SUPPLEMENTARY INFORMATION

A hybrid algorithm for dental artifact detection in large computed tomography datasets

Figures

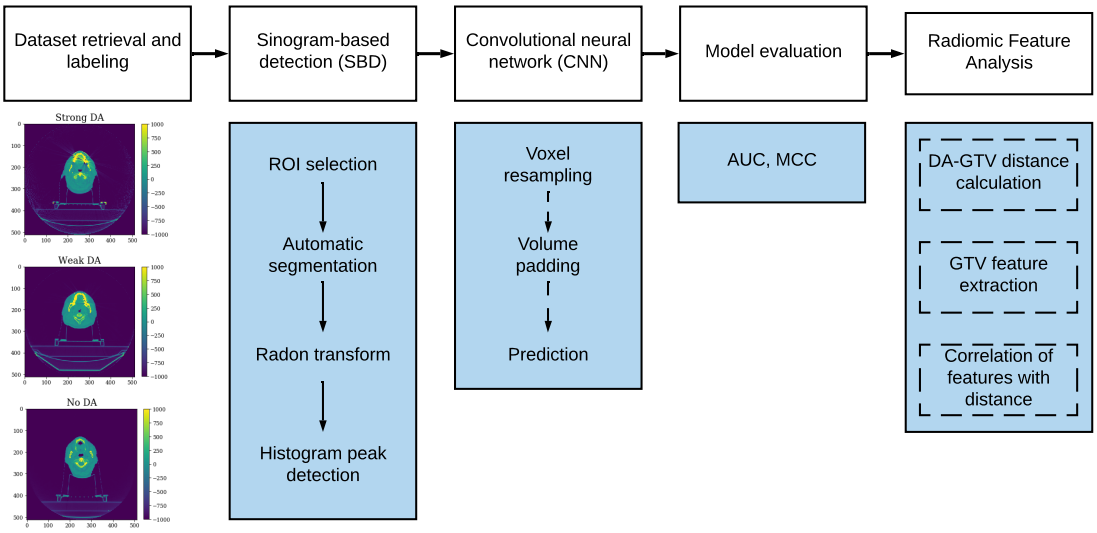


Figure 1: The study design includes five main steps: (1) retrieval of head and neck CT imaging volume dataset and labelling of DA; (2) initial classification of DA using a sinogram-based detection (SBD) method; (3) secondary classification of SBD-classified dental artifacts using a previously trained CNN; (4) model evaluation; and (5) exploration of the effect of DA magnitude and its distance from the GTV on radiomic features.

