## Title

Development and interinstitutional validation of an automatic vertebral-body misalignment error detector for cone-beam CT-guided radiotherapy



## Journal

Medical Physics, 49(10)

## ISSN

0094-2405

## Authors

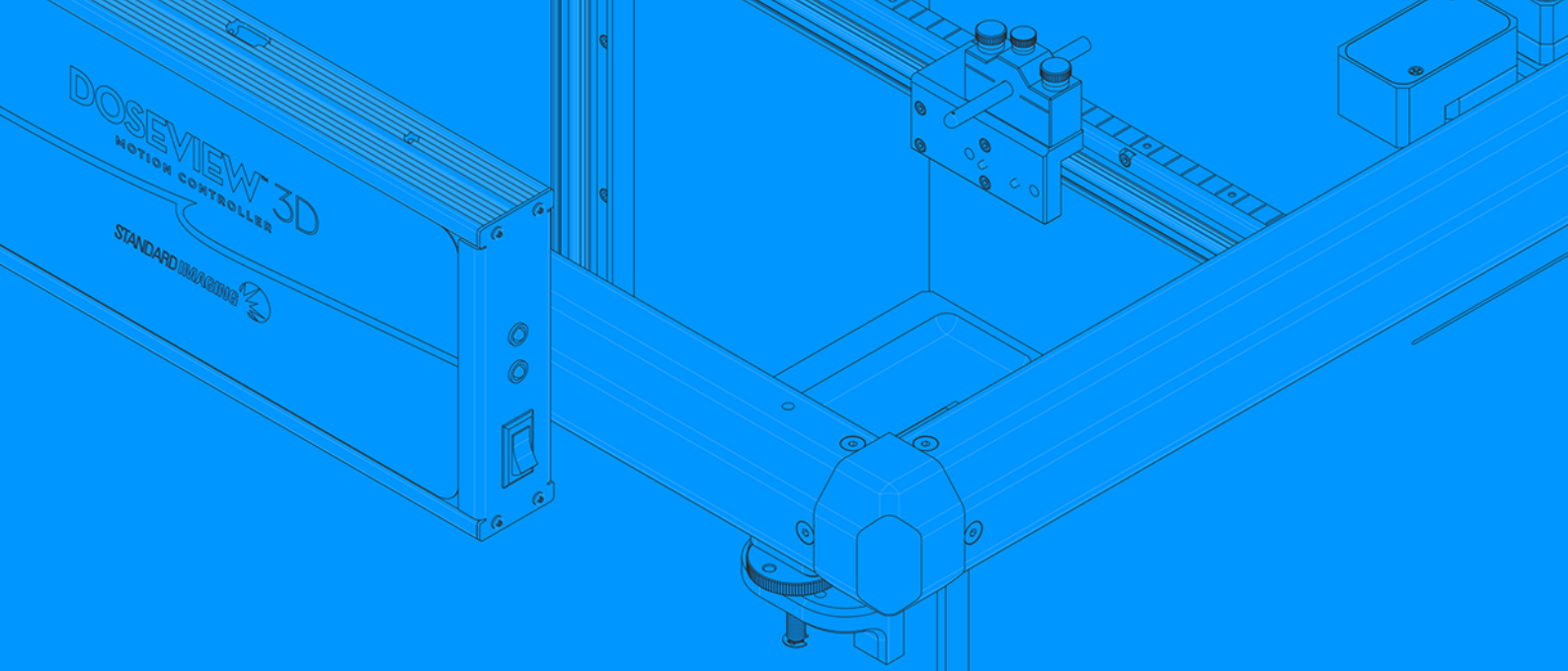
Luximon, Dishane C
Ritter, Timothy
Fields, Emma
et al.

## Publication Date

2022-10-01

## DOI

10.1002/mp.15927




# IT'S YOUR TIME
# BE PRECISE

STANDARD IMAGING

Every day we spend **our time** optimizing ways to make QA easy and reliable.

**Ask us how our solutions can benefit you.**

WWW.STANDARDIMAGING.COM

Received: 12 April 2022 | Revised: 20 July 2022 | Accepted: 5 August 2022

DOI: 10.1002/mp.15927

RESEARCH ARTICLE

MEDICAL PHYSICS

# Development and interinstitutional validation of an automatic vertebral-body misalignment error detector for cone-beam CT-guided radiotherapy

**Dishane C. Luximon**[1] | **Timothy Ritter**[2] | **Emma Fields**[2] | **John Neylon**[1] | **Rachel Petragallo**[1] | **Yasin Abdulkadir**[1] | **John Charters**[1] | **Daniel A. Low**[1] | **James M. Lamb**[1]

[1]Department of Radiation Oncology, David Geffen School of Medicine, University of California, Los Angeles, California, USA

[2]Department of Medical Physics, Virginia Commonwealth University, Richmond, Virginia, USA

**Correspondence**
Dishane C. Luximon, Department of Radiation Oncology, David Geffen School of Medicine, University of California, 200 UCLA Medical Plaza Suite B265, Box 956951, Los Angeles, CA 90095-6951, USA.
Email: dluximon@mednet.ucla.edu

**Funding information**
Agency for Healthcare Research and Quality (AHRQ), Grant/Award Number: 1R01HS026486

## Abstract

**Background:** In cone-beam computed tomography (CBCT)-guided radiotherapy, off-by-one vertebral-body misalignments are rare but serious errors that lead to wrong-site treatments.

**Purpose:** An automatic error detection algorithm was developed that uses a three-branch convolutional neural network error detection model (EDM) to detect off-by-one vertebral-body misalignments using planning computed tomography (CT) images and setup CBCT images.

**Methods:** Algorithm training and test data consisted of planning CTs and CBCTs from 480 patients undergoing radiotherapy treatment in the thoracic and abdominal regions at two radiotherapy clinics. The clinically applied registration was used to derive true-negative (no error) data. The setup and planning images were then misaligned by one vertebral-body in both the superior and inferior directions, simulating the most likely misalignment scenarios. For each of the aligned and misaligned 3D image pairs, 2D slice pairs were automatically extracted in each anatomical plane about a point within the vertebral column. The three slice pairs obtained were then inputted to the EDM that returned a probability of vertebral misalignment. One model ($EDM_1$) was trained solely on data from institution 1. $EDM_1$ was further trained using a lower learning rate on a dataset from institution 2 to produce a fine-tuned model, $EDM_2$. Another model, $EDM_3$, was trained from scratch using a training dataset composed of data from both institutions. These three models were validated on a randomly selected and unseen dataset composed of images from both institutions, for a total of 303 image pairs. The model performances were quantified using a receiver operating characteristic analysis. Due to the rarity of vertebral-body misalignments in the clinic, a minimum threshold value yielding a specificity of at least 99% was selected. Using this threshold, the sensitivity was calculated for each model, on each institution's test set separately.

**Results:** When applied to the combined test set, $EDM_1$, $EDM_2$, and $EDM_3$ resulted in an area under curve of 99.5%, 99.4%, and 99.5%, respectively. $EDM_1$ achieved a sensitivity of 96% and 88% on Institution 1 and Institution 2 test set, respectively. $EDM_2$ obtained a sensitivity of 95% on each institution's test set. $EDM_3$ achieved a sensitivity of 95% and 88% on Institution 1 and Institution 2 test set, respectively.

**Conclusion:** The proposed algorithm demonstrated accuracy in identifying off-by-one vertebral-body misalignments in CBCT-guided radiotherapy that

was sufficiently high to allow for practical implementation. It was found that fine-tuning the model on a multi-facility dataset can further enhance the generalizability of the algorithm.

**KEYWORDS**
deep learning, patient safety, radiation therapy

## 1 | INTRODUCTION

The technologies behind external beam radiation therapy (EBRT) are continuously evolving to enhance treatment planning and beam delivery. The use of image-guided radiotherapy (IGRT), for example, has allowed for more precise and highly conformal beam delivery and treatment planning.[1] Although these technologies promise to reduce setup uncertainties, they also bring more complexities to the EBRT processes, which may increase the risk of incidents in the absence of safeguards.[2,3] Lack of experience, inadequate procedures, inattention, and miscommunications between therapists may result in setup and treatment errors.[4–7]

In a study covering 336 treatment facilities in the United States, 396 critical events were identified between 2014 and 2016, where 6.3% of those were due to wrong manual shifts or wrong IGRT-generated shifts.[8] This report also highlighted a T12-L5 spine case where the automatic registration of the cone-beam computed tomography (CBCT) was incorrect by 3 cm in the superior–inferior direction for the first two fractions of a five-fraction treatment. The error was only captured on the third fraction when the therapists realized that something was wrong and called the physicist for a review. Although the outcome of this treatment is unknown, this incident demonstrates the risks involved when relying solely on human perception to catch errors. In the thoracic region particularly, there is a higher risk of these types of errors occurring due to the similarity between adjacent the vertebral bodies, which are often used as landmark during the registration process in IGRT. This region is also prone to motion artifacts, which can complicate the registration process. Shah et al. have shown, for example, that the anatomical variations and anomalies in the thoracic vertebra and surrounding regions can cause improper labeling of vertebral bodies and contribute to wrong-level spine surgery.[9]

Hence, with these new evolving technologies comes the need for error-mitigating systems that can reduce the risk of setup errors and make EBRT safer for the patient. Although some have been trying to solve this problem with real-time monitoring systems using camera tracking,[10–12] others have proposed the use of automated processes to detect setup errors by analyzing IGRT images acquired before beam delivery.[13] As no additional equipment is required in the latter solution beyond what is used for IGRT, it is more cost-effective and potentially accessible to a larger number of facilities.

