Development and evaluation of a deep learning based artificial intelligence for automatic identification of gold fiducial markers in an MRI-only prostate radiotherapy workflow

Christian Jamtheim Gustafsson, Johan Swärd, Stefan Ingi Adalbjörnsson, Andreas Jakobsson and Lars E Olsson

1 Department of Hematology, Oncology and Radiation Physics, Skåne University Hospital, Lund, Sweden
2 Department of Translational Sciences, Medical Radiation Physics, Lund University, Malmö, Sweden
3 Centre for Mathematical Sciences, Mathematical Statistics, Lund University, Lund, Sweden
4 Spectronic Medical AB, Helsingborg, Sweden

E-mail: christian.JamtheimGustafsson@skane.se

Keywords: gold fiducial marker, MRI only prostate, MRI only radiotherapy, MRI radiation therapy, prostate cancer, deep learning

Abstract

Identification of prostate gold fiducial markers in magnetic resonance imaging (MRI) images is challenging when CT images are not available, due to misclassifications from intra-prostatic calcifications. It is also a time consuming task and automated identification methods have been suggested as an improvement for both objectives. Multi-echo gradient echo (MEGRE) images have been utilized for manual fiducial identification with 100% detection accuracy. The aim is therefore to develop an automatic deep learning based method for fiducial identification in MRI images intended for MRI-only prostate radiotherapy. MEGRE images from 326 prostate cancer patients with fiducials were acquired on a 3T MRI, post-processed with N4 bias correction, and the fiducial center of mass (CoM) was identified. A 9 mm radius sphere was created around the CoM as ground truth. A deep learning HighRes3DNet model for semantic segmentation was trained using image augmentation. The model was applied to 39 MRI-only patients and 3D probability maps for fiducial location and segmentation were produced and spatially smoothed. In each of the three largest probability peaks, a 9 mm radius sphere was defined. Detection sensitivity and geometric accuracy was assessed. To raise awareness of potential false findings a ‘BeAware’ score was developed, calculated from the total number and quality of the probability peaks. All datasets, annotations and source code used were made publicly available. The detection sensitivity for all fiducials were 97.4%. Thirty-six out of thirty-nine patients had all fiducial markers correctly identified. All three failed patients generated a user notification using the BeAware score. The mean absolute difference between the detected fiducial and ground truth CoM was 0.7 ± 0.9 [0 3.1] mm.

A deep learning method for automatic fiducial identification in MRI images was developed and evaluated with state-of-the-art results. The BeAware score has the potential to notify the user regarding patients where the proposed method is uncertain.

1. Introduction

The superior soft tissue contrast of magnetic resonance imaging (MRI) compared to computed tomography (CT) has made MRI a widespread image modality for target delineation in prostate cancer radiotherapy (Blomqvist et al 2013, Metcalfe et al 2013, Menard et al 2018). Conventionally, CT and MRI images are used in combination where the images are registered into a common frame of reference. Image registration between CT and MRI for the prostate can lead to systematic spatial uncertainties in the radiotherapy planning (Nyholm et al 2009) where a spatial uncertainty of 1.7–2 mm for patients with and without fiducial markers has been reported (Roberson et al 2005, Nyholm et al 2009, Korsager et al 2016, Wegener et al 2019). Such a systematic registration error could result in a systematic treatment error, and lead to a displacement of
the dose distribution away from the target, and compromise tumor control (van Herk et al 2000, van Herk 2004).

MRI-only radiotherapy workflows, i.e. workflows where the CT is excluded and the prostate external beam radiotherapy workflow are solely based on MRI images avoid such image registration errors (Jonsso et al 2019) and have now been introduced into the clinic (Tyagi et al 2017, Tenhunen et al 2018, Greer et al 2019, Persson et al 2020). Further, such a workflow avoids anatomical differences, such as bladder and rectum filling between CT and MRI examinations, which can affect delineation of the target and the organ at risk (Kerkhof et al 2008, Schmidt and Payne 2015).