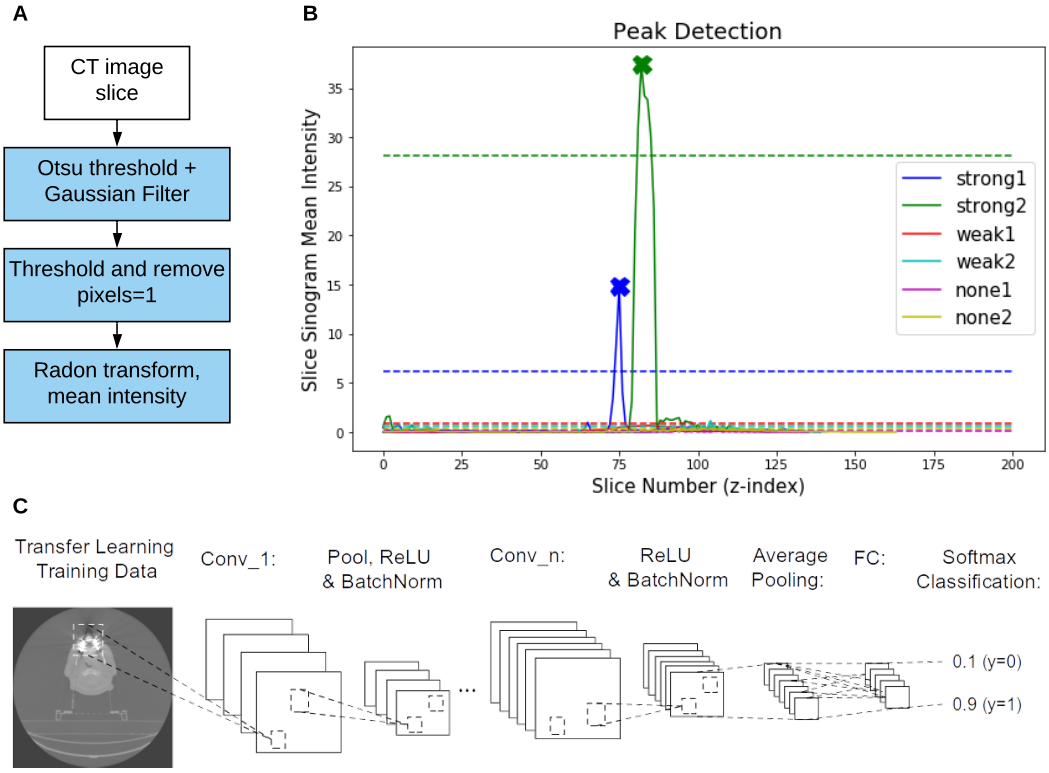


Figure 2: An illustration of the two binary DA classifiers used in this study. (A) Two steps in the sinogram-based detection (SBD). First, one slice from a CT volume is thresholded and blurred, before being thresholded again to remove pixels in the body of the patient. The remaining pixels are thresholded again, revealing the streaks outside the patient's body. The image is then transformed to the sinogram domain and the mean sinogram pixel intensity is computed. (B) An example of the 'mean sinogram intensity' for each slice in six CT volumes (each image represented with a different colour). A peak detection algorithm is applied to this plot for a given patient to detect slices likely to contain DAs. We annotate the detected slices with Xs to show that the algorithm detected one peak from each of the green and blue curves (both images labelled as 'strong DA'). The dashed lines represent the peak detection threshold for each patient. (C) The CNN architecture used in the study. The network consisted of 5 convolutional layers (conv_1 to conv_5) creating a total of 64 filters.

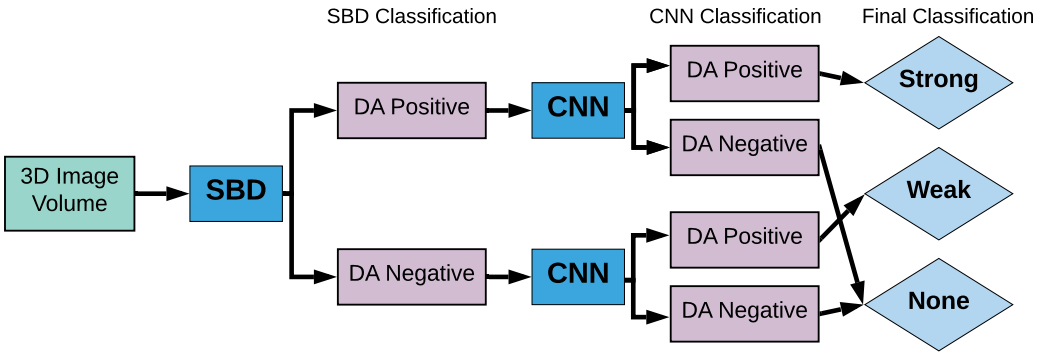


Figure 3: Flowchart of the SBD-CNN hybrid algorithm for dental artifact detection. Images were annotated manually and then first binned using SBD (Sinogram based detection) based on the average intensity of the corresponding sinogram. Subsequently, the original images were classified using the CNN model. Images that were labelled as artifact positive by both the SBD and CNN were categorized as having strong dental artifacts. Images labelled as artifact negative by both methods were labelled as having no artifacts. This way our hybrid model is capable of labelling images based on the strength of artifact presence.

