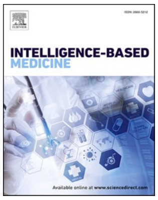
Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine



Machine learning prediction of Dice similarity coefficient for validation of deformable image registration

Yun Ming Wong [a,1,*], Ping Lin Yeap [b,c], Ashley Li Kuan Ong [b], Jeffrey Kit Loong Tuan [b], Wen Siang Lew [a], James Cheow Lei Lee [a,b], Hong Qi Tan [a,b]

[a] *Division of Physics and Applied Physics, Nanyang Technological University, Singapore*
[b] *Division of Radiation Oncology, National Cancer Centre Singapore, Singapore*
[c] *Department of Oncology, University of Cambridge, United Kingdom*

ARTICLE INFO

Keywords:
Deformable image registration
Adaptive radiotherapy
Machine learning
Dice similarity coefficient
DVF-based metrics

ABSTRACT

Introduction: Deformable image registration (DIR) plays a vital role in adaptive radiotherapy (ART). For the clinical implementation of DIR, evaluation of deformation accuracy is a critical step. While contour-based metrics, for example Dice similarity coefficient (DSC), are widely implemented for DIR validation, they require delineation of contours which is time-consuming and would cause hold-ups in an ART workflow. Therefore, this work aims to accomplish the prediction of DSC using various metrics based on deformation vector field (DVF) by applying machine learning (ML), in order to provide an efficient means of DIR validation with minimised human intervention.

Methods: Planning CT image was deformed to the cone-beam CT images for 20 prostate cancer patients. Various DVF-based metrics and DSC were calculated, and the former was used as input features to predict the latter using three ML models, namely linear regression (LR), Nu Support Vector Regression (NuSVR) and Random Forest Regressor (RFR). Four datasets were used for analysis: 1) prostate, 2) bladder, 3) rectum and 4) all the organs combined. Average mean absolute error (MAE) was computed to evaluate the model performance. The classification performance of the best-performing model was further evaluated, and the prediction interval and feature importance were calculated.

Results: Overall, RFR achieved the lowest average MAE, ranging between 0.045 and 0.069 for the four datasets, while LR and NuSVR had slightly poorer performances. Analysis on the results of best-performing model showed that sensitivity and specificity of 0.86 and 0.51, respectively, were obtained when a prediction threshold of 0.85 was used to classify the fourth dataset. Jacobian determinant was found to be a significant contributor to the predictions of all four datasets using this model.

Conclusion: This study demonstrated the potential of several ML models, especially RFR, to be applied for prediction of DSC to speed up the DIR validation process.

1. Introduction

Adaptive radiotherapy (ART) is an important part of the radiotherapy workflow for certain clinical situations where treatment plans are modified based on the anatomic changes of patients [1]. This is especially crucial for highly conformal treatment techniques such as proton therapy [2], in which a small variation can bring about a large impact to the daily dose distribution, as well as hypofractionated radiotherapy.

Depending on the time scale at which adaptation is performed, three classes of ART have been defined: offline ART, online ART, and real-time (in-line) ART [2,3]. Offline ART takes place over the course of a few days and aims to address systematic variations between treatment fractions. Due to the inherent lag of offline ART to accommodate to the latest change, online ART presents as a superior approach by performing adaptation right before a treatment fraction. Real-time ART, on the other hand, extends to an even shorter time frame by adapting the plan based on real-time changes detected directly before the treatment itself.

* Corresponding author. 21 Nanyang Link, 637371, Singapore.
E-mail address: yunmingwong07@gmail.com (Y.M. Wong).

[1] First authors.

https://doi.org/10.1016/j.ibmed.2024.100163
Received 13 March 2023; Received in revised form 2 July 2024; Accepted 7 August 2024
Available online 8 August 2024

For a successful implementation of ART, image registration, defined as the process of transforming one image to another, is an indispensable step. There are three major categories of image registration, namely rigid image registration, affine image registration and deformable image registration. Among all, deformable image registration (DIR) offers the greatest number of degrees of freedom (could be as large as three times the number of voxels) as each individual voxel can be separately transformed in the x, y and z directions [4]. Therefore, it plays an important role to account for the non-rigid anatomic changes of organs and tumours. The collection of vectors denoting the deformation magnitude and direction for each voxel forms the deformation vector field (DVF).

Since its introduction, DIR has seen an escalating application in radiotherapy. Some examples include contour propagation and dose accumulation [5], which are of particular interest for an ART workflow. To ensure a reliable clinical translation, there is a need for validation of DIR to verify the deformation accuracy. Validation of DIR can be done by means of physical phantom, virtual phantom or patient-specific metrics [6]. The lattermost option can be further sub-divided into image-based metrics, among which contour-based metrics are the most widely used [7], and DVF-based metrics in terms of geometric accuracy evaluation [6]. Geometric accuracy evaluation is important for assessing the quality of contour mapping, which serves to propagate contours from one image to another.

Several contour-based metrics have been well-defined and frequently studied in the literature, for instance, Dice similarity coefficient (DSC) [8–13] and Hausdorff distance (HD) [8,13]. However, these contour-based metrics require the manual delineation of contours on the daily CBCT images by radiation oncologists, which is laborious in nature and would pose a hindrance to the smooth integration of ART into the daily clinical workflow. DVF-based metrics, on the contrary, have the advantage of being readily available after completion of a DIR. Yet, this type of metrics did not gain as much popularity due to the lack of a strongly proven correlation with the contour-based metrics. To leverage the potential of DVF-based metrics for DIR validation, machine learning could be a promising approach in building a model for prediction of contour-based metrics using DVF-based metrics.

Very limited studies could be found in the literature that applied machine learning to predict DIR accuracy. One of the earlier works involved the use of support vector machine (SVM) to classify the registration as a failed or successful registration based on similarity measures [14]. Another study predicted registration error obtained from manually annotated landmarks using registration-based and intensity-based features [15]. In addition, Dushepa created training and test sets using Monte Carlo simulation and subsequently quantified registration accuracy using features such as bootstrap features, pixel intensity differences and spectrum width [16]. All of these, however, do not involve the prediction of contour-based metrics, which are the key elements for evaluating contour mapping accuracy of a DIR algorithm. As discussed earlier, manual delineation of contours is a tedious process and because of that, contour-based metrics are not always easily obtainable. On that account, this work aims to accomplish the accuracy evaluation step without the hassle of generating manual contours, by applying machine learning to identify a model for predicting DSC, a contour-based metric, using DVF-based metrics as input features. With this model being established, it would be possible to realise a high throughput evaluation of DIR accuracy, so that the quality of daily adaptation could be verified for every single fraction during offline or even online ART without causing much delay to the treatment workflow.