Gold fiducial markers are according to clinical routine inserted into the prostate to allow high precision target positioning. Such fiducials are challenging to identify and differentiate towards intra-prostatic calcifications when only MRI images are available (Kapanen et al 2013, Ghose et al 2016, Dinis Fernandes et al 2017, Maspero et al 2017, 2018). This challenge has been reported to be the greatest obstacle for a successful implementation of MRI-only prostate radiotherapy (Tyagi et al 2017, Tenhunen et al 2018). Gold fiducial markers can be detected by exploiting the difference in magnetic susceptibility between the surrounding tissue and the fiducial. By using gradient echo based acquisition sequences an increased detection sensitivity can be achieved (Schenck 1996, Port and Pomper 2000, Nosrati et al 2019).

In our previous paper we introduced the multi-echo gradient echo sequence (MEGRE) for gold fiducial marker identification where multiple slices could be visualized for multiple echo times (Gustafsson et al 2017a). Differentiation between fiducial and intra-prostatic calcifications could thereby be facilitated due to their difference in magnetic susceptibility. This was possible as the expansion of the gold fiducial marker signal void took place more rapidly compared to the signal void change for calcifications with respect to echo time. This effect can also be easily recognized by the human eye when scrolling through the images.

Dinis Fernandes et al (2017) found that image distortion from rectal gas was an issue for their fiducial detection method, using a single echo balanced steady-state free precession acquisition sequence. The multiple echoes available from MEGRE images offers the possibility to use the lower echo time data of the series and thereby avoid problems with rectal gas disturbing the image (Gustafsson et al 2017a), figure 1.

Previous studies using MEGRE images for manual gold fiducial marker identification showed a 100% true positive detection rate (TPR) for 40 patients using two observers and a TPR of 99.6% for 44 patients using four observers (Jamtheim Gustafsson 2019, Persson et al 2020). This is an improvement compared to earlier manual detection studies performed, using one or more other acquisition sequences which reported a TPR between 91% and 98% (Kapanen et al 2013, Ghose et al 2016, Maspero et al 2017, 2018, Tenhunen et al 2018, Pathmanathan et al 2019). A workflow where the identification of the gold fiducial markers occur automatically has been suggested to save time and resources and eliminate inter-observer differences but the TPR for these automatic methods has only been reported to be 84%–95% (Ghose et al 2016, Dinis Fernandes et al 2017, Maspero et al 2017, Gustafsson et al 2017a). The methods have therefore not yet been considered ready for clinical implementation, due to the necessary improvements needed in method and quality control (QC) procedures to avoid or raise warning flags for misclassifications.

Development of deep learning convolutional neural networks has positioned itself as one of the most promising methods for development of medical image based artificial intelligence (AI) for radiotherapy (Meyer et al 2018, Sahiner et al 2019). MEGRE images builds a proven and solid foundation for gold fiducial...
marker detection in an MRI-only workflow and is therefore a suitable image material to use in exploration of deep learning based automated detection methods.

The aim of this work was therefore to (1) develop and evaluate an automatic method for gold fiducial marker detection using a deep learning convolutional neural network AI and (2) implement a QC method that issue notifications for patients with challenging fiducial detection where the results may be associated with uncertainties.

2. Methods

For development of deep learning neural network models, it is common to divide the available data into three separate datasets, namely a ‘training dataset’, a ‘validation dataset’ and a ‘test dataset’. The training dataset was used to train and fit the parameters of the model, and the model was continuously validated on a validation dataset during training. The validation dataset guided manual tuning of the model hyperparameters in the training process and monitored under- and over fitting of the training data. Random subsampling validation was used to ensure generalization throughout the development process. The test dataset was then used to evaluate the performance of the model on unseen data and thus estimated the generalization error. This process is referred to as inference. The test dataset was never exposed to the neural network in the training process (Gibson et al 2018). Scripting source code, configuration, final model file and software description used in this study to pre-process data or to train and apply the model is available on GitHub at https://github.com/jamtheim/MEGRE-Net. All datasets and annotations used in this study are publicly available and instructions for data access can be found at https://datasets.aida.medtech4health.se/10.23698/aida/megre (Jamtheim Gustafsson 2020).

2.1. Patient selection

For the training and validation dataset, a total of 326 prostate cancer patients without hip prosthesis referred to primary external beam radiation therapy (EBRT) were included in the study. Mean body weight was 86.3 ± 14.0 (1SD) [55–135] kg, mean age was 71.7 ± 5.6 (1SD) [52–85] yr. The MEGRE MRI-data was acquired between January 2016 and May 2019. The training and validation datasets were defined as an 85%/15% split for a 5-time random subsampling validation, i.e. 85% of all subjects were randomly selected to the training dataset and 15% were randomly selected to the validation dataset. No subjects were allowed to exist in more than one validation dataset.