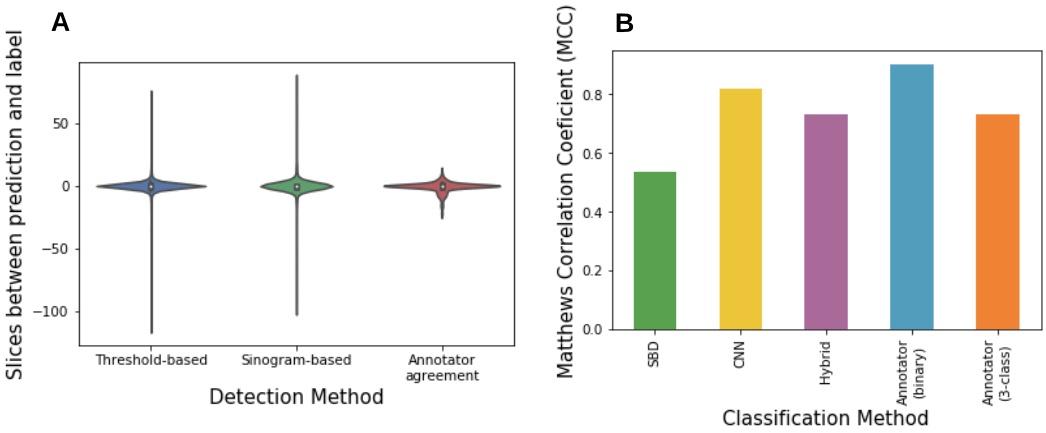


Figure 4: Performance of DA classification. (A) Distributions of how close the predicted slice index is to the labelled index for the threshold-based and sinogram based-detection methods (e.g. $\|i_{\text{predicted}} - i_{\text{labelled}}\|$). The difference in slice label between two human annotators for a set of 482 CT volumes is also shown. (B) Performance (MCC) of the DA magnitude classification techniques used in this study. The p-value of the MCC for all classifiers was < 0.001). The sinogram-based detection (SBD) and convolutional neural network (CNN) are both binary classifiers. The SBD was tested on 3,211 CT image volumes and the CNN binary classifier was tested on a subset of 2,319 image volumes. The SBD-CNN hybrid algorithm is a three-class classifier and the three-class MCC is therefore displayed here.