2. Methods

2.1. Patient Characteristics and data

This study has been approved by the Institutional Review Board. Retrospective data from 20 low-risk prostate cancer patients, treated between 2016 and 2019 at National Cancer Centre Singapore (NCCS), were obtained. These patients underwent 37 to 39 fractions throughout their entire treatment course. For each patient, the planning CT (pCT) and cone-beam computed tomography (CBCT) images taken before each treatment fraction along with the manually drawn contours were imported into RayStation 10A (RaySearch Laboratories, Stockholm, Sweden). The pCT images have a slice thickness of 2/2.5 mm and pixel spacing of 0.7734 mm–1.0742 mm, while the CBCT images have a slice thickness of 2.5 mm and pixel spacing of 1.1719 mm. The CBCT images were taken using Varian machine (Siemens Healthineers, Forchheim, Germany). Three organs were included for analysis, namely prostate, bladder and rectum.

2.2. Parameter optimisation and exploratory data analysis

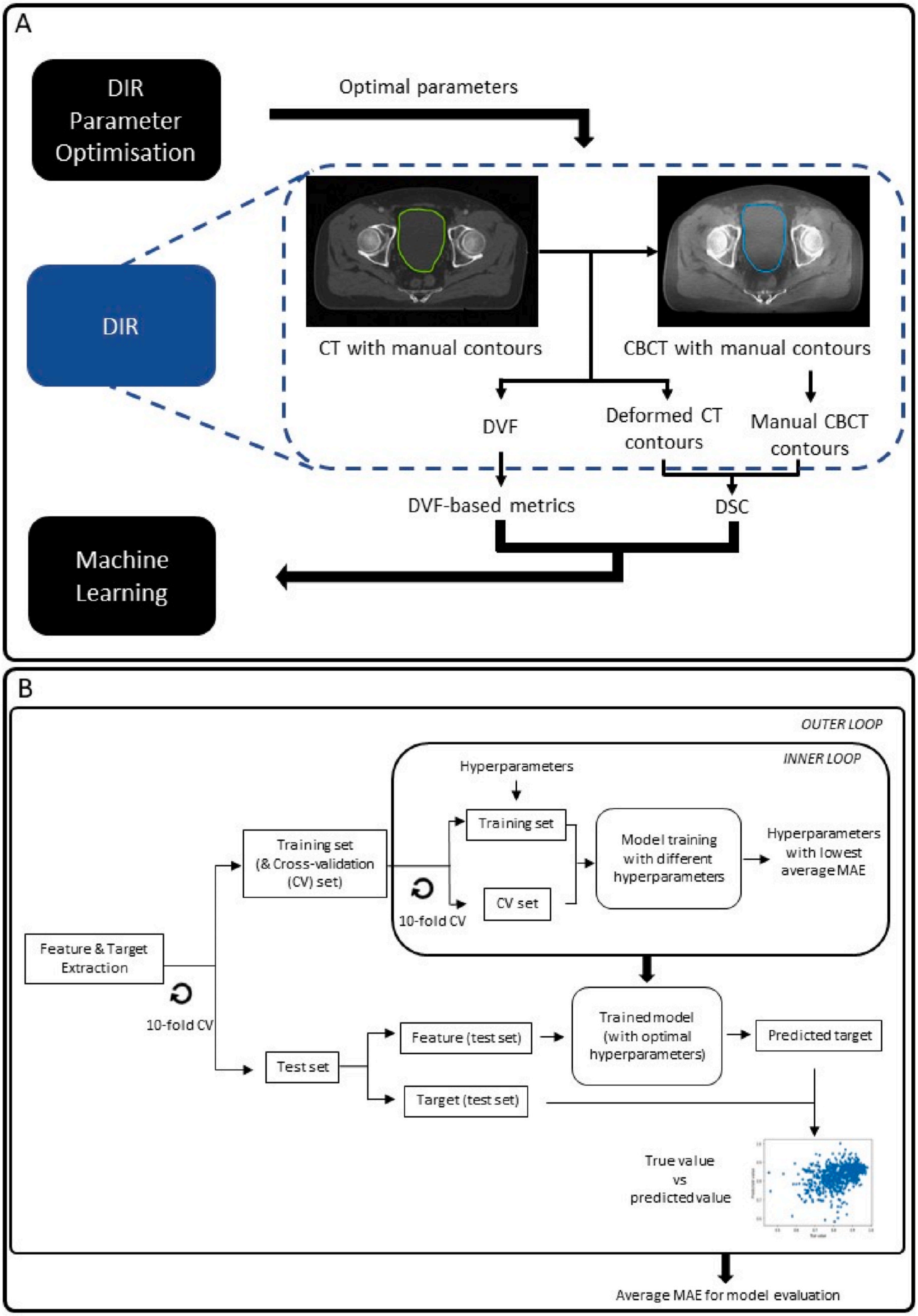
Before extracting the data for machine learning, optimisation of parameters for the DIR in RayStation was performed using data from five patients, selected randomly from the complete set of 20 patients. Hybrid deformable registration, optimised based on an objective function comprising of an image similarity term, grid regularization terms, and anatomical penalty terms, was done. The pCT image was assigned as the reference image while the fractional CBCT images taken at an interval of three fractions served as the target images. Four different final resolutions (0.1 cm, 0.2 cm, 0.3 cm, 0.4 cm) and two different similarity measures (Correlation Coefficient (CC), Mutual Information (MI)) were tested within RayStation to determine the parameters that give the best registration quantified by the DSC and image similarity metrics.

Using the optimal parameters, hybrid deformable registrations were performed using the pCT image as the reference image while all the fractional CBCT images constituted the target images. The distributions of DSC for all patients and each individual patient were illustrated based on sites. Subsequently, Kruskal-Wallis test, was done to test the null hypothesis that all the patients have similar distributions of DSC values. On top of that, Spearman correlation test was performed between DSC and treatment fractions, organ volume and treatment fractions, as well as DSC and organ volume, for each of the three organs. A *p*-value of 0.05 was used for both two-tailed tests to mark the significance of the null hypotheses.