The test dataset consisted of 39 prostate cancer patients without hip prosthesis, which previously had been included in an MRI-only treatment study (Persson et al 2020). Mean body weight for the test dataset was 85.4 ± 12.2 (1SD) [62–113] kg, mean age was 71.3 ± 5.4 (1SD) [49–81] yr. The MEGRE MRI-data was acquired between April 2017 and May 2018.

All patients in all datasets were subject to insertion of three inferior-superior long axis-oriented cylinder-shaped intra-prostatic gold fiducial markers with length 5.0 mm and diameter 1.0 mm. The fiducials were inserted perineally according to clinical routine by an oncologist 2 weeks prior to image acquisition. Five patients in the training/validation dataset lost one fiducial after insertion and was therefore left with only two fiducials.

2.2. MRI acquisition

All MRI scans were performed using a 3T wide bore 70 cm MRI system equipped with a flat table top (GE Discovery 750 W, software versions DV25.0R02-1549b and DV26.0R03-1831b, General Electric Healthcare, Milwaukee, WI, USA). The MRI system was subjected to monthly QC using vendor-specific coil tests and a large volume geometry phantom to assess geometric accuracy (Spectronic Medical AB, Helsingborg, Sweden) (Gustafsson et al 2017b, Wyatt et al 2018). A GE GEM Anterior Array 16 channel receiver array coil was placed on stiff coil bridges over the pelvic area of the patient. First order field map based automatic shimming was performed prior to imaging, where the shim volume was determined from the imaging volume, and automatically selected by the MRI system. An automatic MRI vendor-based image homogeneity correction (SCIC) was applied to the MRI images after image acquisition.

Patients were MRI scanned according to clinical routine where the MEGRE sequence was included for gold fiducial marker visualization. As outlined in our previous paper the MEGRE has a scan time of 5 min and acquires images of multiple slices for multiple echo times (same for all patients) with a slice thickness of 2.8 mm and an in-plane resolution of 1.46 mm × 1.46 mm (reconstructed to 0.47 mm × 0.47 mm), yielding an image matrix size of 512 × 512 with 28–34 slices for each patient (Gustafsson et al 2017a). The signal void, created from the fiducial, is depicted circular and increases in size with increasing echo time, thereby enabling fiducial visualization, figure 1.
2.3. Creation of ground truth label data
For 100 out of the 326 patients in the training/validation dataset the MEGRE images was imported into an in-house MATLAB software (v.2017b, Mathworks Inc. Natick, MA, USA) where the center of mass (CoM) coordinates of the gold fiducial marker signal voids were defined in the first echo images by one observer (medical physicist, CJG) and exported to a text file. For the remaining 226 of the patients in the training/validation dataset and the 39 patients in the test dataset MEGRE images containing echo number one or two was exported to the treatment planning system Eclipse v.15.1 (Varian Medical Systems, Palo Alto, CA, USA) and the CoM points of the gold fiducial marker signal voids were defined and exported as DICOM RT-structures (figures 2 and 3(a)). For 50 of these 226 patients the fiducial CoM definitions was manually determined by one observer (CJG). For the remaining 176 patients the fiducial CoM was manually determined by six observers (four radiotherapy- and two MRI technologists) with a distribution of 1, 6, 23, 24, 56, 66 patients respectively each. The fiducial CoM for the 39 patients in the test dataset was manually determined by two observers (both medical physicists, 23 (CJG) and 11 (other medical physicist) patients each, five together). All fiducial CoM definitions were restricted to exist within one physical slice (and not in between), regardless of used method, and was confirmed to be correct using CT images. The occurrence of intra-prostatic calcifications $\geq 2 \text{ mm}$ in the CT images corresponding to the test dataset was also recorded.

Using a Python- or Linux bash script the fiducial CoM coordinates for each patient were extracted from the text file or DICOM RT-structures. 3D binary masks with the dimensions of the MEGRE image volume for one echo were created using Convert3D (v.1.1.0 [https://sourceforge.net/projects/c3d/]) where each fiducial CoM coordinate defined the center of an inserted sphere (with a predetermined radius) assigned with a signal value equal to 1, background signal was 0 (figures 2 and 3(a)). These 3D masks defined the ground truth labels in the training/validation and test datasets and were created equal for all eight echo times. To further clarify, each 3D mask for each patient contained the ground truth for all three fiducials.