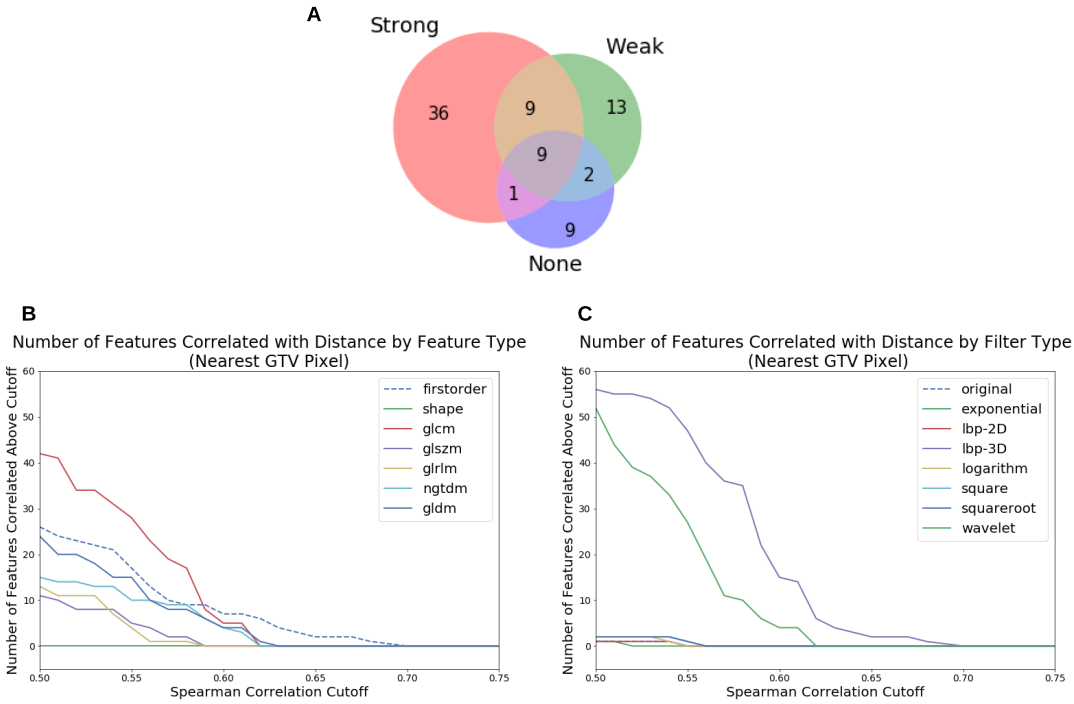


Figure 5: Correlation between GTV-DA distance and feature values, based on the partial correlation using Spearman correlation. (A) Venn diagram showing the number of features with $|r| > 0.55$ calculated from patients from each DA class. This diagram only includes significant correlations ($p < 0.05$). For instance, 36 features had $|r| > 0.55$ and were found in patients with strong DAs (pink region), but those features had $|r| < 0.55$ when calculated from weak or no-DA images). Nine features had $|r| > 0.55$ when calculated for all three DA groups (grey region). (B) The number of features with DA-GTV distance correlation above a given cutoff, grouped by feature type. (C) These correlations grouped by filter type. (B) and (C) only include significant features ($p < 0.05$).

Supplementary Methods

0.1 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.

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.

0.2 Confounding Factors in PyRadiomic Feature Analysis

0.2.1 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.

0.2.2 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).

0.2.3 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 Table 5. 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 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.

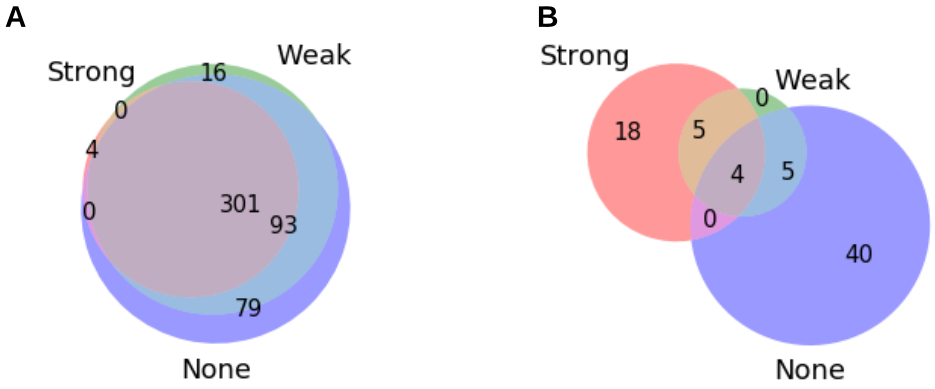
		Annotator 1		
		Strong	Weak	None
Annotator 2	Strong	114	1	4
	Weak	11	51	0
	None	2	65	234

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

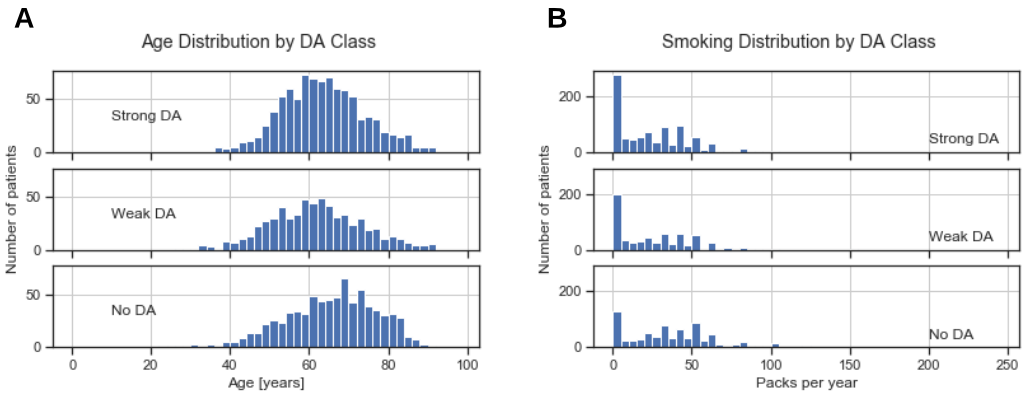
Clinical Variable	χ^2 Statistic	P Value
Sex	1.46	0.23
Smoking Status	38.8	1.90 $\times 10^{-8}$
Primary Disease Site	29.9	1.67 $\times 10^{-3}$
Stage	5.70	0.46

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.