2.3. Machine learning

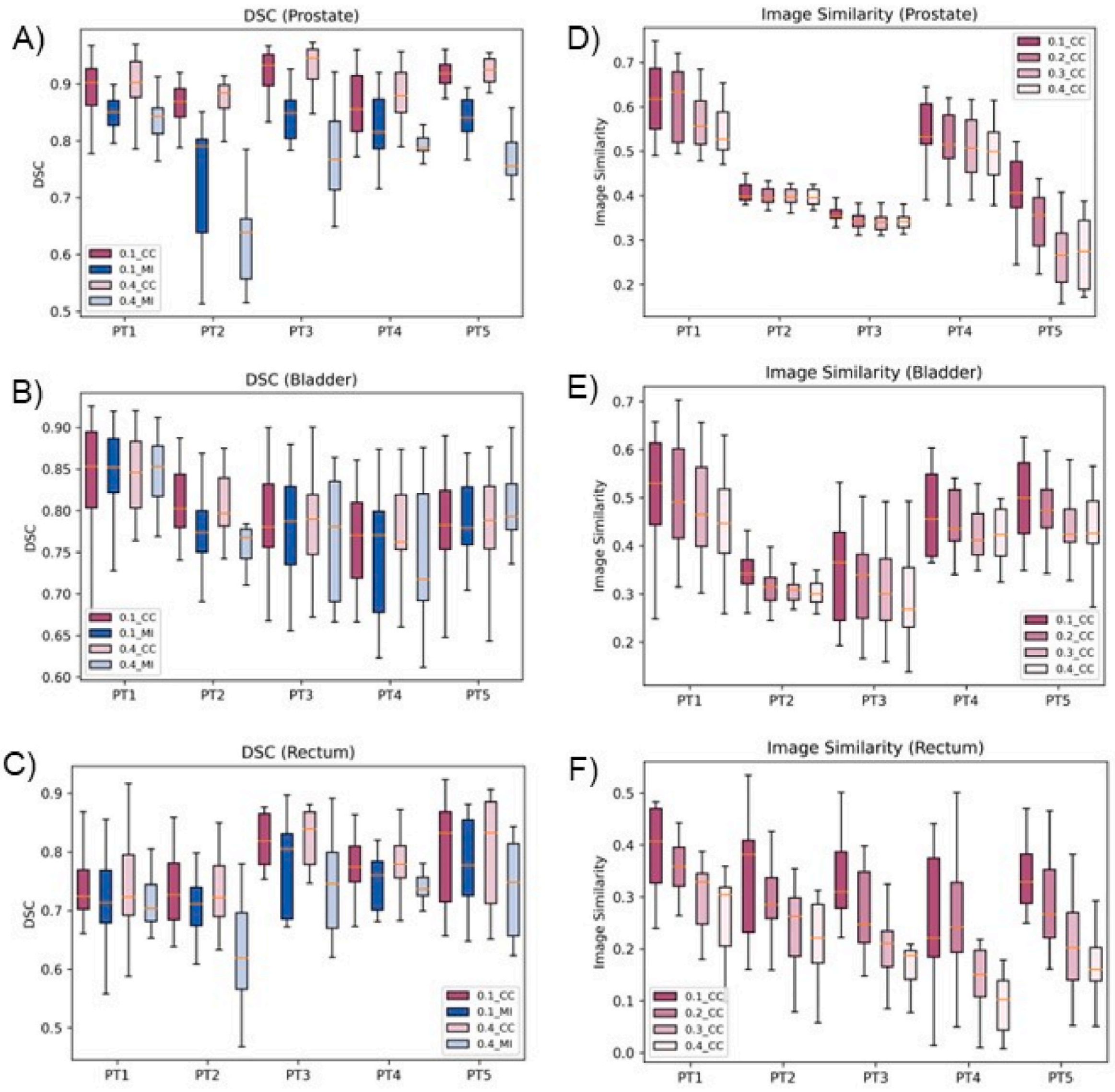
A total of 50 DVF-based metrics were extracted from RayStation, normalized and used as the features to predict the corresponding DSC for each registration, which is the target variable in this case. The DVF-based metrics include 40 descriptive statistics (for example, minimum and mean) of the DVF magnitude and 10 descriptive statistics (mean and various percentiles) of the Jacobian determinant, as summarised in Table S1. The former represents the deformation extent of each voxel in a specific region, whereas the latter signifies the volume change of that region. New regions of interest, collectively known as the “organ ring”, were created on the pCT image through expansion and shrinkage of the contour for each organ under analysis by 0.25 cm, resulting in a total thickness of 0.5 cm. This was set to be slightly greater than the bladder and rectum wall thickness, which is typically close to 3 mm or less [17,18], and served to account for deformation happening at the surface of each organ.

Analysis was done on four sets of data: 1) prostate only, 2) bladder only, 3) rectum only, and 4) all the organs combined. The first three sets have the same total number of samples (761) while the last set has three times as many (2283). Machine learning algorithms were implemented using the scikit-learn v1.0.2 package [19]. Three different models, linear regression (LR), Nu Support Vector Regression (NuSVR) and Random Forest Regressor (RFR) were used for predictive modelling. LR fits a linear model and aims to minimise the residual sum of squares between the target and prediction [20]. NuSVR is a regression model based on SVM [21] and uses a parameter called nu to control the number of


A
DIR Parameter Optimisation
Optimal parameters
DIR
CT with manual contours
CBCT with manual contours
DVF
Deformed CT contours
Manual CBCT contours
DVF-based metrics
DSC
Machine Learning
B
OUTER LOOP
INNER LOOP
Hyperparameters
Training set
CV set
10-fold CV
Model training with different hyperparameters
Hyperparameters with lowest average MAE
Training set (& Cross-validation (CV) set)
Feature & Target Extraction
10-fold CV
Test set
Feature (test set)
Target (test set)
Trained model (with optimal hyperparameters)
Predicted target
True value vs predicted value
Average MAE for model evaluation

Acronyms: DIR, deformable image registration; CT, computed tomography; CBCT, cone-beam computed tomography; DVF, deformation vector field; DSC, Dice similarity coefficient; MAE, mean absolute error.

Fig. 1. (A) Outline of major steps performed in this study. The machine learning pipeline is shown in detail in (B).



Acronyms: DSC, Dice similarity coefficient; CC, correlation coefficient; MI, mutual information.

Fig. 2. Boxplots for DIR parameter optimisation.
Figures A, B and C show the DSC values achieved using different combinations of final resolution and similarity measure for prostate, bladder, and rectum of five patients, respectively. For clarity, only two final resolutions (0.1 cm and 0.4 cm) are displayed. Figures D, E and F show the image similarity achieved using CC as the similarity measure and different final resolutions for prostate, bladder, and rectum of five patients, respectively.

support vectors. Lastly, RFR [22] uses a number of decision trees with various sub-samples, and predicts the target by averaging the results given by all the trees in the forest.

A nested cross-validation approach is used to evaluate the model performance. To attain the best performance for NuSVR and RFR, the hyperparameters were optimised through an inner 10-fold cross-validation. The hyperparameters varied are summarised in Table S2. In a single run, every different combination of hyperparameters would give a corresponding mean absolute error (MAE). The resulting 10 sets of results were then averaged to determine the combination of hyperparameters with the lowest average MAE, which would be the optimal hyperparameters in this case. This step of hyperparameter tuning constitutes the inner loop, where its output (the optimal hyperparameters) was fed into the outer loop to predict the test set. Similarly, 10-fold validation was used in the outer loop and the average MAE across the test set in each of the fold was computed to evaluate the model performance.

Using this nested cross-validation approach, the training set constituted 81 % of the whole dataset, while the cross-validation set and test set took up 9 % and 10 %, respectively. These contain 616, 69, and 76 samples, respectively for the first three datasets for each individual organ, and 1849, 206, and 228 samples, respectively for the fourth dataset with all three organs combined. A schematic showing the overview of the major steps employed in this study can be found in Fig. 1.