2.4. MEGRE data pre-processing
The MEGRE data for each patient, containing images with eight different echo times, was automatically sorted into echo time specific folders using a Linux bash script. The image data for each patient for each echo consisted of an image volume with 28–34 images ($512 \times 512 \text{ matrix}$) and was in this process also converted from DICOM to NIfTI format (Cox et al 2004) using Convert3D and stored in separate NIfTI files.

The vendor-based image homogeneity correction (SCIC) was unintentionally omitted for a large number of patients in the training/validation dataset after a software upgrade of the MRI console. To minimize differences between the subjects in the training/validation data an N4 bias field correction image filter (Sled et al 1998, Tustison et al 2010) was applied to all training/validation and test data for each patient and for each echo using the Simple ITK framework (Lowekamp et al 2013) implemented in Python.
Z-score normalization was enabled in the training/validation process. The same normalization was applied to the test dataset for each patient volume and for each echo after N4 bias field correction using a MATLAB script. Z-score normalization was computed by

$$Z = \frac{imageVolume - mean(imageVolume)}{std(imageVolume)}$$

i.e. each image volume was normalized to have mean of 0 and unit variance. Please see figure 3(b) for block diagram.

2.5. Data augmentation

To increase the variability in the training/validation dataset, offline data augmentation was applied individually to all 326 patients in the training/validation dataset after N4 bias field correction using an in-house MATLAB software (v.2019a, Mathworks Inc. Natick, MA, USA). The data augmentation consisted of a random image rotation around the superior-inferior patient axis for each patient in the interval $-15^\circ$ to $15^\circ$ using linear interpolation. Image data for each echo for each patient time was handled separately but subjected to the same amount of rotation. The same rotation was applied to the ground truth labels using nearest neighbor interpolation (to avoid producing non-binary mask values). The augmented data and labels for each patient data was saved as a new subject, producing a total of 652 training/validation subjects (original + augmented data) which equaled a total of 1946 fiducial objects (figure 3(b)).

2.6. Training and applying the deep learning model

The training and evaluation of the model was aimed for semantic segmentation and configured in NiftyNet v.0.5.0 (Gibson et al 2018), an open-source platform for convolutional neural networks based on TensorFlow v.1.12.0, where multiple popular deep learning models are available (Abadi et al 2015). Experiments were performed using two NVIDIA GeForce RTX 2080 Ti graphical processing units (GPUs), an AMD 2920X Ryzen Threadripper CPU and 128 GB DDR4 RAM, running on an Linux Ubuntu 18.04.1 server with CUDA version 10.0 and NVIDIA driver version 410.78. Image data from each separate echo time was defined as separate input channels to the neural network (total of eight channels). The HighRes3DNet (Li et al 2017) neural network model was used together with a relu activation function and an unweighted cross-entropy loss function, which measures the classification performance and outputs probability pixel values between 0 and 1. Adaptive moment estimation (Adam) was chosen as optimizer with a learning rate of 0.001 (Kingma and Ba 2014). Z-score normalization was enabled, called whitening in NiftyNet (equation (1)). The NiftyNet
platform utilized patch-based medical image analysis and a patch window size of $96 \times 96 \times 32$ pixels was used as input, where each label (0 or 1 in the binary mask) in the sampling prior had the same probability of being sampled. To further clarify, the probability of the patch window center voxel being label 1 was 0.5. The cross-entropy loss was calculated on the input patches. All input image volumes were prior to patch sampling zero padded with a width of two pixels in-plane, added to both ends of each dimension. Four pixels were added in the image slice direction in the same way with the objective to have at least 32 slices, thereby matching the patch window size in slice direction.