Jani et al. have developed an automated system for the detection of patient identification and setup errors in EBRT using setup kilovoltage computed tomography (CT) images and planning CT images.[14] Their work made use of image similarity metrics as features, which were applied to a linear discriminant analysis for the error classification. Although this classical machine learning method produced acceptable results in classifying wrong-vertebral-body errors, it was limited by the feature selection, which relied on human observation for pattern recognition. Deep learning (DL), on the other hand, can automatically determine and extract high-level features from raw data, which allows it to obtain patterns undiscernible by human observation.[15] Convolutional neural networks[16] have previously been used for image classification problems and this has huge potential in the field of medical imaging.[17–19] Several DL methods have been proposed for disease or tissue characterization, diagnosis, and prognosis.[20–22] However, to this day, the use of DL has yet to be applied to CBCT-guided radiotherapy setup error detection.

In this study, we propose a DL–based algorithm that can detect off-by-one vertebral misalignment errors in CBCT-guided radiotherapy by using the planning CT and the CBCTs, focusing on the thoracic and abdominal regions where vertebrae are often used as registration landmarks.[23] Due to the similarity of vertebral bodies in these regions, increased organ motion, and the lower image quality of the CBCT compared to the planning CT, one potential and clinically impactful mistake that could occur is the misregistration of the CBCT with respect to the planning CT by one vertebral body. This particular mistake may go unnoticed and lead to significant harm to the patient by missing the targeted tumor and causing excess damage to healthy tissues.

The long-term goal of this project is to develop a fully automated error detection system that can act as a real-time secondary barrier to prevent off-by-one vertebral-body misalignments from occurring in the clinic. Additionally, by analyzing all the treatment scans performed within a user-defined time and flag possible anomalies, this tool could potentially aid and supplement regular chart checks performed by medical physicists for quality assurance (QA) of CBCT-guided EBRT. However, for successful clinical implementation, it is essential to have a tool that minimally disrupts the clinical workflow due to false positives. Hence, in the development of our tool, a large focus was placed on the model's ability to catch off-by-one vertebral-body misalignments with

a threshold value that leads to less than 1% of false positives, which can be deemed acceptable in comparison to other false-positive interrupts and interlocks in the clinical workflow.

Interinstitutional validation is also key in assessing a DL model's generalizability power on a variety of patients, registration practices, image quality, and scanning protocols. In this study, which included patient data from two different institutions, the performance of the tool on cross-institutional data was investigated. Such experiment could be helpful in determining the ability to apply the tool to other facilities, or otherwise, the need for further data to enhance the generalizability of the model for effective error-catching power and minimal false positives at other facilities.

## 2 | METHOD

### 2.1 | Dataset

Under an IRB-approved protocol, planning CTs and CBCTs were collected from 380 patients undergoing radiotherapy treatment in the thoracic or abdominal region at the University of California, Los Angeles Medical Center (Institution 1). The treatments at Institution 1 had been performed on three TrueBeam and one Novalis Tx linear accelerator treatment machines (Varian Medical Systems, CA, United States). From those 380 patients, 1316 clinically aligned planning CT–CBCT pairs were obtained and used in our work. Additionally, 100 patient datasets were collected from the Virginia Commonwealth University Medical Center (Institution 2). The patients at Institution 2 had been treated on Varian Trilogy and TrueBeam linear accelerator treatment machines. The acquisition protocol used to acquire the CBCTs at each institution is described in Table A1. For each CBCT acquired from the two facilities, a registration (REG) file in the DICOM format was extracted to obtain the clinically applied alignment. Additionally, the RT structure file for each planning CT was collected.

The patient data from Institution 1 was collected using an in-house DICOM query and retrieval (DQR) application programming interface using the pynetdicom[1] Python package. Our custom DQR software allowed automatic retrieval of patient data from the ARIA image management system (Varian Medical Systems) based on user-defined date ranges, plan names, and image types. This tool was built to fully automate our data acquisition protocol, thereby allowing the possibility of a fully automated error detection pipeline.

The planning CT–CBCT pair obtained from each treatment fraction was used as true-negative (aligned) cases. Due to the scarcity of off-by-one vertebral-body misalignment cases in the clinic, the true-positive (misaligned) cases were manually generated. In the misalignment generation process, the planning CT–CBCT pair from the earliest treatment fraction of each patient, together with its corresponding clinically applied REG file, were selected and imported into MIM (MIM Software Inc, OH, United States). For the 480 selected pairs, off-by-one vertebral-body misalignments were simulated on MIM by manually shifting the CBCT by one vertebral body in the cephalic–caudal direction with respect to the planning CT. Two misalignments were produced for each individual patient; one in the superior direction, and the other in the inferior direction. For these misaligned cases, the CBCT and CT were carefully matched as much as possible to produce errors that had the potential of being overlooked in the clinical setting, as shown in Figure 1. The new misaligned REG files were then exported in the DICOM format to obtain the true-positive (i.e., error-simulating) registrations.

The datasets from each institution were separately and randomly split into their respective training, validation, and test sets. As scans from multiple fractions were used as true-negative (aligned) cases in our study, the dataset split was performed based on the patients' unique anonymized identifiers to avoid having scans from the same patient on both the training and test sets. The number of scans used in the training, validation, and testing phase is described in Table 1.

### 2.2 | Image preprocessing

The REG files, both aligned and misaligned, were consequently used to match the coordinates of the CBCT volume with those of the planning CT volume. To ensure uniformity over the whole dataset, all volume pairs were resampled using a $1 \times 1 \times 1.5\text{-mm}^3$ grid.

The couch position from the planning CT is very rarely aligned to the couch from the CBCT due to differences in material and structure. Hence, the positional and structure differences in the images are of trivial importance in our error detection system and can even be misleading in the detection of wrongly aligned patients. In order to remove the couch from the images, the body contours found in the structure files were used to clean up both the CT and CBCT volumes such that the couch and other irrelevant regions outside of the body were assigned voxel values equivalent to the Hounsfield unit (HU) of air (−1000 HU).

The eventual goal of this project is to develop a tool that could run in real time simultaneously with treatment delivery processes, and hence, run time and memory footprint were primary considerations. Therefore, instead of using the entire 3D image volumes as inputs to the DL model, orthogonal 2D slices were used, as shown in Figure 2. By extracting one slice in each orthogonal plane, the memory requirement of our model was considerably minimized, whereas the important features of the patients' anatomy were kept and used by our model to analyze the patient alignment. Selection of an