2.4. *Clinical evaluation using best performing model*

To classify the quality of contour mapping as either good or bad, a prediction threshold can be determined for the datasets predicted by the best performing model. Any DSC values below the threshold will be considered a bad mapping, hence requiring physician to manually review the registered contour. In contrast, a DSC value above the threshold will signify a good registration, that is, a trustable registered contour. As

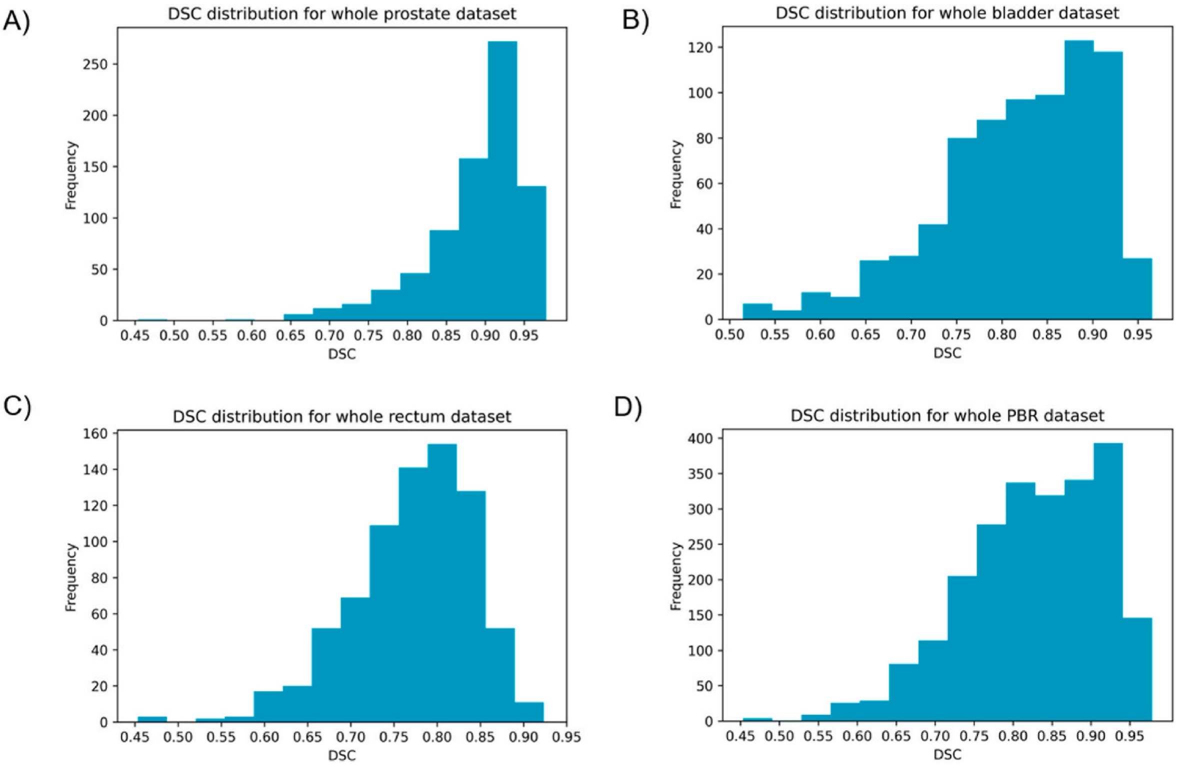


Fig. 3. DSC distributions of the complete dataset for (A) prostate, (B) bladder, (C) rectum, and (D) all three organs.

the commonly used standard for an acceptable DSC value ranges from around 0.8 to 0.9 [4], analysis was performed using prediction thresholds close to this range: 0.75, 0.80, 0.85 and 0.90.

Four performance metrics, namely sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), were calculated. A positive case denotes a DSC value below the threshold (bad contour mapping) and vice versa for a negative case. From a clinical perspective, it is critical to flag a bad contour mapping, so that appropriate action can be taken to rectify the poorly registered contour. Therefore, sensitivity would be given a higher priority compared to specificity. Additionally, 68 % prediction interval (PI) was calculated using the range between 16th percentile and 84th percentile of out-of-bag errors generated by RFR. Lastly, feature importance was also calculated to identify the features with a greater impact on the predictions for each dataset.

3. Results

3.1. Parameter optimisation and exploratory data analysis

The boxplots depicting the DSC results for DIR using different combinations of final resolution and similarity measure for five patients are shown in Fig. 2A–C. For clearer visualisation, only results for final resolutions of 0.1 cm and 0.4 cm are plotted. The complete DSC results can be found in Fig. S1. From Fig. 2A–C, it is obvious that CC outperforms MI in most of the cases. As our current focus is on the application of DIR for contour mapping, DSC was given precedence for the performance evaluation and hence MI was ruled out. To select the best final resolution to be used for the registrations, the boxplots for image similarity were examined. Specifically, the boxplots corresponding to registrations using CC as the similarity measure were compared, as shown in Fig. 2D–F. A generally decreasing trend was observed for the image similarity when the final resolution increased. Therefore, 0.1 cm and CC were chosen as the optimal parameters for DIR in the subsequent parts of the study.

Table 1
Average MAE for different prediction models, classified based on different datasets. P, B, R, and PBR denote dataset for prostate, bladder, rectum, and all three organs, respectively.

	P	B	R	PBR
LR	0.049 ± 0.016	0.068 ± 0.013	0.056 ± 0.009	0.063 ± 0.009
NuSVR	0.045 ± 0.017	0.072 ± 0.014	0.056 ± 0.011	0.061 ± 0.009
RFR	0.045 ± 0.013	0.069 ± 0.008	0.053 ± 0.009	0.060 ± 0.009

Fig. 3 shows the DSC distributions for all the patients, while Fig. S2 shows the distributions for each individual patient, classified based on sites. The Kruskal-Wallis test gave a significant statistic ($P < 0.05$) that there is a difference between DSC distributions of all the patients. The H scores with their corresponding *p*-values for each dataset are tabulated in Table S3. The subsequent Spearman correlation test revealed a higher occurrence of significant correlation between DSC and organ volumes, with at least half of the total patients exhibiting either a positive or negative correlation. Interestingly, a statistically significant ($P \ll 0.05$) strongly positive correlation ($r_s > 0.80$) was observed for a number of patients in the bladder dataset. These findings are depicted with a correlation matrix as shown in Fig. S3.

