Implementation

We used a pre-trained five-layer convolutional neural network (CNN) [8] as the secondary binary classifier in the detection pipeline (Figure 2C). This pre-trained model is available openly as a Docker through modelhub.ai (https://modelhub.ai); for compatibility, our patient CT volumes were preprocessed to contain isotropic voxels ($1\times1\times1$ mm) and a resampled matrix size of $256\times256\times256$, retaining the image aspect ratio. Details regarding training of the CNN can be found in our previous work which explored the impact of CNN depth and image resolution (https://pubmed.ncbi.nlm.nih.gov/31851961/). Additionally, the CNN was validated on external datasets prior to being made openly available [8].

0.8 Detailed Description of the Thresholding-based DA Location Algorithm

The thresholding-based algorithm works by first clipping the intensity values between the maximum intensity in one patient's CT image volume and 200 HU above that maximum. The standard deviation of each axial slice is then computed and peak detection is performed on the standard deviations of each axial slice using the Scipy `find_peaks` function in a similar manner to the SBD peak detection step. A peak, defined as a slice with a standard deviation higher than its adjacent slices, is detected if it has a value $\sigma_i > \mu_V + 1.5\sigma_V$, where μ_V is the mean of the slice standard deviations for that patient volume and σ_V is the standard deviation of the slice standard deviations for that patient. If any peaks are detected, the algorithm simply returns the indices of those peaks for that patient. Otherwise, the lower bound of the clipping range is decreased by 50 HU and the process is repeated for the patient until at least one peak is found.

0.9 Performance of the CNN and SBD algorithms as stand-alone binary classifiers

The class-weighted AUC was used to assess the accuracy of the CNN binary classifier based on the prediction scores produced by the network. The `roc_auc_score` function from Scikit-learn (version 0.22.1) was used with the "weighted" option [9]. This computes the metrics for each class, weighted by the number of true instances for each label. For all AUC and MCC values we also estimated a P-value from 5000 iterations of a randomized permutation test. We then estimated the confidence interval of these metrics by bootstrap sampling the test set with a sample size of 200 for 5000 iterations.

0.9.1 Sinogram Based Detection (SBD)

The sinogram-based method had an accuracy (correct detection rate) of 90.5% for the strong DA image volumes and an accuracy of 24.9% for the weak DA image volumes. Combining the three DA magnitude classes to binary labels (DA positive, DA negative) the sinogram method had an overall true positive rate of 65.7% (95% CI [57.6%, 73.7%]), a false positive rate of 7.4% (95% CI [1.6%, 14.5%]), and a false negative rate of 34.3% (95% CI [26.5%, %42.5]), AUC=0.79 (95% CI [0.73, 0.84];

P-value=0.0002), and MCC=0.55, (95% CI [0.45, 0.65]; P-value=0.0002; Figure 4B). In images that were classified as DA positive by the algorithm, 44.5% of DA location predictions exactly matched the labelled slice index. In 89.0% of cases, the predicted location was within 5 slices of the labelled location, 92.3% were within 10 slices of the label, and 92.1% were within 15 slices of the label (Figure 4A). The slice thickness varied between 2.0 and 3.0 mm for different patients and had a median of 2.0 mm.

0.9.2 Convolutional Neural Network Detection

The published CNN [8] was tested on 2319 CT image volumes (945 strong, 606 weak, and 768 without artifacts). These 2319 images are a subset of the full 3211 image volume set that are independent from the training of the published CNN. When used on its own to make binary classifications (DA positive or DA negative) of entire patient CT volumes, the CNN yielded an MCC of 0.82 (P-value=0.0002; Figure 4B) and an AUC of 0.97 (95% CI [0.94, 0.99]; P-value=0.0002; Supplementary Figure 3), in line with the performance of the CNN found in the original study (an AUC = 0.91 $\pm$ STD 0.01) [8].

Supplementary Tables

	Minimum Value	Maximum Value	Median Value
Slice Thickness (mm)	2.0	3.0	2.0
Pixel Spacing (mm)	0.656	1.195	0.976
X-Ray Tube Current (mA)	200	540	300
Number of Slices per Patient	90	333	181

Supplementary Table 1: Details of the acquisition parameters for RADCURE.