[1] https://pydicom.github.io/pynetdicom/stable/#

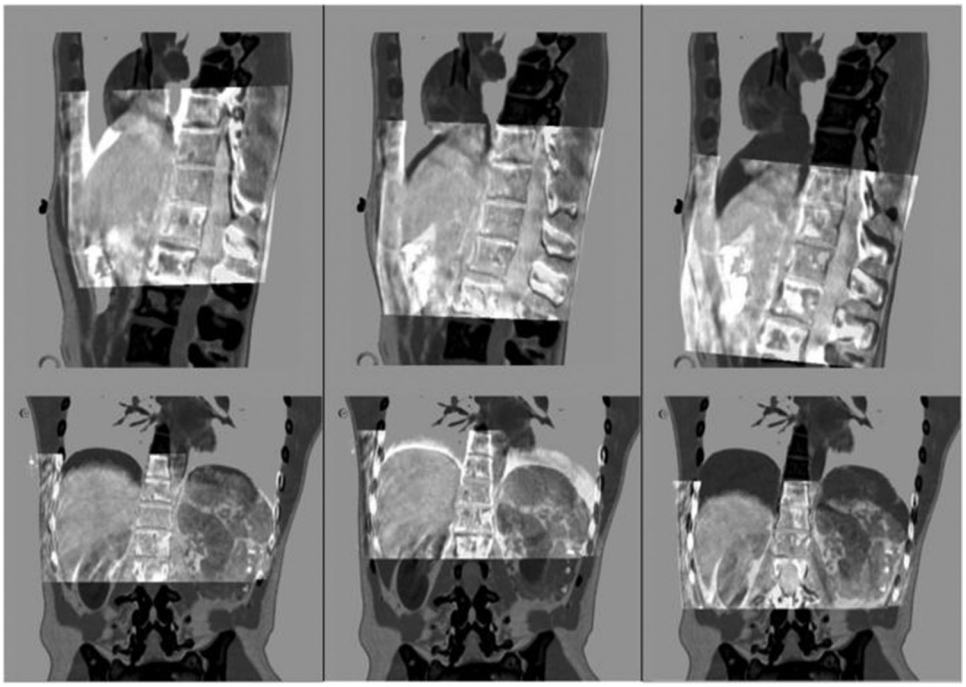



**FIGURE 1** Image fusions to demonstrate the manually generated off-by-one vertebral-body misalignments. In column (a), the cone-beam computed tomography (CBCT) was upshifted by one vertebral body with respect to the planning computed tomography (CT). Column (b) shows the correct clinical alignment, and column (c) shows a misalignment where the CBCT was downshifted by one vertebral body with respect to the planning CT.

**TABLE 1** Description of the dataset partitioned into the training, validation, and testing sets for each institution

| | | Number of patients | CBCT image pairs | | |
|---|---|---|---|---|---|
| | | | Aligned | Misaligned[a] | Total |
| **Institution 1** | Training | 304 | 1069 | 608 | 1677 |
| | Validation | 29 | 98 | 58 | 156 |
| | Testing | 47 | 149 | 94 | 243 |
| **Institution 2** | Training | 70 | 70 | 140 | 210 |
| | Validation | 10 | 10 | 20 | 30 |
| | Testing | 20 | 20 | 40 | 60 |
| **Total** | | 480 | 1416 | 960 | 2376 |

*Note*: The total number of patients and scans used in our work is also shown.
[a]Two misaligned image pairs were manually generated for each patient in the dataset.
Abbreviation: CBCT, cone-beam computed tomography.

appropriate origin for the coordinate axes was important to assure that the relevant image features were present. In the clinic, for thoracic and abdominal cases, the spine is often used as a marker during the registration and patient alignment step. An automated process was therefore used to select axial, sagittal, and coronal planes that intersected at the approximate center of the vertebral bodies at a location midway through the image in the cranio-caudal direction. This particular point was chosen, rather than the treatment plan isocenter, as it offers more details about the vertebral location in the detection of off-by-one vertebral-body misalignments.

A binary mask of the patient body was first extracted from the CBCT using a thresholding method. A morphological dilation followed by erosion was applied on the binary mask to fill any gaps after the thresholding operation. The dilation and erosion operations used $20 \times 20$ and $5 \times 5$-pixel$^2$ rectangular structuring element, respectively. The axial slice index was then extracted by locating the middle slice of the mask containing the patient body on the CBCT, denoted as $X_{Ax}$.

Using the axial slice index obtained in the previous step, the corresponding axial slice images were extracted from both the CT scan and the CBCT scan.

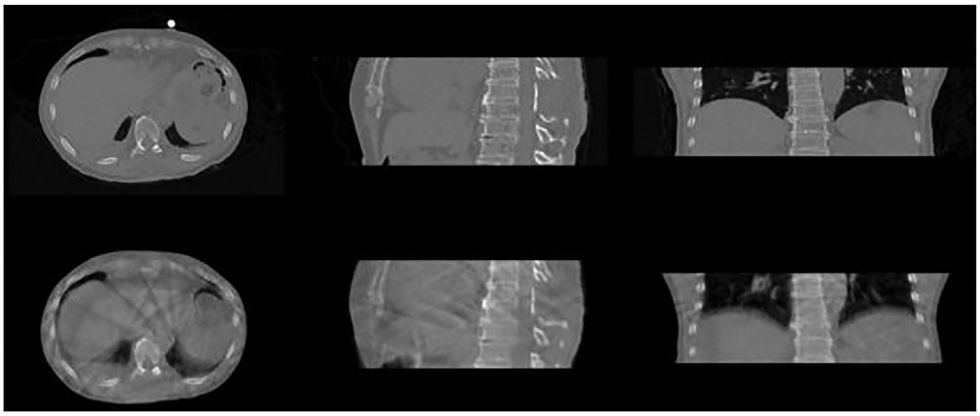


**FIGURE 2** Orthogonal 2D slices extracted from the planning computed tomography (CT) (top row) and its corresponding cone-beam computed tomography (CBCT) (bottom row)

The vertebral-body location in the coronal and sagittal planes, denoted as ($X_{Cor}$ and $X_{Sag}$), was then obtained by applying a constant 10-pixel translation in the anterior direction from the central point of the spinal canal. The spinal canal location was derived from the spinal canal structure in the treatment plan if existing, or from a dedicated UNet-based[23] spinal canal segmentation algorithm (see Appendix) if the plan did not contain a spinal canal contour. This vertebral-body coordinates ($X_{Cor}, X_{Sag}, X_{Ax}$) were then used as the coordinate origin for the coronal and sagittal 2D slices extracted from the planning CT scan and the CBCT scan.

Following the orthogonal slice extraction, the 2D images in the coronal and sagittal planes were cropped to reduce the empty regions around the patient body and hence minimize the number of unnecessary computations in our error detection model (EDM). The coronal and sagittal slices were cropped to 400 × 110 and 280 × 110 images about the center of the CBCT image, which was found using the binary mask of the patient body. The axial slice images were down-sampled using a linear interpolation method to obtain 256 × 256 arrays. By repeating our experiment on the original 512 × 512 axial images, we found that the downsizing step performed did not have any adverse effect on the accuracy of our EDM.

After the orthogonal slice extraction, the 2D arrays from the planning CT and CBCT were then concatenated with respect to their plane to obtain one 256 × 256 × 2 axial array, one 400 × 110 × 2 coronal array, and the other 280 × 110 × 2 sagittal array. For each of the orthogonal arrays, the first channel is the planning CT image, and the second channel is the respective CBCT image. These three arrays were then used as inputs to our EDM.

## 2.3 | Error detection model (EDM)

The EDM was based on the Dense-Net architecture[24] and was composed of three branches that processed the three orthogonal images separately before merging into a final densely connected layer, as shown in Figure 3. This three-branch EDM made use of densely contracting paths to capture contextual information from the three inputs before outputting a misalignment probability.

In the clinical setting, the manual image registration process is often performed using all three orthogonal views. However, coronal and sagittal planes may be more sensitive to cranio-caudal misregistrations such as one vertebral-body displacements. Therefore, more convolutional filters were placed in the coronal and sagittal branches such that the model extracts a higher number of features from these two planes, as compared to the axial branch. Hence, this results in the EDM placing higher weights on the coronal and sagittal plane during the off-by-one vertebral-body misalignment detection.

## 2.4 | Training and testing configuration

One EDM was trained on the training set from Institution 1 only ($EDM_1$). During training, $EDM_1$ was validated after each epoch using a validation set from Institution 1 only. $EDM_1$ was further trained by updating all the weights in the model using a lower learning rate on the training dataset from Institution 2 to produce $EDM_2$. Some studies have shown that this method of unfreezing and fine-tuning all layers can outperform the traditional transfer learning method where most of the network's layers are kept frozen and the final layers are updated.[25] This fine-tuning method has also been proven to be an effective method of training on imbalanced data that is present in our dataset due to the higher number of patients obtained from Institution 1 as compared to Institution 2, as shown in Table 1.[26] After the fine-tuning step, $EDM_2$ was validated during training using a validation set containing data from both institutions.