3.2. Machine learning

The hyperparameters selected in each loop during training of NuSVR and RFR models are shown in Table S4. Despite there being no discernible trend in the hyperparameters, each loop produced comparable results, indicated by the standard deviation close to 0.01 as shown in Table 1. Fig. 4 shows the stacked histograms illustrating the distributions of the absolute error against the true DSC values for RFR (Please refer to Figs. S4–5 for similar histograms for LR and NuSVR.). A substantial portion of the errors is smaller than 0.05 and these errors are found close to the peak of each distribution, ranging from approximately 0.75 to 0.95. Conversely, errors greater than 0.15 are mostly found at

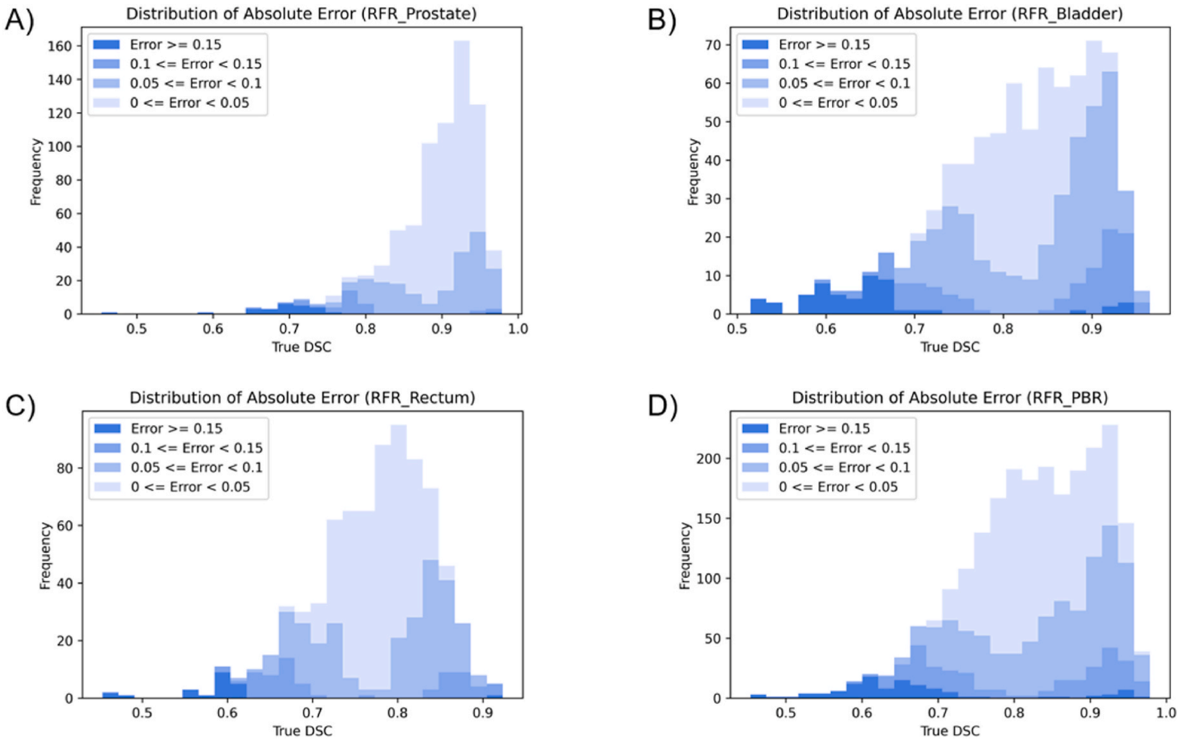


Fig. 4. Distribution of absolute error for predictions using RFR model for (A) prostate, (B) bladder, (C) rectum, and (D) all three organs.

the extreme ends of the distribution.

The average MAE for each model were reported in Table 1, for a more quantitative comparison. It is observed that RFR performed best among the three models, with the lowest average MAE in general. On the other hand, LR and NuSVR had slightly higher average MAE compared to RFR in most of the cases, while showing a comparable accuracy to each other. Despite these variations, the standard deviation of the score for all three models remained below 0.02, indicating the stability of the models.

3.3. *Clinical evaluation using best performing model*

The sensitivity, specificity, PPV and NPV to evaluate the classification performance with prediction thresholds of 0.75, 0.80, 0.85 and 0.90 are shown in Table 2.

PBR_0.85 was found to give a sensitivity of 0.86, specificity of 0.51, PPV of 0.69 and NPV of 0.74. This means that 86 % of the bad mapping cases are picked out for further review, while 51 % of the good mapping cases are spared the extra check, thus saving time and effort required during an ART workflow.

The PBR dataset was divided into its constituent datasets of prostate, bladder and rectum for further analysis. Targeting a sensitivity of at least 80 %, it was found that P_const_0.90 gave a sensitivity of 0.93, specificity of 0.16, PPV of 0.48 and NPV of 0.74, B_const_0.85 gave a sensitivity of 0.87, specificity of 0.21, PPV of 0.60 and NPV of 0.55, and R_const_0.80 gave a sensitivity of 0.79, specificity of 0.31, PPV of 0.65 and NPV of 0.48. All the PPV and NPV are either near to 50 % or even higher.

Table 3 shows the average predicted DSC values by the RFR model with their corresponding average 68 % PI for each range of true DSC values with an interval of 0.10, classified based on the four datasets. It is noticeable that for the lower ranges of true DSC values (0.40–0.50, 0.50–0.60) across all the datasets, the predicted DSC values and their PI showed higher discrepancies compared to the true DSC values in the test set. This finding agrees with what has been reported in section 3.2, where larger errors are found at the extreme ends of the true DSC distributions. Of particular note is that the prostate dataset with true DSC range of 0.40–0.50 gave a predicted DSC and PI of around 0.90, which is a pronounced contradiction. This is due to the range containing merely one sample, making it hard for the model to predict this value without prior "training".

The feature importance calculated using each dataset is shown in Fig. 5, focusing on the top 10 most important features. It is noteworthy that the percentiles of the Jacobian determinant occupied a large proportion of this region. For three out of the four datasets (prostate, bladder and all three organs), the 10th, 20th, and 30th percentiles were always found within the first sixth most important features among all the 50 features fed into the model. This suggests that the information about expansion and shrinkage served as a good indicator of the contour mapping accuracy.

Considering the metrics in the separate directions, it is noticed that metrics in the z-direction (superior-inferior, SI) influenced the predictions for the prostate and bladder datasets more, while those in the y-direction (posterior-anterior, PA) and x-direction (right-left, RL) had a greater impact on the datasets of rectum and all three organs, respectively.

4. Discussion

In this study, three machine learning models (LR, NuSVR and RFR) were trained and their DSC prediction accuracies were quantified using average MAE. We found that RFR could predict DSC with the highest accuracy. Using RFR and a prediction threshold of 0.85 to classify the dataset with all three organs, a sensitivity of 0.86 and specificity of 0.51 were achieved, while both PPV and NPV were close to 0.70. Separate analysis on the constituent datasets identified a differential threshold in order to achieve a minimum sensitivity of 0.80 for all organs, which is presumably due to the distinct DSC distributions of each dataset. The relatively low specificity, however, would entail a less efficient workflow than what is desired. This presents a room for improvement which could be accomplished with an increase in prediction accuracy. The availability of a larger dataset for training is expected to be helpful in

Table 2
Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for each dataset using a prediction threshold of 0.75, 0.80, 0.85 and 0.90. P, B, R, and PBR denote dataset for prostate, bladder, rectum, and all three organs, respectively. In the main text, the term "(Site)_(Threshold)" and "(Site)_const_(Threshold)" will be used to represent the results using the dataset of a certain site and the stated prediction threshold for (A) the original datasets and (B) the constituent datasets of PBR, respectively.