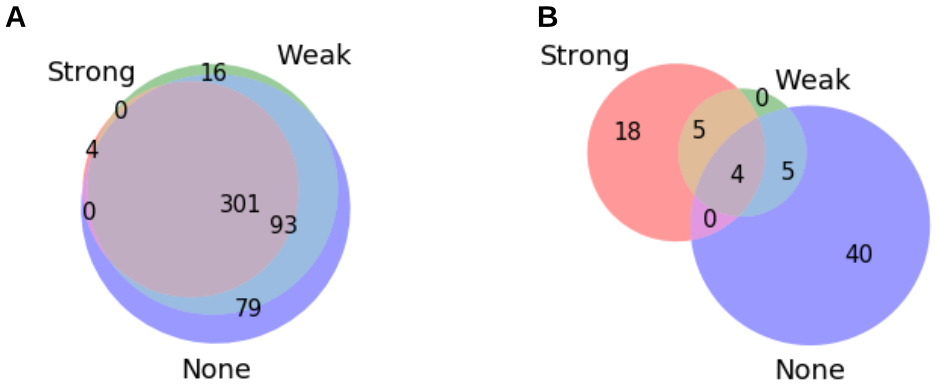
		Annotator 1		
		Strong	Weak	None
Annotator 2	Strong	114	1	4
	Weak	11	51	0
	None	2	65	234

Supplementary Table 2: Contingency table of annotator agreement for DA class.

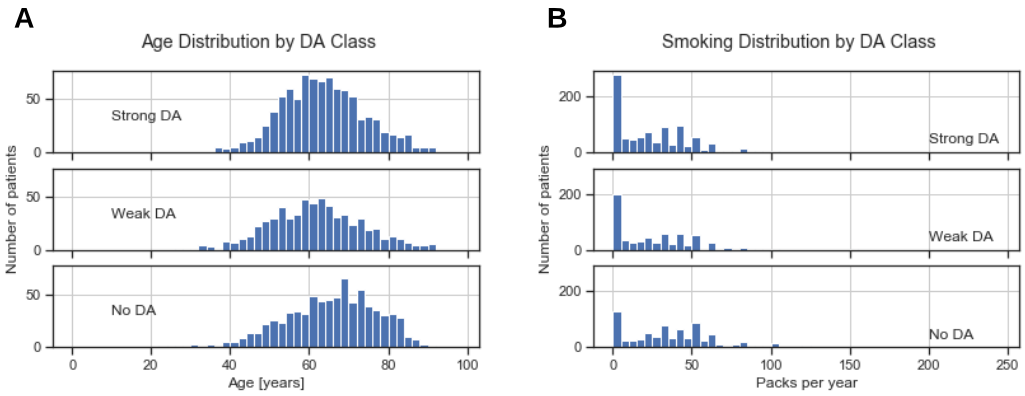
Clinical Variable	χ^2 Statistic	P Value
Sex	1.46	0.23
Smoking Status	38.8	1.90 $\times 10^{-8}$
Primary Disease Site	29.9	1.67 $\times 10^{-3}$
Stage	5.70	0.46

Supplementary Table 3: The results of a χ^2 test for various categorical clinical variables. The test was performed by grouping the data by DA status (strong, weak, none) and testing the distributions of the given clinical variable between each DA group.