Training of the neural network depended on the value of multiple hyperparameters, thus a parameter search was performed. Investigated hyperparameters were size of ground truth radius, batch size, amount of L1 or L2 regularization on model trainable parameters and number of iterations. Hyperparameter optimization was guided by maximizing the dice similarity score between the ground truth 3D binary masks and the generated 3D binary masks after image inference (directly from NiftyNet) in the training and validation data. The metric was automatically calculated using a Python script. Tensorboard also provided guidance in avoiding overtraining by monitoring the cross-entropy loss (Abadi et al. 2015). Optimization experiments were conducted with a 5-time random subsampling validation, where the subject distribution was randomly chosen under the condition that no overlap between the validation datasets was allowed. The final hyperparameters after optimization consisted of a ground truth radius of 9 mm, a batch size of 2, no L1 or L2 regularization and trained for 40,000 iterations.

It was desirable to use as much data as possibly available for training a final model. Training of the final model was therefore performed with a 100%/0% split using the final model parameters which took 20 h to train. A stochastic component is always present in deep learning methods as model training is initiated with randomly assigned values for the neural network layer weights. Investigation of the train-retrain reliability of the method was therefore conducted for five training sessions, using the final model parameters and a fixed, fixed distribution of training/validation subjects.

Image inference on the test dataset, aimed to produce spherical segmentation masks around the CoM positions of the fiducials, was performed using the final model running on one GPU with an image inference patch windows size of $256 \times 256 \times 32$. For patients with more than 32 slices overlapping patches were created. However, once a voxel had been assigned a segmentation label in the output image volume, the voxel label value was not overwritten by information from an overlapping patch. The probability map for each patient after the SoftMax normalization in the neural network was exported with original image volume size (by setting the argument output_prob = true in the NiftyNet configuration file). Please see figure 3(b) for block diagram.

### 2.7. Evaluation of the final model

The probability map from the inference on the test dataset was spatially smoothed in a post-processing step by convolving the probability map with a spherical kernel (9 mm radius) where every voxel was assigned the value one using an in-house MATLAB software. This operation enabled better separation of overlapping probability segments as separate probability peaks were generated in the convolution process. The final location of the gold fiducial markers in the test dataset were determined by selecting only the three largest probability peaks from the convolved probability data. This corresponded to the number of inserted fiducials, known a-priori. These peak coordinates where used to insert three spherical objects with 9 mm radius, creating a final 3D binary mask for each patient in the test dataset, denoted ‘the final 3D binary mask’ below. The sphere CoM thereby defined the detected gold fiducial marker CoM. Evaluation of the final model applied to the test dataset was performed by calculating the dice similarity score between the final 3D binary masks and ground truth. Dice was thereby calculated for all three fiducials combined, and not for each individual fiducial. The geometric integrity of the applied model was assessed by comparing the detected gold fiducial marker CoM coordinates against the ground truth gold fiducial marker CoM coordinates for each individual fiducial for all spatial directions. Gold fiducial marker detection performance metrics were calculated by visually comparing the final 3D binary masks to ground truth 3D binary masks using ITK-SNAP v3.8.0 (Yushkevich et al. 2006). Please see figure 3(c) for block diagram.

### 2.8. QC method—BeAware score calculation

The existence of calcifications or other objects in the prostate giving rise to similar signal behavior as a fiducial marker could result in more than three large probability peaks in the probability map. Uncertainties could therefore occur in the post-processing steps where the three largest peaks were selected and possibly result in selection of false positive objects. To mitigate this, a ‘BeAware’ score reaching from 0 to 1 was developed as a QC method and calculated for every patient in the training/validation dataset. To notify for such events two BeAware score thresholds were determined from the training/validation dataset subject population. The first threshold corresponded to the case where the results should be discarded and the
Figure 4. BeAware score for the patients in the training/validation data for an 85%/15% split (random subsampling validation 1) visualized as blue asterisks plotted against corresponding dice similarity score. Two separate BeAware score thresholds were determined (dotted black lines) and identified patient results which should be discarded (BeAware score $\leq 0.1$) or manually checked ($0.1 < \text{BeAware score} \leq 0.7$). The BeAware scores and corresponding dice similarity score for the 39 patients in the test dataset is shown in filled green circles.