A)

Threshold	Metric	P	B	R	PBR
0.75	**Sensitivity**	0.00	0.04	0.24	0.10
	Specificity	1.00	0.99	0.87	0.97
	PPV	–	0.50	0.48	0.45
	NPV	0.96	0.80	0.69	0.81
0.80	**Sensitivity**	0.14	0.35	0.93	0.58
	Specificity	0.99	0.78	0.16	0.78
	PPV	0.71	0.49	0.64	0.60
	NPV	0.91	0.67	0.58	0.77
0.85	**Sensitivity**	0.18	0.88	1.00	0.86
	Specificity	0.94	0.19	0.00	0.51
	PPV	0.44	0.59	0.89	0.69
	NPV	0.81	0.54	–	0.74
0.90	**Sensitivity**	0.75	1.00	1.00	0.99
	Specificity	0.44	0.00	0.00	0.12
	PPV	0.53	0.80	0.99	0.77
	NPV	0.68	0.00	–	0.73

B)

Threshold	Metric	P	B	R
0.75	**Sensitivity**	0.03	0.01	0.17
	Specificity	0.99	1.00	0.90
	PPV	0.20	0.33	0.46
	NPV	0.96	0.79	0.68
0.80	**Sensitivity**	0.24	0.32	0.79
	Specificity	0.94	0.83	0.31
	PPV	0.31	0.52	0.65
	NPV	0.92	0.67	0.48
0.85	**Sensitivity**	0.41	0.87	0.96
	Specificity	0.74	0.21	0.05
	PPV	0.30	0.60	0.90
	NPV	0.82	0.55	0.14
0.90	**Sensitivity**	0.93	1.00	1.00
	Specificity	0.16	0.01	0.00
	PPV	0.48	0.80	0.99
	NPV	0.74	0.50	–

this.

It is noticeable that the prostate dataset always has the lowest MAE for all the models and vice versa for the bladder dataset. This indicates that the prostate DSC values predicted are more accurate. This could be attributed to the distribution of DSC values across the whole dataset, as shown in the histograms in Fig. 3. As there are generally less data points with DSC values smaller than 0.75, it is difficult for the models to predict the low DSC values for any organ accurately. This could explain the higher MAE of predicted DSC values for bladder.

Analysis on the feature importance revealed that the Jacobian determinant is an important factor in determining the predictions of DSC using the best performing model, that is, RFR. Indeed, Jacobian determinant have been broadly investigated in past studies where it acted as a measure of the physical plausibility of a DIR [13,23–27], which should reflect the DIR quality to a certain extent. On the other hand, the greater influence of the metrics in different directions for different datasets could be linked to the preferred motion direction of the organs. According to a deformation model described by Stoll et al. [28], the bladder mainly deforms in the SI direction. Studies have also shown that the prostate motion is usually in the PA and SI directions [29–33], with a number suggesting a stronger tendency in the SI direction [30–32]. Meanwhile, it has long been recognized that the prostate motion in the PA direction is more strongly correlated to the rectal volume than the bladder volume [34], from which we deduce the substantiality of PA rectum deformation due to rectal filling. Using a similar reasoning, this could explain the higher significance of metrics in the PA direction for predictions in the rectum dataset. Nonetheless, metrics in these two directions carried relatively less weights when all the datasets are combined, as the model attempted to account for datasets of a greater variety. This could be the reason that more metrics in the RL direction are brought to places with a higher importance instead.

Table 3
Average predicted DSC with its corresponding average 68 % prediction interval for true DSC range with an interval of 0.10, using RFR model. L indicates the lower bound while U indicates the upper bound of the prediction interval. P, B, R, and PBR denote dataset for prostate, bladder, rectum, and all three organs, respectively.

True DSC Range	P	B	R	PBR
0.40–0.50	0.94 0.90 (L) 0.98 (U)	–	0.75 0.70 (L) 0.81 (U)	0.80 0.74 (L) 0.87 (U)
0.50–0.60	0.82 0.78 (L) 0.87 (U)	0.80 0.73 (L) 0.87 (U)	0.77 0.71 (L) 0.82 (U)	0.79 0.73 (L) 0.86 (U)
0.60–0.70	0.87 0.83 (L) 0.92 (U)	0.81 0.74 (L) 0.88 (U)	0.76 0.71 (L) 0.82 (U)	0.80 0.74 (L) 0.86 (U)
0.70–0.80	0.87 0.82 (L) 0.91 (U)	0.81 0.74 (L) 0.88 (U)	0.77 0.71 (L) 0.83 (U)	0.80 0.74 (L) 0.86 (U)
0.80–0.90	0.88 0.84 (L) 0.93 (U)	0.82 0.75 (L) 0.89 (U)	0.78 0.72 (L) 0.84 (U)	0.82 0.76 (L) 0.89 (U)
0.90–1.00	0.89 0.85 (L) 0.94 (U)	0.83 0.76 (L) 0.90 (U)	0.78 0.72 (L) 0.84 (U)	0.86 0.80 (L) 0.92 (U)

Several past studies have demonstrated the time taken for contour delineation in the order of minutes. In their study, Yedekci et al. [35] reported an average manual contouring time of 14.0 ± 0.4 min per fraction for cervical cancer patients. Besides, other work reported a mean contouring time of 37.4 ± 5.9 min [36] and a median contouring time of 20 min [37] per patient for breast cancer and lung cancer radiotherapy, respectively. Using our trained model, it is possible to eliminate the step of manual contouring and achieve a routine check on the propagated contours in a matter of seconds, given the DVF-based metrics resulting from DIR between new pairs of images.

At this point in the discussion, there is a caveat that we wish to highlight: the volume dependence of DSC, as has been reported in several studies [38,39]. In our work, the differential thresholds identified for each organ were all close to 0.85, the selected threshold using the complete dataset with all three organs. Hence, it can serve as a representative average threshold for the entire dataset. When dealing with a new dataset containing various organs, it is recommended to perform a separate analysis for each organ, to identify if there is a need for organ-specific thresholds. Based on the established threshold(s) and DSC distributions (examples shown in Fig. 3), contours could also be sampled randomly from each side of the threshold to justify whether a DSC lower than the threshold actually corresponds to a low contour overlap, and vice versa. These steps should be performed during the "commissioning" phase of the model, before clinical implementation could take place. A regular update and quality assurance (QA) of the model [40] will also be imperative to include more patients' data and ensure that the model performance is being monitored.