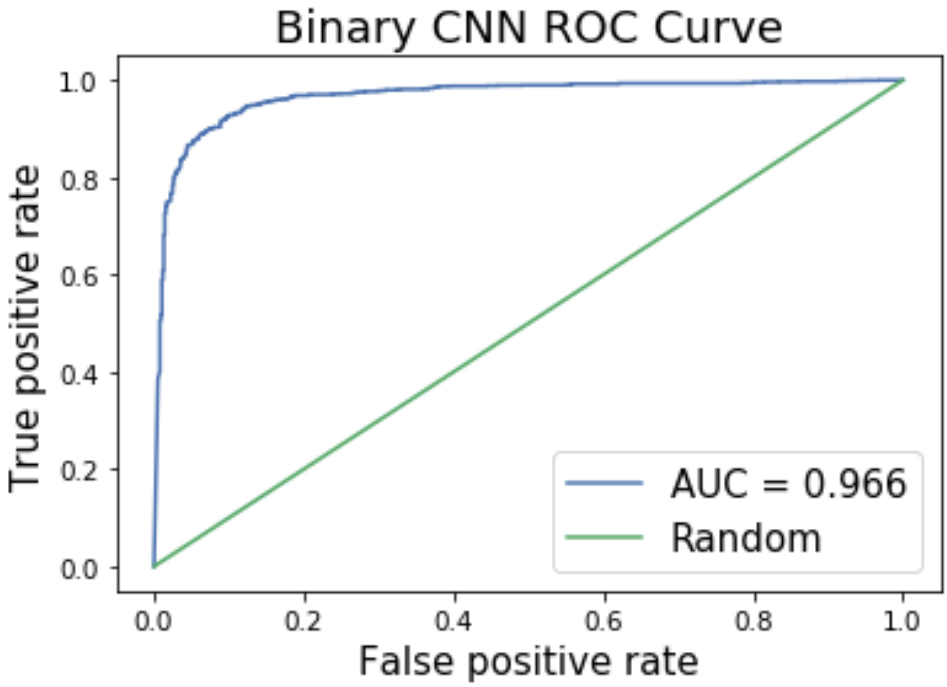
Supplementary Figures



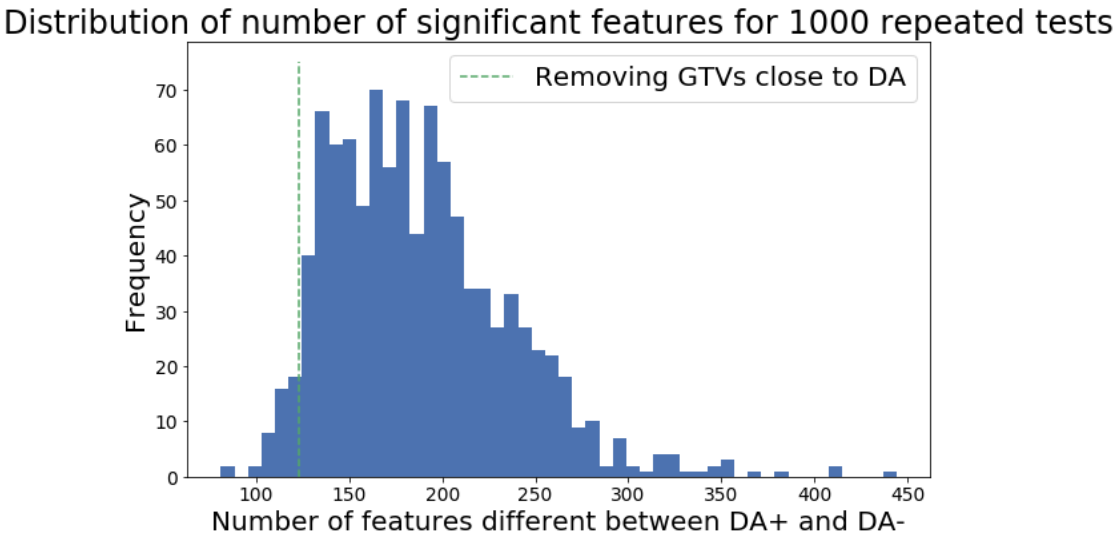
Supplementary Figure 1: Number of PyRadiomic features correlated with distance between the DA and the GTV pixel nearest to the DA. This correlation is computed using the Spearman rank correlation between distance and feature value, computed separately for each DA class. Unlike the results in figure 5, 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

Supplementary Files

Supplementary File 1. xxx

Supplementary File 2. xxx

Supplementary File 3. xxx