second threshold corresponded to the case where the results could contain uncertainties and should be manually checked (figure 4). The determined BeAware score thresholds were applied to the individual patient inference results of the test dataset (figure 3(c)). QC assessment for each individual patient could thereby be achieved. The BeAware score for one patient was calculated according to

$$
\text{BeAware score} = \left( \frac{\sum_{i,j,k \in F_1} \rho(i,j,k)}{\sum_{i,j,k \in \text{Total}} \rho(i,j,k)} \right)^{\alpha} \left( \frac{1}{T_1} \sum_{i,j,k \in F_1} \rho(i,j,k) \cdot \frac{1}{T_2} \sum_{i,j,k \in F_2} \rho(i,j,k) \cdot \frac{1}{T_3} \sum_{i,j,k \in F_3} \rho(i,j,k) \right)^{\beta} \tag{2}
$$

where $\rho$ is the probability values from 0 to 1 in the 3D un-convolved probability map, $i, j, k$ is a specific point, $F_{1,2,3}$ define the voxels within three respective objects in the final 3D binary mask (i.e. perfect inserted spheres) and $T_{1,2,3}$ is the amount of voxels in these objects respectively, which here is of equal size. 'Total' refers to all the available voxels $i, j, k$ in the probability map. The denominator sum where Total is used is therefore the sum of all probability values in the 3D un-convolved probability map. The parameters $\alpha$ and $\beta$ were in this study both set to 1.

The first factor of equation (2) represents a penalty from the existence of more than three probability peaks in the un-convolved data. To exemplify: Consider if all voxels within $F_1$, $F_2$ and $F_3$ are assigned with probability value 1 by the neural network (i.e. 100% confidence that the voxels are part of the fiducial marker sphere volume). Further, consider a calcification that has been mistaken by the neural network for a fiducial marker and also produced probability values larger than 0. The first factor in equation (2) will therefore equal a value less than 1 as there are probability values larger than 0 outside of the fiducial marker sphere volumes $F_1$, $F_2$ and $F_3$. The BeAware score will therefore be penalized. The second, third and fourth factor in equation (2), represents multiplication of the average probability per voxel in each defined fiducial sphere volume in the un-convolved data. This average probability for each fiducial sphere volume can be regarded as a measure of the quality to which the individual fiducial marker has been depicted by the neural network. If
Table 1. Dice similarity score for each random subsampling validation when applying the final model parameters on an 85%/15% split. The final model was trained on a 100%/0% split and applied to the test dataset.

| Training data | Validation data |
|---------------|-----------------|
| Subsampling validation 1 | 0.985 | 0.961 |
| Subsampling validation 2 | 0.986 | 0.975 |
| Subsampling validation 3 | 0.985 | 0.969 |
| Subsampling validation 4 | 0.981 | 0.967 |
| Subsampling validation 5 | 0.984 | 0.962 |
| Mean ± 1SD [min max] | 0.98 ± 0.002 [0.98 0.99] | 0.97 ± 0.005 [0.96 0.98] |

Table 2. Dice similarity score for five different training sessions with the identical 85%/15% training/validation split (randomly chosen) using the final model parameters.

| Training data | Validation data |
|---------------|-----------------|
| Session 1 | 0.986 | 0.966 |
| Session 2 | 0.985 | 0.963 |
| Session 3 | 0.984 | 0.964 |
| Session 4 | 0.983 | 0.965 |
| Session 5 | 0.979 | 0.962 |
| Mean ± 1SD [min max] | 0.98 ± 0.002 [0.98 0.99] | 0.96 ± 0.001 [0.96 0.97] |

all voxels within $F_1$, $F_2$ and $F_3$ are assigned with probability value 1 by the neural network, the average probability will also equal one. If not every voxel in each detected fiducial volume has a probability of one, the BeAware score will be further penalized with multiplication of factors less than 1.

3. Results

In table 1, the dice similarity score for the 5-time random subsampling validation with the 85%/15% split is presented for the training/validation data using the final model parameters. Results from the train-retrain reliability of the method using a randomly chosen identical 85%/15% split five times is presented in table 2. Standard deviation in tables 1 and 2 are calculated as population standard deviation.

Inference on the test dataset of 39 patients using the final model took 20 s per patient and another 12 s for post-processing. One-hundred and fourteen out of a total of 117 gold fiducials were correctly identified, yielding a TPR (or detection sensitivity) of 97.4%. Intra-prostatic calcifications $\geq 2$ mm, measured in the CT images, were found in 22/39 (56%) patients. All fiducial markers were correctly identified for 36/39 patients. The mean dice similarity score for all 39 patients after post-processing was $0.92 \pm 0.07$ (1SD) [0.64 1.00] while it was $0.94 \pm 0.04$ (1SD) [0.84 1.00] after excluding the three failed patients, calculated from the final 3D binary masks. The three failed patients had one calcification each mistaken for a fiducial marker.