Finally, a third model, $EDM_3$, was trained from scratch using the training data from both institutions. Similar

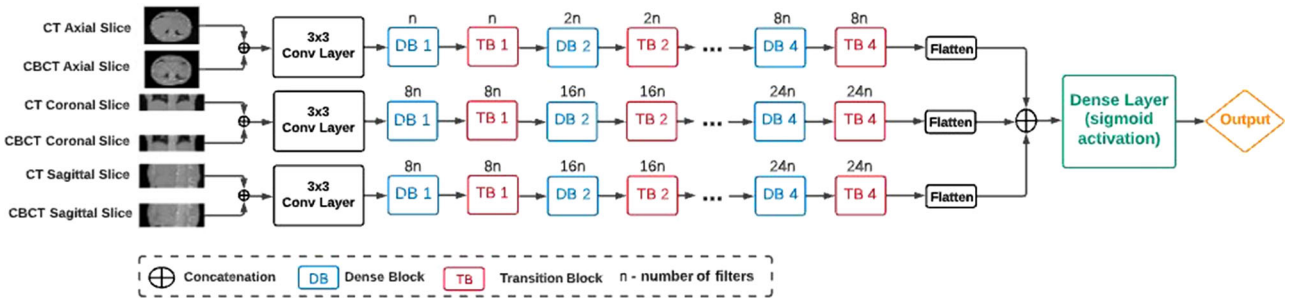



**FIGURE 3** Depiction of the network architecture used in the proposed work ($n = 4$). The dense block consists of two densely connected layers connected in a feed-forward mode (each composed of two convolutional layers, two batch normalization layers, two activation layers, and one dropout layer) and the transition block of three layers (batch normalization layer, convolutional layer, and max pooling layer).

validation was done during training as $EDM_2$. The three models were then tested on a randomly selected and unseen patient dataset composed of images from both institutions, for a total of 303 image pairs. These experiments and comparisons were performed to assess whether $EDM_1$ could be used by another facility or whether it was necessary to fine-tune the model or retrain the model from scratch using data from the other facility before implementation.

In our experiments, we refrained from using pretrained classification networks that are trained on natural images, such as ResNet50,[27] and fine-tuned the model using our dataset. As natural image classification tasks are essentially very different from medical image classification tasks in terms of image characteristics, dataset sizes, and number of classes, transfer learning using powerful pretrained network has shown to offer little benefit in the medical imaging domain as compared to training the network from scratch using medical images.[28]

The proposed EDMs were implemented using TensorFlow 2.2 with Keras backend. $EDM_1$ and $EDM_3$ were trained using an Adam Optimizer[29] with a starting learning rate of $5 \times 10^{-5}$. During training, the models were evaluated after each epoch using their respective validation set, and the learning rate was reduced by a factor of 0.75 if the validation loss did not improve for 15 consecutive epochs. Both models were trained until the validation AUC did not improve for 50 consecutive epochs, or for a maximum of 200 epochs. The model achieving the highest validation accuracy was then saved. $EDM_1$ achieved convergence after 84 epochs and $EDM_3$ converged after 82 epochs. $EDM_1$ was fine-tuned using an Adam Optimizer with a starting learning rate of $2 \times 10^{-5}$ to produce $EDM_2$. Again, the model was evaluated during training on its validation set, and the learning rate was reduced by a factor of 0.75 if the validation loss did not improve for 10 consecutive epochs. This model was trained until the validation AUC did not improve for 20 consecutive epochs, or for a maximum of 100 epochs, and the model achieving the highest validation accuracy was saved. $EDM_2$ achieved convergence after 49 epochs.

## 2.5 | Loss function and evaluation metrics

During the model training, the binary cross-entropy (BCE) loss was used as the loss function, as shown in Equation (1). BCE has been shown to be an effective loss function for binary classification problems[30]:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} -\left(y_i \times \log\left(p_i\right) + \left(1 - y_i\right) \times \log\left(1 - p_i\right)\right) \quad (1)$$

where $y$ is the ground-truth label, and $p$ is the predicted probability of misalignment.

The receiver operating characteristic (ROC) curve was used to assess the performance of our models in classifying the registrations from our validation dataset.[31] The areas under each ROC curve (AUC) were used to quantify the performance of our models.

Although the principal target of our proposed algorithm is to catch misalignment errors, due to the rarity of the event in the clinic, it is crucial to minimize unnecessary disruption in the clinical workflow due to false positives. Based on our analysis on the patient load at Institution 1, which approximates to 300–350 treatments per week, it was deduced that a specificity of ≥99% would be equivalent to about one false positive per treatment machine per week, which was deemed acceptable in comparison to other false-positive interrupts and interlocks in the clinical workflow. Although this false-positive rate may vary from institution to institution based on the patient load, we believe that the chosen specificity is reasonable enough to limit clinical disruptions at most facilities. Hence, in our evaluation, a base threshold value yielding a specificity of at least 99% was chosen. From the binary results obtained, the true positive ($tp$), false positive ($fp$), false negative ($fn$), and true negative ($tn$) counts were obtained. These were

then used to calculate the sensitivity, *F*-1 score, and Matthews correlation coefficient (MCC). Student *t*-test was used to assess statistical significance of differences in the results from the three models on the whole test dataset, with a $p$-value $< 0.05$ being considered statistically significant.

The *F*-1 score combines both the precision and recall of a binary classifier and is shown in the following equation[32]:

$$F-1 = \frac{2tp}{2tp + fp + fn} \quad (2)$$

MCC, shown in Equation (3), is another metric used to quantify the performance of a binary classifier and has been shown to be a balanced measure in the case of class imbalances.[33,34] MCC can take a value between −1 and +1, where +1 means perfect positive correlation between prediction and ground truth, and −1 means perfect negative correlation:

$$MCC = \frac{(tp \times tn) - (fn \times fp)}{\sqrt{(tp + fn) \times (tn + fp) \times (tp + fp) \times (tn + fn)}} \quad (3)$$

Additionally, the mean model prediction probability was calculated for varying caudal-cranial distances between the planning CT and CBCT. For each image pair in the test set, the CBCT was automatically misaligned in the caudal-cranial direction by ±10, ±20, and ±40 mm with respect to the planning CT. These image pairs were then inputted to the best performing model to obtain the misalignment prediction probabilities. Provided that the human thoracic vertebral body is on average 20 mm in length,[35] this test can add value to the clinical utility of our algorithm by validating its potential at catching misalignment errors that are off by less than one vertebral body, and also misalignment errors that are greater than one vertebral body in magnitude.

## 3 | RESULTS AND ANALYSIS

$EDM_1$, $EDM_2$, and $EDM_3$ were tested on the 243 image pairs from Institution 1 test set and 60 image pairs from Institution 2 test set. Figure 4 represents the ROC analysis performed to assess the classification ability of $EDM_1$, $EDM_2$, and $EDM_3$ on the test sets. For each model and analysis, a threshold yielding a specificity of at least 99% was chosen. The sensitivity, *F*-1 score, and MCC were then calculated and are described in Table 2.

$EDM_2$ was found to be the superior model with the highest sensitivity, *F*-1 score, and MCC on the combined test set as compared to $EDM_1$ and $EDM_3$, with their score differences being statistically significant. $EDM_2$ was then used to plot the mean model prediction probability as a function of caudal-cranial distances between the planning CT and CBCT, as shown in Figure 5.

## 4 | DISCUSSION

In this work, a deep-learning-based vertebral-body misalignment error detection algorithm for cone-beam CT-guided radiotherapy was presented. Automated extraction of 2D slices from the planning CT and corresponding CBCT in each anatomic plane about a point within the vertebral column was performed as a preprocessing step. The three slice pairs were then input to our EDM that was composed of three branches. Each branch was used to extract features from one of the orthogonal images planes and was joined in a final densely connected layer, before returning a misalignment probability. Using an Nvidia Quadro P1000 4-GB graphics processing unit (GPU) (Nvidia Corporation, Santa Clara, CA, USA) system with a 16-GB RAM, our algorithm takes an average of 6.8 s to preprocess the input images, run through the EDM, and output a probability of misalignment. If the system were implemented as a third-party system independent of the clinical record and verify (R&V) system, the images would have to be retrieved from the treatment machine or R&V system, thereby increasing the runtime. In our implementation in a Varian environment, the retrieval of CBCT and alignment (REG file) from the ARIA servers using our DQR software required an additional 58 s, on average. Ideally, the proposed algorithm would be incorporated into the R&V system, obviating this data transfer contribution to the runtime. Although run-time optimization could further decrease the execution time of the algorithm, we believe that it can be clinically implemented as-is.

The EDM model was trained and tested using data from two institutions. The EDM trained on the single-institution data only, $EDM_1$, showed great ability in identifying vertebral-body misalignments from the same institution's test set, while limiting the number of false positives, which is key for successful clinical implementation. $EDM_1$ was also fine-tuned by training the model on a combination of two institution's data, resulting in another model called $EDM_2$. A third model, $EDM_3$, was trained from scratch using data from both institutions. Our results demonstrated that $EDM_2$ and $EDM_3$ performed better on the second institution's test dataset, with $EDM_2$ (fine-tuned model) being the superior model out of the three when it came to the combined test set. Although $EDM_1$ produced moderately accurate results on an external facility's data, the results from this experiment showed that incorporating using interinstitutional data into the training data could further enhance the classification capabilities and sensitivity of the model when applied to the respective facility's scans.