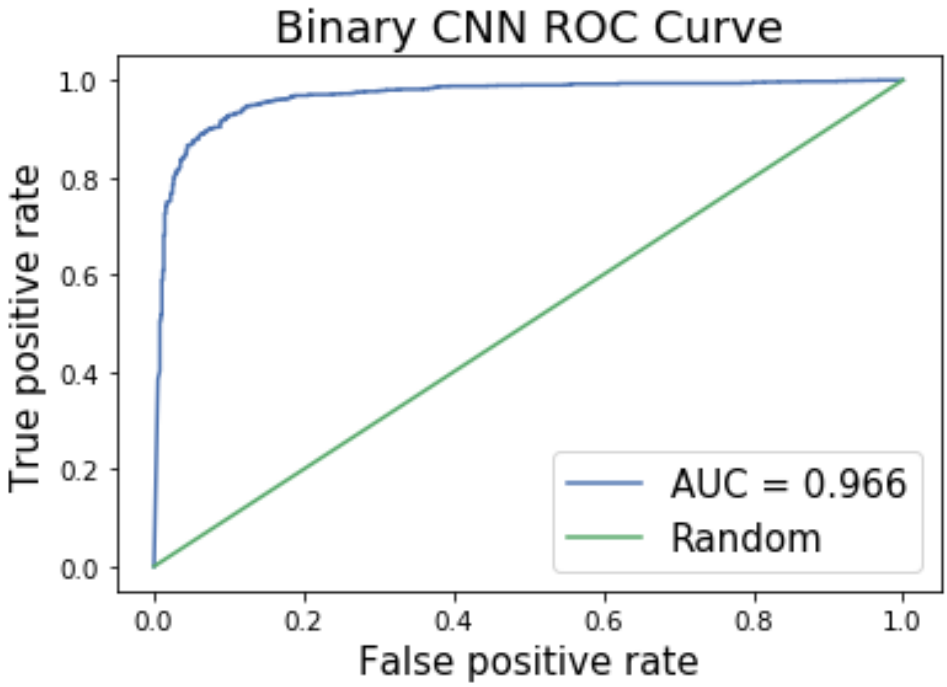
Supplementary Figures



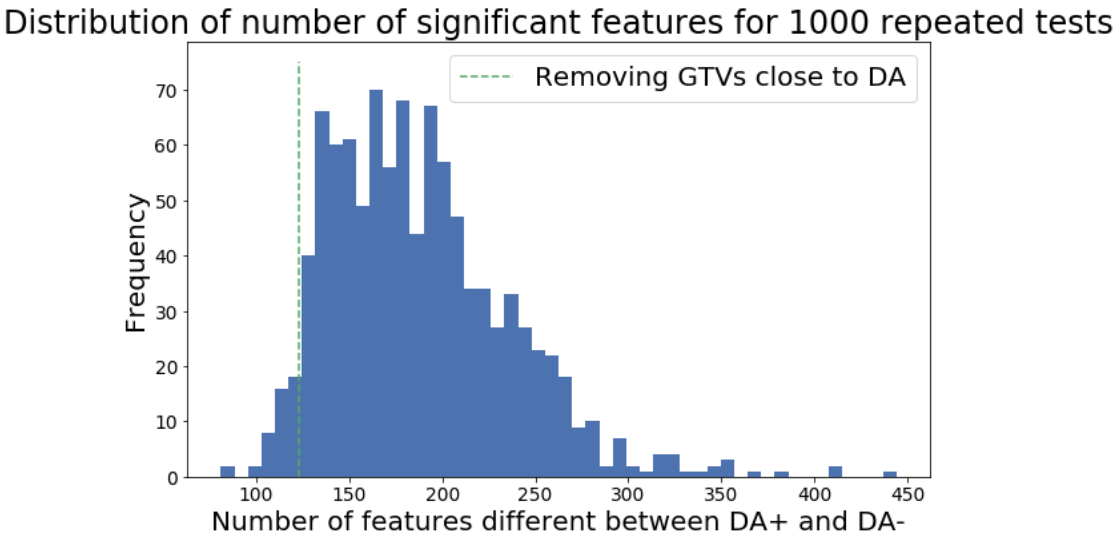
Supplementary Figure 1: Number of PyRadiomic features correlated with distance between the DA and the GTV pixel nearest to the DA. This correlation is computed using the Spearman rank correlation between distance and feature value, computed separately for each DA class. Unlike the results in figure 5, these correlations do not control for volume. We include two thresholds of Spearman R, 0.5 (A) and 0.65 (B) in order to illustrate that volume confounds many distance-correlated radiomic features. This is particularly true for images with no DAs, where correcting for volume removes 31 of the 40 features correlated with DA-GTV distance displayed in this plot.



Supplementary Figure 2: Distributions of two continuous clinical variables between DA classes from the RADCURE dataset. The age distribution is shown on the left, while the number of packs per year is shown on the right.



Supplementary Figure 3: ROC curve for binary DA detection CNN with AUC 0.966.



Supplementary Figure 4: The distribution of the number of significant features between strong-DA and no-DA images for 1000 repeated tests. Each test randomly selected 1006 patients from the full dataset and performed a Wilcoxon rank sum test for each feature between strong-DA and no-DA images. The number of significant features was then calculated for each test with a cutoff of $p < 0.05$. Selectively removing images with the GTV and DA overlapping resulted in 123 significant features, shown with the green dashed line. This value was in the bottom fifth percentile of the repeated random test distribution.

Acronyms

PM	Princess Margaret Cancer Centre
UHN	University Health Network

References

[1] Diehn FE, Michalak GJ, DeLone DR, Kotsenas AL, Lindell EP, Campeau NG, et al. CT Dental Artifact: Comparison of an Iterative Metal Artifact Reduction Technique with Weighted Filtered Back-Projection. Acta Radiol Open. 2017 Nov;6(11):2058460117743279.

[2] Kotsenas AL, Michalak GJ, DeLone DR, Diehn FE, Grant K, Halaweish AF, et al. CT Metal Artifact Reduction in the Spine: Can an Iterative Reconstruction Technique Improve Visualization? AJNR Am J Neuroradiol. 2015 Nov;36(11):2184–2190.

[3] Croxford A, Fajnwaks P, Botkin C, Oliver D, Nguyen N, Osman M. Prevalence and patterns of metal artifacts in FDG PET/CT. J Nucl Med. 2010 May;51(supplement 2):2123–2123.

[4] 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 Jun;5:4006.

[5] van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017 Nov;77(21):e104–e107.

[6] Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern. 1979;9(1):62–66.

[7] van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. PeerJ. 2014 Jun;2:e453.

[8] Welch ML, McIntosh C, Purdie TG, Wee L, Traverso A, Dekker A, et al. Automatic classification of dental artifact status for efficient image veracity checks: effects of image resolution and convolutional neural network depth. Phys Med Biol. 2020 Jan;65(1):015005.

[9] Pedregosa F. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–2830.