The main limitation of this work lies in the restricted range of the training data. As this study aims to simulate a real clinical workflow where the optimal DIR parameters would be ideal for use, the DSC values resulting from the DIR largely fall between 0.5 and 1, with more

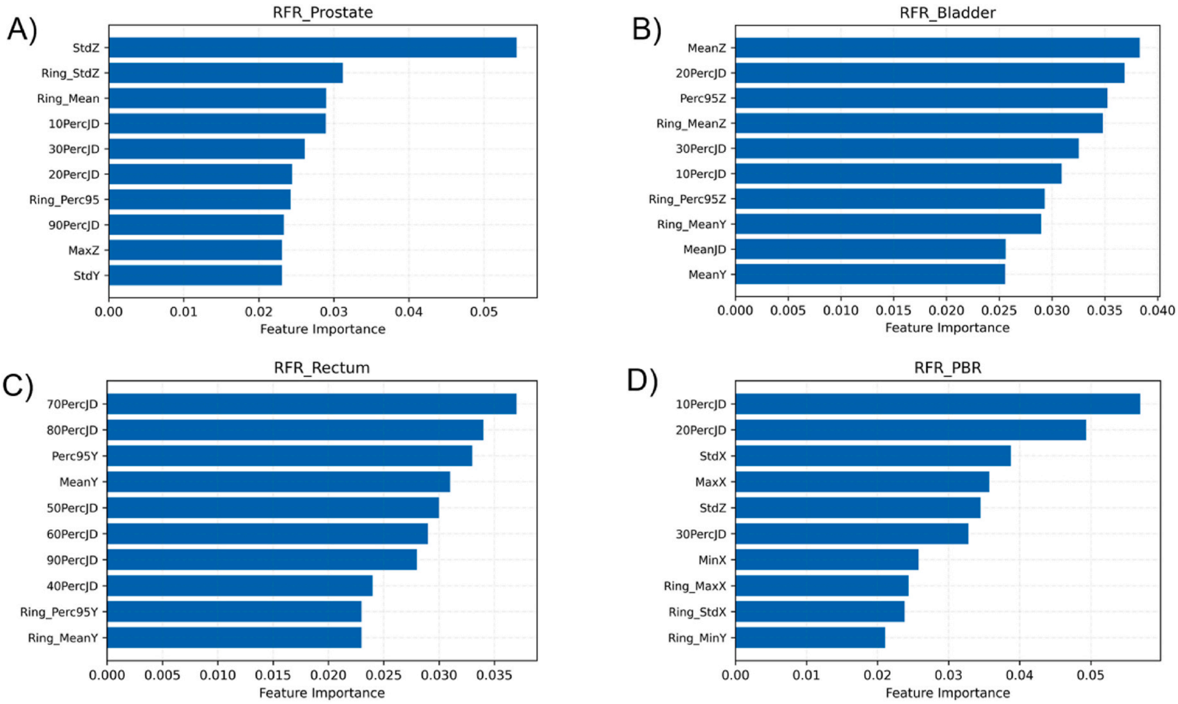


Fig. 5. Feature importance for the top 10 features when RFR model is used to predict dataset of (A) prostate, (B) bladder, (C) rectum, and (D) all three organs.

data points clustering around the higher-range values. This skewed distribution causes an imbalance in the prediction accuracy of all the models tested, as shown in Fig. 4 and Figs. S4–5. As discussed earlier in section3.2, a high percentage of errors smaller than 0.05 is found close to the peak of the distribution, while greater errors are observed nearer to the tails. At the initial phase of implementation, regular data collection along the way to include patients from more cohorts would prove useful in getting the model to be more “prepared” to out-of-distribution data, hence boosting its performance.

It is important to note that this proposed ML pipeline is limited to DIR validation for contour propagation. In other words, the reliability of DVF is confined to the edges of the organs considered, and it does not guarantee the registration accuracy within the organ contours. Hence, further QA procedures must be performed for the application of DIR in dose accumulation, which is another interesting aspect that we are looking to explore.

While artificial intelligence (AI)-based autocontouring software is increasingly being introduced, with some offering CBCT autocontouring function, these are not yet available in all centres worldwide, and the results vary based on the quality of CBCT images. It is undoubted that AI-based method will continue advancing, and with the enhancement of CBCT image quality, DIR-based contour propagation may become obsolete one day. That being said, we believe that this ML workflow could serve as a useful initial exercise in the DIR validation process, before a full transition towards AI-based autocontouring, even on CBCT images, could take place.

5. Conclusion

This study has adopted machine learning to predict contour-based metric using more than one DVF-based metric. The performance of the three models tested are promising, with RFR giving the best performance, signified by the overall lowest average MAE. As a precautionary step, the feature distributions can be used to flag any potential failure of the model prediction for clinical data that do not belong to the current dataset. It is foreseen that this work will provide a significant speed-up for the DIR validation process, which will be extremely advantageous for assimilation of ART into clinical practice.

Funding support

Hong Qi Tan is supported by the Duke-NUS Oncology Academic Program Goh Foundation Proton Research Programme (08/FY2021/EX (SL)/92-A146), Clinical & Systems Innovation Support – Innovation Seed Grant (08/FY2022/P2/02-A68)

CRediT authorship contribution statement

Yun Ming Wong: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Ping Lin Yeap:** Data curation, Formal analysis. **Ashley Li Kuan Ong:** Data curation. **Jeffrey Kit Loong Tuan:** Data curation. **Wen Siang Lew:** Supervision. **James Cheow Lei Lee:** Supervision. **Hong Qi Tan:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Writing – review & editing.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmed.2024.100163.

References

[1] Yan D, Vicini F, Wong J, Martinez A. Adaptive radiation therapy. 1997. p. 11.
[2] Glide-Hurst CK, et al. Adaptive radiation therapy (ART) Strategies and technical considerations: a state of the ART review from nrg Oncology. Int. J. Radiat. Oncol. Mar. 2021;109(4):1054–75. https://doi.org/10.1016/j.ijrobp.2020.10.021.
[3] Paganetti H, Botas P, Sharp GC, Winey B. Adaptive proton therapy. Phys Med Biol Nov. 2021;66(22):22TR01. https://doi.org/10.1088/1361-6560/ac344f.

[4] Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: report of the AAPM radiation therapy committee task group No. 132. Med Phys Jul. 2017;44(7):e43–76. https://doi.org/10.1002/mp.12256.

[5] Rigaud B, et al. Deformable image registration for radiation therapy: principle, methods, applications and evaluation. Acta Oncol Sep. 2019;58(9):1225–37. https://doi.org/10.1080/0284186X.2019.1620331.

[6] Paganelli C, Meschini G, Molinelli S, Riboldi M, Baroni G. Patient-specific validation of deformable image registration in radiation therapy: overview and caveats. Med Phys 2018:15.

[7] Loi G, et al. Performance of commercially available deformable image registration platforms for contour propagation using patient-based computational phantoms: a multi-institutional study. Med Phys Feb. 2018;45(2):748–57. https://doi.org/10.1002/mp.12737.

[8] Mee M, Stewart K, Lathouras M, Truong H, Hargrave C. Evaluation of a deformable image registration quality assurance tool for head and neck cancer patients. J. Med. Radiat. Sci. 2020;67(4):284–93. https://doi.org/10.1002/jmrs.428.

[9] Nie K, Pouliot J, Smith E, Chuang C. Performance variations among clinically available deformable image registration tools in adaptive radiotherapy — how should we evaluate and interpret the result? J Appl Clin Med Phys Mar. 2016;17(2):328–40. https://doi.org/10.1120/jacmp.v17i2.5778.

[10] Nobnop W, Chitapanarux I, Neamin H, Wanwilairat S, Lorvidhaya V, Sanghangthum T. Evaluation of deformable image registration (DIR) methods for dose accumulation in nasopharyngeal cancer patients during radiotherapy. Radiol Oncol Sep. 2017;51(4):438–46. https://doi.org/10.1515/raon-2017-0033.