As compared to a similar work that uses non–DL techniques[14] to find vertebral misalignment errors in thoracic CBCT-guided radiotherapy treatments, $EDM_2$ resulted in higher sensitivity (0.95 vs. 0.90) for a fixed specificity of 99%. Additionally, our model was

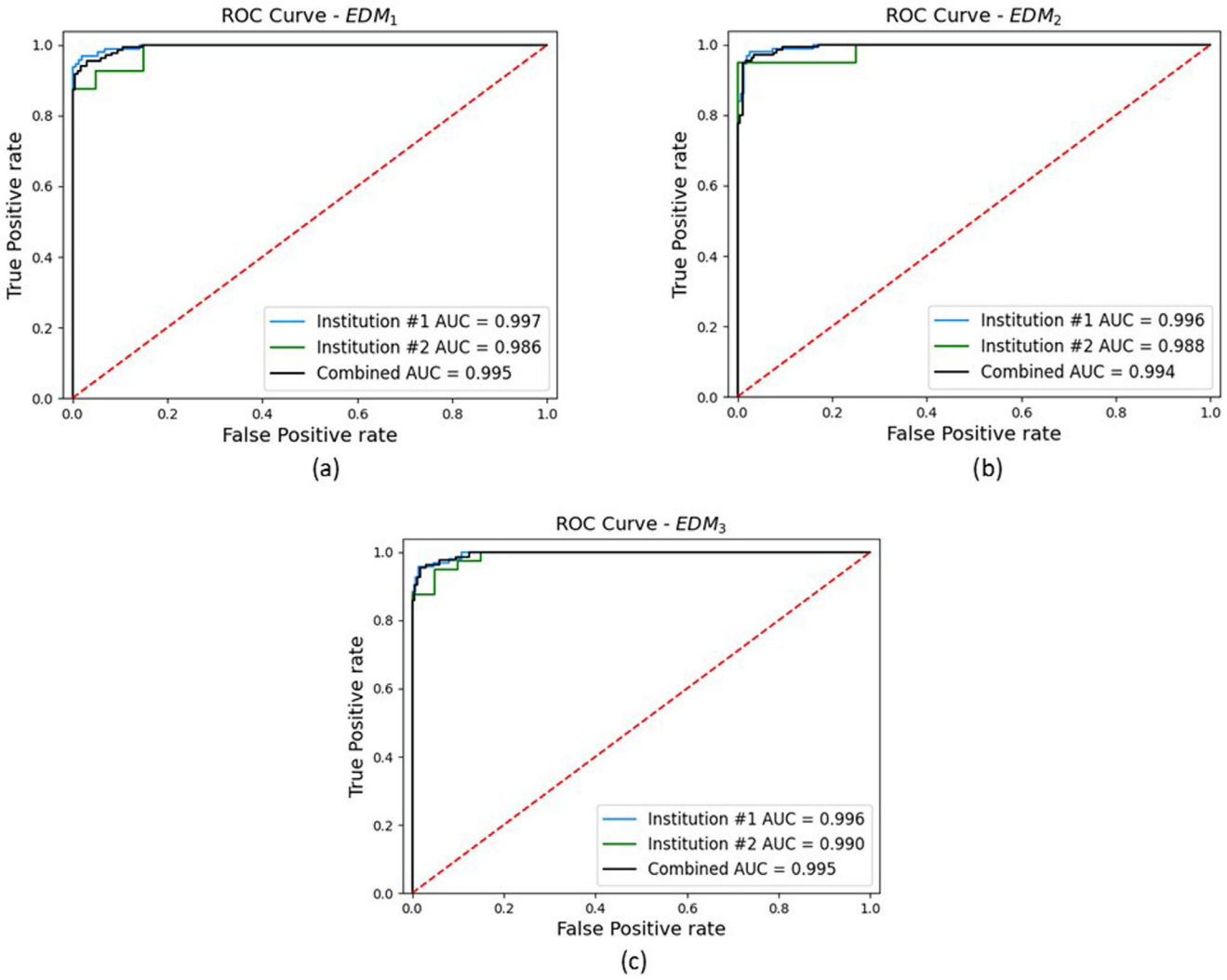



**FIGURE 4** Receiver operating characteristic (ROC) curves to represent the classification performance of $EDM_1$ (a), $EDM_2$ (b), and $EDM_3$ (c) on our test dataset. The three curves on each graph represent the performances of the model on the test set from Institution 1 only (blue), the test set from Institution 2 only (green), and the combination of both sets (black). The area-under-curve is also shown for each curve.

**TABLE 2** Classification results of the three models on the test datasets using a threshold that yields at least 99% specificity

| Model | Test set | Specificity | Sensitivity | *F*-1 score | MCC |
|---|---|---|---|---|---|
| **$EDM_1$** | Institution 1 | 0.99 | **0.96** | **0.97** | **0.95** |
| | Institution 2 | 0.99 | 0.88 | 0.93 | 0.84 |
| | Combined set | 0.99 | 0.93 | 0.95 | 0.92 |
| **$EDM_2$** | Institution 1 | 0.99 | 0.95 | 0.96 | 0.94* |
| | Institution 2 | 0.99 | **0.95** | **0.97** | **0.93*** |
| | Combined set | 0.99 | **0.95** | **0.97** | **0.94*** |
| **$EDM_3$** | Institution 1 | 0.99 | 0.95 | 0.96 | 0.94* |
| | Institution 2 | 0.99 | 0.88 | 0.93 | 0.84* |
| | Combined set | 0.99 | 0.93 | 0.95 | 0.92* |

*Note*: The numbers in bold represent the better score obtained for each respective test set.
*Results from corresponding rows were found to be statistically significant ($p$-value <0.05).
Abbreviation: MCC, Matthews correlation coefficient.

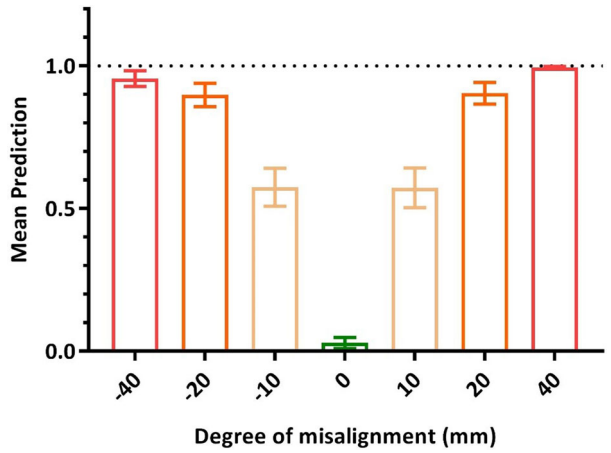



**FIGURE 5** Column bars to represent the mean misalignment prediction of $EDM_2$ on the combined test dataset as a function of caudal-cranial misalignment distances. The error bars represent the 95% confidence interval of the mean value.

validated on a larger test set composed of unseen data from two different institutions as compared to the non-DL techniques that were validated using a 10-fold cross-validation method for a training-testing dataset composed of 57 patients from a single institution. As compared to a commonly used image similarity metric (mutual information[36]), our model has also been shown to produce more discriminative scores for off-by-one vertebral-body misalignment error detection, as shown in Figure A1.

$EDM_2$ obtained significantly higher mean prediction scores for 10, 20, and 40-mm caudal-cranial misalignments as compared to the correct clinical alignments. This demonstrates the potential of $EDM_2$ in detecting misalignments smaller and larger in magnitude than 1 vertebral body, in addition to the off-by-one vertebral-body misalignments, which validates the appreciable value that the model can add to the clinical workflow and to the patients' safety.

When applied to the 303 test images, $EDM_2$ resulted in two false-positives and seven false-negatives using a $\geq$99% specificity threshold. In one of the false-positives, the CBCT clinical setup instructions indicated the prioritization of the soft tissue alignment over the bony alignment. Therefore, on the CBCT–CT registration, misalignments were present at the vertebral bodies, as shown in Figure 6 (Case 1), which we believe likely triggered the misclassification. For the other false-positive case, considerable streak artifacts were observed on the CBCT image, which may have affected the model output. Of the seven false-negative cases, four had a limited field of view, where part of the patient anatomy was not captured on the CBCT as shown in Figure 6 (Case 2). The other three cases showed considerable streak artifacts on the CBCT (see Figure 6, Cases 2 and 3), which could be due to beam hardening effects, photon starvation, or exponential edge gradient effects.[37] The image properties discussed before may have contributed to the wrong classification of those few cases; however, further tests on a larger and more diverse dataset are required to verify the exact causes of failure. Future work could include a dedicated model that flags lower quality scans such that the results of EDM can be interpreted accordingly. Alternatively, attention gates[38] could be incorporated in EDM such that the model focuses on targeted regions instead of irrelevant regions that may contain artifacts.