The mean absolute difference between the detected fiducials and ground truth CoM was $0.73 \pm 0.93$ (1SD) [0.31 mm ($n = 36$, excluding the three failed patients). A more detailed analysis of the non-absolute spatial difference components revealed a mean difference of $-0.20 \pm 0.35$ (1SD) [$-0.94$ 0.47] mm, $-0.12 \pm 0.37$ (1SD) [$-1.4$ 0.94] mm and $0.13 \pm 1.04$ (1SD) [$-2.8$ 2.8] mm for anterior-posterior, left-right, and inferior-superior (slice) direction respectively.

The patient based QC assessment was calculated by the BeAware score and are plotted against the dice similarity score for all patients in the training/validation dataset for random subsampling validation 1 (figure 4). Two BeAware score thresholds were defined, based on the BeAware score results from patients in the training/validation dataset (inference results from the test dataset were not available). The first BeAware score threshold of $\leq 0.1$ was selected as it identified patient results that should be discarded, demonstrated by two patients in the training/validation 85%/15% split with BeAware scores close to 0 (lower dotted black line in figure 4). The second BeAware score threshold was selected to be $>0.1$ but $\leq 0.7$ and identified patient results, which could contain uncertainties (upper dotted black line in figure 4). The value 0.7 was chosen as it identified two patients in the training/validation 85%/15% split with dice similarity score $<0.65$, and both patients had false positive fiducial marker identifications (figure 4).

In the test dataset the mean BeAware score was $0.79 \pm 0.11$ (1SD) [0.392–0.892], hence no patients had a BeAware score $\leq 0.1$ which was the lower threshold. BeAware score notifications were issued for the larger threshold for seven out of the 39 patients where three of these patients corresponded to the three patients where the fiducial detection failed (BeAware scores 0.640–0.670). BeAware score notifications were thereby
issued for four patients in spite of their high dice similarity score (>0.90), who all had successful fiducial detection (figure 4).

4. Discussion

The proposed method for automatic gold fiducial marker detection in MEGRE images was based on a deep learning convolutional neural network, which was developed from the concept that the task could be defined as a semantic segmentation problem for the neural network to solve. A total of 326 patients were included in the study for training of the neural network, where the ground truth fiducial CoM for the training/validation data had been defined by seven observers. Subject data where subjected to N4 bias field filtering, data augmentation, Z-score normalization and a 5-time random subsampling validation process, where a final model was trained and applied to a test dataset of 39 patients, data not previously seen by the neural network.

A true positive fiducial detection rate of 97.4% (or sensitivity) was achieved on the test dataset with a mean geometric absolute deviation of 0.73 ± 0.93 mm (1SD) compared to ground truth. A QC approach to estimate the level of fiducial detection uncertainty from the deep learning model was implemented and evaluated. The results show that the developed model together with its QC methodology and minor adjustments can be implemented in our clinic as a decision support software in order to reduce observer bias and time consumption.

The performance of the model was comparable to the manual observer results in Gustafsson et al (2017a) for a similar patient group but it is of importance to notice that the results in Gustafsson et al (2017a) reflects the fiducial detection performance in a T2-weighted image volume, were fiducial detection was aided by MEGRE images as support. The performance for fiducial detection in the corresponding MEGRE image geometry has thereafter been reanalyzed and published in Jamtheim Gustafsson (2019) and yielded a mean TPR of 99.6 [99–100]% (n = 4 observers) and a mean geometric accuracy of 0.93 ± 0.88 (1SD) [0.00–5.88] mm.

Concerning this study, it should be noted that all fiducials in the test dataset were previously detected with 100% success by two observers in the MR-PROTECT study (Persson et al 2020). The results from the deep learning method presented in this study are very similar to the human observer level in terms of detection rate and geometric accuracy. The TPR of 97.4% in this study was better compared to currently existing non-deep learning methods for automatic fiducial detection in MRI images which have reported corresponding values of 84%–95% (Ghose et al 2016, Dinis Fernandes et al 2017, Maspero et al 2017, Gustafsson et al 2017a). Deep learning methods have previously also been applied for automatic detection of fiducial markers in x-ray images or radioactive brachytherapy seeds in MRI images, with a close to perfect TPR (Mylonas et al 2019, Sanders et al 2019), similar to the results in this study. Despite the differences in fiducial and image modality type, the deep learning approach seem to constitute a powerful methodology for this kind of task.