[11] Saleh Z, et al. A multiple-image-based method to evaluate the performance of deformable image registration in the pelvis. Phys Med Biol Aug. 2016;61(16):6172–80. https://doi.org/10.1088/0031-9155/61/16/6172.

[12] Shi L, et al. Benchmarking of deformable image registration for multiple anatomic sites using digital data sets with ground-truth deformation vector fields. Pract. Radiat. Oncol. Sep. 2021;11(5):404–14. https://doi.org/10.1016/j.prro.2021.02.012.

[13] Varadhan R, Karangelis G, Krishnan K, Hui S. A framework for deformable image registration validation in radiotherapy clinical applications. J Appl Clin Med Phys Jan. 2013;14(1):192–213. https://doi.org/10.1120/jacmp.v14i1.4066.

[14] Christoph V, Ali K, Parmeshwar K, Rudiger W. A learning-based approach to evaluate registration success. In: MIAR 2010 med. Imaging augment. Real., vol. 6326; 2010. p. 429–37.

[15] Sokooti H, Saygili G, Glocker B, Lelieveldt BPF, Staring M. Quantitative error prediction of medical image registration using regression forests. Med Image Anal Aug. 2019;56:110–21. https://doi.org/10.1016/j.media.2019.05.005.

[16] Dushepa V. A machine learning approach for image registration accuracy estimation. In: 2020 IEEE Ukrainian microwave week (UkrMW). Kharkiv, Ukraine: IEEE; Sep. 2020. p. 368–72. https://doi.org/10.1109/UkrMW49653.2020.9252810.

[17] Hakenberg OW, Linne C, Manseck A, Wirth MP. Bladder wall thickness in normal adults and men with mild lower urinary tract symptoms and benign prostatic enlargement. Neurourol Urodyn 2000;19(5):585–93. https://doi.org/10.1002/1520-6777(2000)19:5<585::aid-nau5>3.0.co;2-u.

[18] Fisher JK. Normal colon wall thickness on CT. Radiology Nov. 1982;145(2):415–8. https://doi.org/10.1148/radiology.145.2.7134445.

[19] Pedregosa F, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825–30.

[20] sklearn.linear_model.LinearRegression." Accessed: January. 1, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[21] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. Apr. 2011;2(3):1–27. https://doi.org/10.1145/1961189.1961199.

[22] Breiman L. Random forests. Mach Learn 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.

[23] Qin A, Ionascu D, Liang J, Han X, O'Connell N, Yan D. The evaluation of a hybrid biomechanical deformable registration method on a multistage physical phantom with reproducible deformation. Radiat Oncol Dec. 2018;13(1):240. https://doi.org/10.1186/s13014-018-1192-x.

[24] Jurkovic I-A, Papanikolaou N, Stathakis S, Kirby N, Mavroidis P. Objective assessment of the quality and accuracy of deformable image registration. J Med Phys 2020;45(3):156–67. https://doi.org/10.4103/jmp.JMP_47_19.

[25] Eiben B, Bertholet J, Menten MJ, Nill S, Oelfke U, McClelland JR. Consistent and invertible deformation vector fields for a breathing anthropomorphic phantom: a post-processing framework for the XCAT phantom. Phys Med Biol Aug. 2020;65(16):165005. https://doi.org/10.1088/1361-6560/ab8533.

[26] Kuang D. On reducing negative jacobian determinant of the deformation predicted by deep registration networks. 2019. https://doi.org/10.48550/arXiv.1907.00068. arXiv, Jun. 28.

[27] Pal S, Tennant M, Ray N. Towards positive jacobian: learn to postprocess diffeomorphic image registration with matrix exponential. arXiv Feb. 01, 2022. https://doi.org/10.48550/arXiv.2202.00749.

[28] Stoll M. Combining motion statistics with patient-specific biomechanical modelling to predict probable interfractional deformations. In: Programme book of the 18th int. Conf. On the use of computers in radiation therapy; 2016.

[29] Britton KR, Takai Y, Mitsuya M, Nemoto K, Ogawa Y, Yamada S. Evaluation of inter- and intrafraction organ motion during intensity modulated radiation therapy (IMRT) for localized prostate cancer measured by a newly developed on-board image-guided system. Radiat Med Feb. 2005;23(1):14–24.

[30] Litzenberg DW, et al. Influence of intrafraction motion on margins for prostate radiotherapy. Int J Radiat Oncol Biol Phys Jun. 2006;65(2):548–53. https://doi.org/10.1016/j.ijrobp.2005.12.033.

[31] Kotte ANTJ, Hofman P, Lagendijk JJW, van Vulpen M, van der Heide UA. Intrafraction motion of the prostate during external-beam radiation therapy: analysis of 427 patients with implanted fiducial markers. Int J Radiat Oncol Biol Phys Oct. 2007;69(2):419–25. https://doi.org/10.1016/j.ijrobp.2007.03.029.

[32] Nederveen AJ, van der Heide UA, Dehnad H, van Moorselaar RJA, Hofman P, Lagendijk JJW. Measurements and clinical consequences of prostate motion during a radiotherapy fraction. Int J Radiat Oncol Biol Phys May 2002;53(1):206–14. https://doi.org/10.1016/s0360-3016(01)02823-1.

[33] Dawson LA, Mah K, Franssen E, Morton G. Target position variability throughout prostate radiotherapy. Int J Radiat Oncol Biol Phys Dec. 1998;42(5):1155–61. https://doi.org/10.1016/s0360-3016(98)00265-x.

[34] Langen KM, Jones DT. Organ motion and its management. Int J Radiat Oncol Biol Phys May 2001;50(1):265–78. https://doi.org/10.1016/s0360-3016(01)01453-5.

[35] Yedekci Y, Gültekin M, Sarı SY, Yıldız F. Automatic contouring using deformable image registration for tandem-ring or tandem-ovoid brachytherapy. J Contemp Brachytherapy Feb. 2022;14(1):72–9. https://doi.org/10.5114/jcb.2022.112814.

[36] Byun HK, et al. Evaluation of deep learning-based autosegmentation in breast cancer radiotherapy. Radiat Oncol Oct. 2021;16(1):203. https://doi.org/10.1186/s13014-021-01923-1.

[37] Lustberg T, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother Oncol Feb. 2018;126(2):312–7. https://doi.org/10.1016/j.radonc.2017.11.012.

[38] Deeley MA, et al. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. Phys Med Biol Jul. 2011;56(14):4557–77. https://doi.org/10.1088/0031-9155/56/14/021.

[39] Kumarasiri A, et al. Deformable image registration based automatic CT-to-CT contour propagation for head and neck adaptive radiotherapy in the routine clinical setting. Med Phys Dec. 2014;41(12):121712. https://doi.org/10.1118/1.4901409.

[40] Hurkmans C, et al. A joint ESTRO and AAPM guideline for development, clinical validation and reporting of artificial intelligence models in radiation therapy. Radiother Oncol Aug. 2024;197:110345. https://doi.org/10.1016/j.radonc.2024.110345.