Our experiments have shown that an error-detection model based on single-institutional data is not sufficiently generalizable to cross-institutional data. We demonstrated that incorporating a small amount of cross-institutional training data recovers some of the performance. However, patient data from two institutions may not be enough to capture the variability in scanning protocol, image quality, and registration techniques across all treatment facilities and treatment machines. The performance of $EDM_3$, which was trained on an imbalanced multi-institutional dataset, has demonstrated that the model does not have similar classification abilities on both institutions test data. Hence, this paper calls for the importance and need for more data across multiple facilities, such that the generalizability power of the model could be improved, and the error detection system could benefit a wider range of facilities. Further work in this direction should include a determination of a minimum diversity of cross-institutional data that would lead to an expectation of similar model performance on data from an unseen institution.

Another limitation of this study is the use of 2D orthogonal slices as input to the EDM, instead of the whole 3D volumes. Although the 2D images lead to faster computation time, the amount of features captured by the model is limited to the selected slices. A 3D model could capture many more useful features from the entire scans, which could further improve the detection of misalignment errors. With the current systems available in the clinic, the 3D model is currently deemed impractical due to its memory requirements. However, with the rise in computation technologies and easier access to high-end GPUs, the 3D EDM could be a more effective and practical approach in the future, as compared to the 2D EDM.

Our algorithm also focuses on one particular type of error that could occur in CBCT-guided radiotherapy. Other subtle errors occurring at the soft tissue level during the registration of the CBCT to the planning CT can possibly lead to suboptimal treatments and must be avoided. Our algorithm is not currently optimized to catch these soft-tissue misalignments. Furthermore, our EDM was trained solely on thoracic and abdominal cases, which only makes a fraction of CBCT-guided radiotherapy treatments. Other sites commonly treated using CBCT-guided radiotherapy include the head and neck and pelvic area. Future works involve expanding

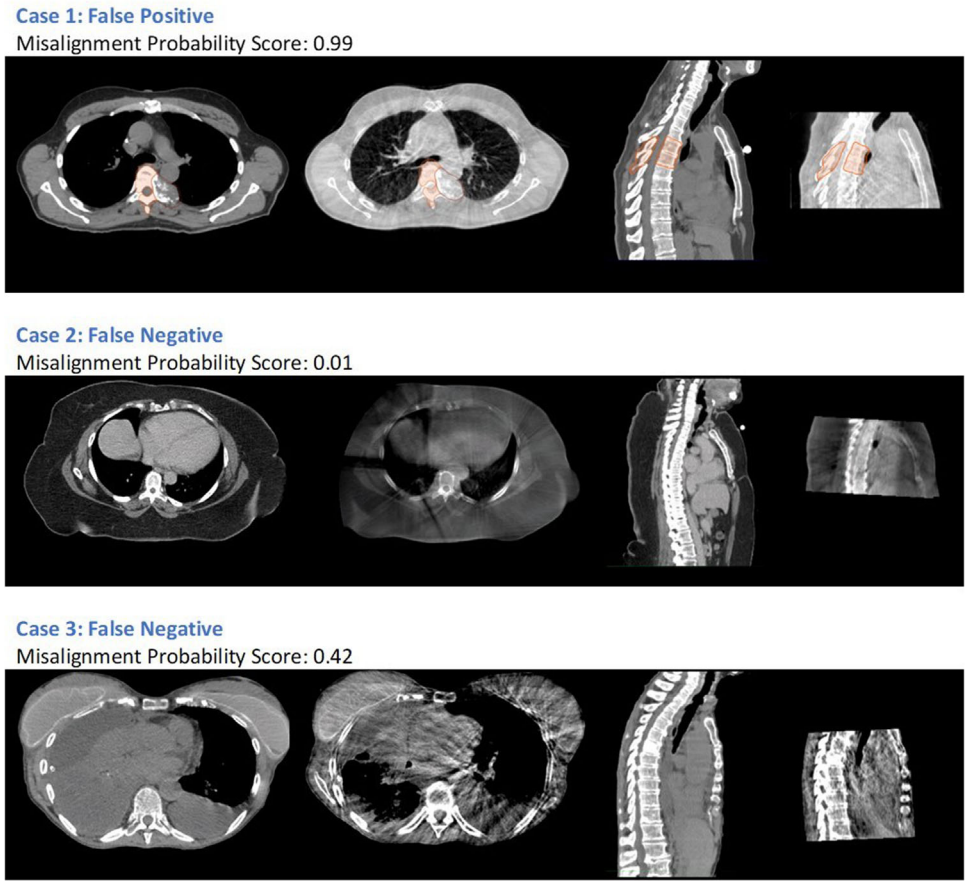



**FIGURE 6** Three examples of misclassification by EDM2 using a ≥99% specificity threshold. For each case, the planning computed tomography (CT) slice is shown to the left of the corresponding cone-beam computed tomography (CBCT) slice. Case 1 shows a correct clinically performed registration where the soft tissue alignment was prioritized over the bony alignment (the contours of the planning target volumes are shown to demonstrate the misalignment present at the vertebral body). Case 2 shows an example where part of the patient body was not present on the CBCT axial scan, in addition to considerable streak artifacts. In Case 3, substantial streak artifacts were observed on the CBCT scan.

our algorithm to catch different types of misalignment errors in the thoracic and abdominal cases, as well as for the other treatment sites mentioned earlier.

Even though wrong-vertebral-body misalignment error occurs very rarely in the clinic, it can have serious consequences to the patient if not detected prior to treatment. With its strong error-catching ability, our algorithm could prove to be useful as a fully automated online secondary safety check to the therapist, minimizing the risk of wrong vertebral-body registration during patient alignment. It can be deemed even more useful in facilities that have a shortage of radiation therapy technologists, which is often seen in underserved communities and developing countries.[39–41] As compared to real-time monitoring systems using surface imaging,[10–12] our software requires minimal external hardware (standalone computer plus interface hardware and software) and a low up front cost for clinical implementation. Hence, our software can be of particular interest to facilities that lack resources for additional equipment for patient safety.

Additionally, our algorithm could be used as an aid or supplement to image review performed by medical physicists as part of weekly chart check QA of external beam radiotherapy treatments. Although physicians are responsible for approving image guidance results, medical physicists commonly spot-check image alignments on a daily or weekly basis, which has a high risk priority number in the radiation therapy workflow[42] and is time-consuming.[43] Our tool offers the possibility of automatically analyzing all of the scans of patients being treated within a particular time frame and flag all of the possible anomalies detected through a time-stamped report. This way, the physicist can effectively review the handful of treatments that have been flagged as highest probability of treatment error, instead of randomly choosing plans or going through all the treatment plans. Hence, our tool can not only make treatment QA more time-effective, but it can also make it more robust to incident detection and improve incident learning.

A thorough description of clinical implementation of the error-detection tool is beyond the scope of this paper

and the subject of ongoing work in our lab. Based on our experience developing this algorithm, we believe that there are three essential aspects of commissioning that would be performed at a clinic implementing this tool. First, the expected false-positive rate should be validated using unmodified clinical data from the site. Second, the sensitivity to IGRT errors should be benchmarked using a standardized process such as a plot of prediction scores for synthetic misalignments of 10, 20, and 40 mm.[2] Production of such a plot could be automated by software provided by the algorithm developer. Third, an end-to-end test using an anthropomorphic phantom should be performed.

## 5 | CONCLUSION

Off-by-one vertebral-body misalignments represent a rare but serious error in IGRT. An automatic deep-learning-based misalignment error detector was proposed, which can flag potential cases of off-by-one vertebral-body misalignment in the registration of the planning CT to the CBCT. Our results have shown that our algorithm has sufficient sensitivity and specificity for routine clinical use. Algorithm robustness was validated by applying it to interinstitutional data. This algorithm can be used as an online safety-check during CBCT-based-guided radiotherapy and can also facilitate the QA of external beam radiotherapy treatments by aiding medical physicists during regular physics chart checks.

### ACKNOWLEDGMENT

The research reported in this study was supported by the Agency for Healthcare Research and Quality (AHRQ) under award number 1R01HS026486.

### DATA AVAILABILITY STATEMENT

Aggregate data is available upon request to the corresponding author. The data is not publicly available due to privacy or ethical restrictions.

### CONFLICT OF INTEREST

The authors have no conflicts to disclose.

### REFERENCES

1. Zelefsky MJ, Kollmeier M, Cox B, et al. Improved clinical outcomes with high-dose image guided radiotherapy compared with non-IGRT for the treatment of clinically localized prostate cancer. *Int J Radiat Oncol Biol Phys*. 2012;84(1):125-129. https://doi.org/10.1016/j.ijrobp.2011.11.047
2. Margalit DN, Chen YH, Catalano PJ, et al. Technological advancements and error rates in radiation therapy delivery. *Int J Radiat Oncol Biol Phys*. 2011;81(4):e673-e679. https://doi.org/10.1016/j.ijrobp.2011.04.036
3. Fraass BA. Impact of complexity and computer control on errors in radiation therapy. *Ann ICRP*. 2012;41(3-4):188-196. https://doi.org/10.1016/j.icrp.2012.06.011
4. Hendee WR, Herman MG. Improving patient safety in radiation oncology. *Med Phys*. 2011;38:78-82. https://doi.org/10.1118/1.3522875
5. Smith S, Wallis A, King O, et al. Quality management in radiation therapy: a 15 year review of incident reporting in two integrated cancer centres. *Tech Innov Patient Support Radiat Oncol*. 2020;14:15-20. https://doi.org/10.1016/j.tipsro.2020.02.001
6. Mazur LM, Mosaly PR, Hoyle LM, Jones EL, Chera BS, Marks LB. Relating physician's workload with errors during radiation therapy planning. *Pract Radiat Oncol*. 2014;4(2):71-75. https://doi.org/10.1016/j.prro.2013.05.010
7. Belletti S, Dutreix A, Garavaglia G, et al. Quality assurance in radiotherapy: the importance of medical physics staffing levels. Recommendations from an ESTRO/EFOMP joint task group. *Radiother Oncol*. 1996;41(1):89-94. https://doi.org/10.1016/S0167-8140(96)91799-5
8. Ezzell G, Chera B, Dicker A, et al. Common error pathways seen in the RO-ILS data that demonstrate opportunities for improving treatment safety. *Pract Radiat Oncol*. 2018;8(2):123-132. https://doi.org/10.1016/j.prro.2017.10.007
9. Shah M, Halalmeh DR, Sandio A, Tubbs RS, Moisi MD. Anatomical variations that can lead to spine surgery at the wrong level: part II thoracic spine. *Cureus*. 2020;12(6):e8684.
10. Pallotta S, Marrazzo L, Ceroti M, Silli P, Bucciolini M. A phantom evaluation of Sentinel™, a commercial laser/camera surface imaging system for patient setup verification in radiotherapy. *Med Phys*. 2012;39(2):706-712. https://doi.org/10.1118/1.3675973
11. Pallotta S, Simontacchi G, Marrazzo L, et al. Accuracy of a 3D laser/camera surface imaging system for setup verification of the pelvic and thoracic regions in radiotherapy treatments. *Med Phys*. 2013;40(1):011710. https://doi.org/10.1118/1.4769428
12. Schöffel PJ, Harms W, Sroka-Perez G, Schlegel W, Karger CP. Accuracy of a commercial optical 3D surface imaging system for realignment of patients for radiotherapy of the thorax. *Phys Med Biol*. 2007;52(13):3949-3963. https://doi.org/10.1088/0031-9155/52/13/019
13. Lamb JM, Agazaryan N, Low DA. Automated patient identification and localization error detection using 2-dimensional to 3-dimensional registration of kilovoltage x-ray setup images. *Int J Radiat Oncol Biol Phys*. 2013;87(2):390-393.
14. Jani SS, Low DA, Lamb JM. Automatic detection of patient identification and positioning errors in radiation therapy treatment using 3-dimensional setup images. *Pract Radiat Oncol*. 2015;5(5):304-311. https://doi.org/10.1016/j.prro.2015.06.004
15. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29. https://doi.org/10.1038/s41591-018-0316-z
16. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:3431-3440.
17. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. Curran Associates. 2012:25.
18. Razzak MI, Naz S, Zaib A. Deep learning for medical image processing: overview, challenges and the future. In: Dey, N., Ashour, A., Borra, S., eds. *Classification in BioApps. Lecture Notes in Computational Vision and Biomechanics*, vol 26. Springer, Cham. 2018. https://doi.org/10.1007/978-3-319-65981-7_12
19. Cai L, Gao J, Zhao D. A review of the application of deep learning in medical image classification and segmentation. *Ann Transl Med*. 2020;8(11):713. https://doi.org/10.21037/atm.2020.02.44
20. Nibali A, He Z, Wollersheim D. Pulmonary nodule classification with deep residual networks. *Int J Comput Assist Radiol Surg*. 2017;12(10):1799-1808. https://doi.org/10.1007/s11548-017-1605-6

[2] We thank the anonymous reviewer of an early version of this manuscript for this suggestion.

21. Gao M, Bagci U, Lu L, et al. Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Comput Methods Biomech Biomed Eng: Imaging Visual*. 2018;6(1):1-6. https://doi.org/10.1080/21681163.2015.1124249
22. González G, Ash SY, Vegas-Sánchez-Ferrero G, et al. Disease staging and prognosis in smokers using deep learning in chest computed tomography. *Am J Respir Crit Care Med*. 2018;197(2):193-203. https://doi.org/10.1164/rccm.201705-0860OC
23. Higgins J, Bezjak A, Franks K, et al. Comparison of spine, carina, and tumor as registration landmarks for volumetric image-guided lung radiotherapy. *Int J Radiat Oncol Biol Phys*. 2009;73(5):1404-1413. https://doi.org/10.1016/j.ijrobp.2008.06.1926
24. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:4700-4708.
25. Romero M, Interian Y, Solberg T, Valdes G. Targeted transfer learning to improve performance in small medical physics datasets. *Med Phys*. 2020;47(12):6246-6256. https://doi.org/10.1002/mp.14507
26. Amiri M, Brooks R, Rivaz H. Fine-tuning U-Net for ultrasound image segmentation: different layers, different outcomes. *IEEE Trans Ultrason Ferroelectr Freq Control*. 2020;67(12):2510-2518.
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
28. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems*. Curran Associates. 2019:32. https://doi.org/10.48550/arXiv.1902.07208
29. Da K. A method for stochastic optimization. 2014. arXiv preprint arXiv:1412.6980.
30. Janocha K, Czarnecki WM. On loss functions for deep neural networks in classification. 2017. arXiv preprint arXiv:1702.05659.
31. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27(8):861-874. https://doi.org/10.1016/j.patrec.2005.10.010
32. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. European Conference on Information Retrieval. Berlin, Heidelberg. Springer; March 2005:345-359.
33. Guilford JP. Psychometric methods. 1954.
34. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1):1-13. https://doi.org/10.1186/s12864-019-6413-7
35. Busscher I, Ploegmakers JJ, Verkerke GJ, Veldhuizen AG. Comparative anatomical dimensions of the complete human and porcine spine. *Eur Spine J*. 2010;19(7):1104-1114. https://doi.org/10.1007/s00586-010-1326-9
36. Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P. Multimodality image registration by maximization of mutual information. *IEEE Trans Med Imaging*. 1997;16(2):187-198.
37. Schulze R, Heil U, Groβ D, et al. Artefacts in CBCT: a review. *Dentomaxillofac Radiol*. 2011;40(5):265-273. https://doi.org/10.1259/dmfr/30642039
38. Schlemper J, Oktay O, Schaap M, et al. Attention gated networks: learning to leverage salient regions in medical images. *Med Image Anal*. 2019;53:197-207. https://doi.org/10.1016/j.media.2019.01.012
39. Agarwal JP, Krishnatry R, Panda G, et al. An audit for radiotherapy planning and treatment errors from a low-middle-income country centre. *Clin Oncol*. 2019;31(1):e67-e74. https://doi.org/10.1016/j.clon.2018.09.008
40. Izewska J, Andreo P, Vatnitsky S, Shortt KR. The IAEA/WHO TLD postal dose quality audits for radiotherapy: a perspective of dosimetry practices at hospitals in developing countries. *Radiother Oncol*. 2003;69(1):91-97. https://doi.org/10.1016/S0167-8140(03)00245-7
41. Izewska J, Vatnitsky S, Shortt KR. Postal dose audits for radiotherapy centers in Latin America and the Caribbean: trends in 1969-2003. *Rev Panam Salud Publica*. 2006;20(2-3):161-172. https://doi.org/10.1590/s1020-49892006000800013
42. Ford E, Conroy L, Dong L, et al. Strategies for effective physics plan and chart review in radiation therapy: report of AAPM Task Group 275. *Med Phys*. 2020;47(6):e236-e272. https://doi.org/10.1002/mp.14030
43. Ford EC, Terezakis S, Souranis A, Harris K, Gay H, Mutic S. Quality control quantification (QCQ): a tool to measure the value of quality control checks in radiation oncology. *Int J Radiat Oncol Biol Phys*. 2012;84(3):e263-e269. https://doi.org/10.1016/j.ijrobp.2012.04.036
44. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2015:234-241.


**How to cite this article:** Luximon DC, Ritter T, Fields E, et al. Development and interinstitutional validation of an automatic vertebral-body misalignment error detector for cone-beam CT-guided radiotherapy. *Med Phys*. 2022;1-14. https://doi.org/10.1002/mp.15927

## APPENDIX

1. Cone-beam computed tomography (CBCT) acquisition protocol
2. UNet-based spinal canal segmentation

For the thoracic and abdominal radiotherapy treatments, it is a common practice to contour the spinal canal as an organ at risk. However, there are a few cases where only part or none of the spinal canal is contoured within the computed tomography (CT) volume. During the orthogonal image extraction discussed in Section 2.2, the error detection algorithm relies heavily on the presence of the cord contour on the selected axial slice to obtain the vertebral-body position that is used to get the sagittal and coronal images. In the absence of the canal contour, the algorithm would fail in extracting the correct slices, leading to an algorithm failure. Hence, the authors decided to implement a 2D UNet-based spinal canal segmentation (SCS) algorithm that could segment the canal from the selected axial image of the planning CT and avoid the error detection algorithm from failing.

The SCS model was based on the UNet architecture,[44] which is composed of a contracting path that captures contextual features from the input, and an expanding path that extracts the features

**TABLE A1** Summary of the protocol used to acquire and reconstruct the cone-beam computed tomography (CBCTs) used in our experiments

| | Reconstruction method | No. CBCT scans | kVp (kV) | No. full-fan scans | No. half-fan scans |
|---|---|---|---|---|---|
| **Institution 1** | Standard | 992 | 100–125 | 990 | 2 |
| | Auto | 210 | 125–140 | 210 | 0 |
| | Sharp | 82 | 100–125 | 80 | 2 |
| | Smooth | 32 | 125 | 19 | 13 |
| **Institution 2** | Standard | 96 | 110–125 | 95 | 1 |
| | Auto | 3 | 125 | 3 | 0 |
| | Sharp | 1 | 125 | 1 | 0 |
| **Total** | | 1416 | – | 1398 | 18 |

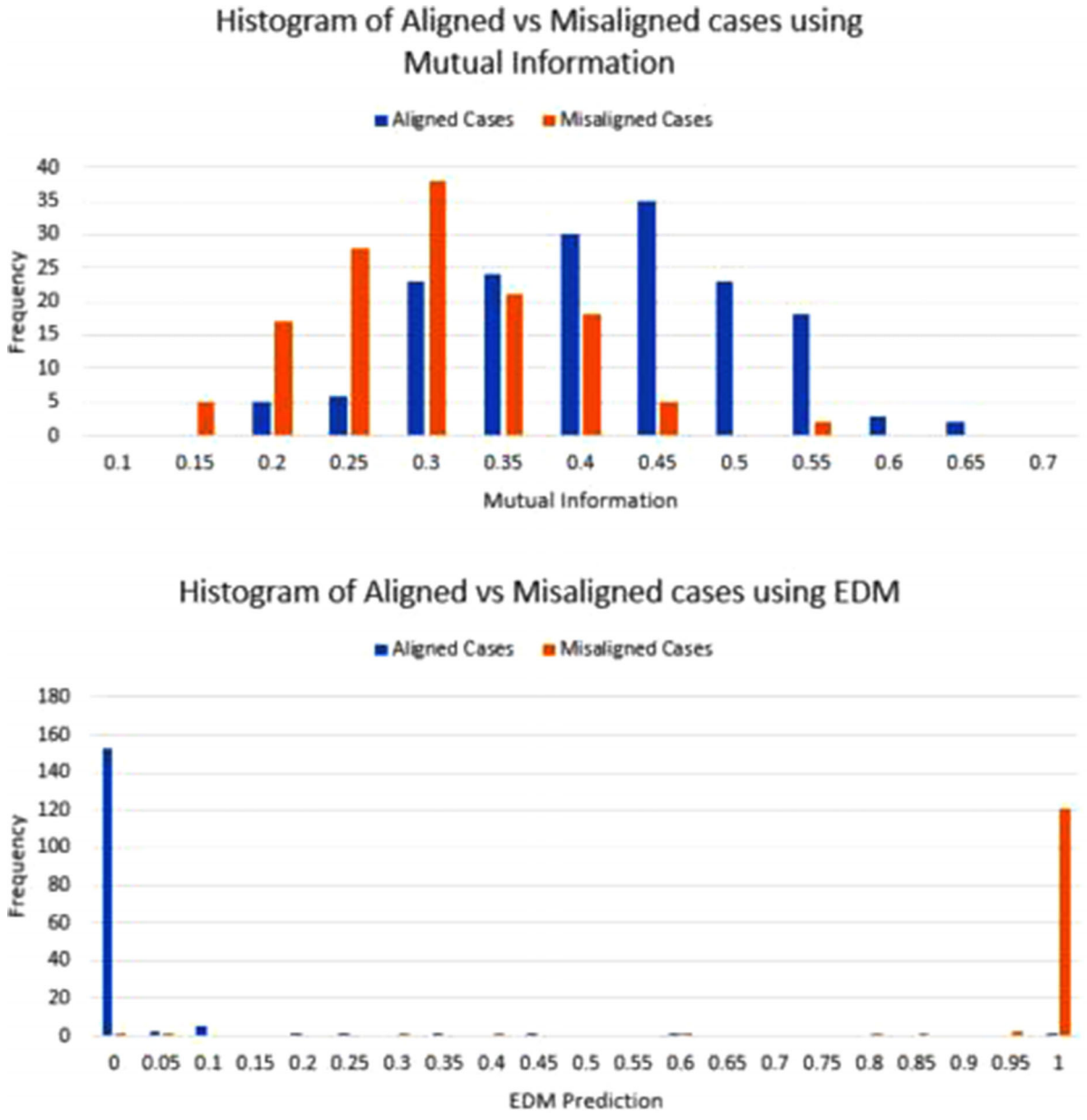



**FIGURE A1** Histograms showing the distribution of scores for the aligned and misaligned cases using (i) mutual information and (ii) our deep learning method ($EDM_2$)

**TABLE A2** Description of the dataset used to train, validate, and test the spinal canal segmentation (SCS) model

| | Number of patients | Number of axial slices |
|---|---|---|
| Training set | 147 | 22 884 |
| Validation set | 15 | 2226 |
| Testing set | 22 | 2914 |

obtained. This model was trained and tested using 184 patients' planning CT from Institution 1. The patient dataset was split into training, validation, and test set, as shown in Table A2. This dataset split was kept consistent to the one performed during the EDM experiment to avoid training SCS on images that would be used during the validation or testing phase of the EDM. The input to the model was a 150 × 150 axial image patch automatically extracted about the center of patient body. The binary mask of the spinal canal was obtained from the RT structure file of each CT dataset and used as ground-truth labels during model training and testing.

The SCS model was implemented using TensorFlow 2.2 with Keras backend. The BCE loss function was used during training. The model was trained using an Adam Optimizer[29] with a starting learning rate of $5 \times 10^{-4}$. During training, the model was evaluated after each epoch using its validation set, and the learning rate was reduced by a factor of 0.8 if the validation loss did not improve for five consecutive epochs. The model was trained until the validation loss did not improve for 20 consecutive epochs, or for a maximum of 200 epochs.

**TABLE A3** Results of the centroid comparisons between the ground-truth contours and the predicted contours

| Average separation (mm) | 1.51 |
|---|---|
| Standard deviation (mm) | 9.49 |
| No. of images with a separation >10 mm | 61 (2.1%) |

The model achieving the highest validation accuracy was then saved. SCM achieved convergence after five epochs.

To test the performance of the model, the distance between the centroid of the ground-truth contour and the centroid of the predicted contour was calculated for each of the 2914 test images. The average and the standard deviation of the calculated distances are reported in Table A3. The number of predictions that led to a centroid separation of more than 1 cm was also calculated.

The number of slices where the ground-truth centroid was found within the region predicted by the SCS model was calculated. Our results show that for 97.4% of the test images, the ground-truth centroid was found within the predicted contour. From the results obtained, the SCS was deemed to produce acceptable results such that it can be incorporated in the error detection algorithm as a secondary and independent method of determining the position of the vertebral body for orthogonal slice extraction.

## 3. EDM prediction versus mutual information