Three out of the 39 patients in the test dataset did not have all fiducial markers estimated correctly, each patient had one fiducial which failed detection due to calcifications mistaken for fiducials. For one of the failed patients the calcification was located in close connection to the fiducial and the signal voids from the objects overlap for larger echo times, i.e. echo 8 produces artefacts with 7–8 mm diameter. This gave rise to more challenging conditions for accurate separation and detection. In spite of this, the results in this study is a major improvement compared to our previous non-deep learning based method, which had a TPR of 84% (Gustafsson et al 2017a). By adding MRI phase data to the input of the neural network a further improved differentiation between fiducials and calcifications could potentially be reached, as suggested and used in Maspero et al (2017).

Data augmentation was in this study performed with a single isolated rotation around the superior-inferior patient axis. Increased degrees of rotational freedom were considered but the appearance of signal voids are dependent on the relative orientation of the gold fiducial marker in the static magnetic field (Jonsson et al 2012). An arbitrary rotation would therefore yield physically incorrect image data. This action was therefore avoided.

The use of random subsampling validation provided a mean to ensure that the size of the training/validation dataset was large enough for generalization. The maximum fluctuations in dice similarity score between the 5-time random subsampling validation in the 85%/15% split was 0.005 dice units, and 0.014 dice units for training and validation data, respectively. This was considered small and reasonable, as the maximum intrinsic fluctuations in the test-retest of an identical 85%/15% split was 0.007 dice units and 0.004 dice units for training and validation data, respectively.

A limitation of the method existed in the choice of the BeAware score thresholds. An upper threshold of 0.7 where chosen, and a notification was issued for four patients with 100% correctly identified fiducials. This was due to multiple probability peaks being produced and was penalized according to equation (2). For
such events, it is suggested that the operator carefully review the MEGRE images to manually confirm a correct identification. To reduce false positives (four patients) a larger dataset is possibly needed together with some small adjustments of the BeAware method and used thresholds. This could in a clinical implementation of the method help to avoid a false fiducial detection, as these results could be corrected by the operator. Further, this study was performed on MEGRE images from a single MRI vendor (GE) at a single magnetic field strength (3T). By publicly sharing the datasets and code for pre-processing and model development the authors promote development of a future multi-vendor solution, where the model can be trained with adequate multi-vendor MRI data, see https://github.com/jamtheim/MEGRE-Net.

The developed method provides a means to eliminate a manual time consuming task in the MRI-only workflow, and also reduce any related observation bias. The method demonstrates excellent detection and geometric performance, and pushes the limit for state-of-the-art automatic MRI-based fiducial detection were previous methods have not been adequate. It should also be emphasized that the method could be applied to conventional workflows where CT and MRI data are combined using image registration, and thereby provide similar benefits. The need for further patient specific QC measures of fiducial identification has previously been highlighted in recent work (Greer et al 2019, Jonsson et al 2019). The BeAware method introduced in this study could after further improvements fulfill some of those QC needs, such as provide warnings for the occurrence of calcifications and potential fiducial misidentifications. If further confirmation of a suggested fiducial location is desired, we propose an MRI-independent method previously developed by our group (Gustafsson et al 2018).

5. Conclusions

A deep learning convolutional neural network AI for automatic fiducial identification in multi-echo gradient echo MRI images was developed and evaluated. The neural network demonstrated a TPR of 97.4% with a mean geometric accuracy of 0.73 mm, and is the best result to date. The results are comparable to human performance, facilitated by the use of state-of-the-art technology. A QC method was also developed which issued notifications for patients with challenging or failed fiducial detection and could call upon the attention of the operator. The software is after minor adjustments considered ready for implementation as a support tool in our clinic, where it can facilitate time and resource minimization for MRI-only prostate radiotherapy.

Acknowledgments

This work was funded by AIDA—Analytic Imaging Diagnostics Arena, Grant Nos. 2017-02447, which is supported by the Swedish strategic innovation program Medtech4Health. The work was supported by VINNOVA Gentle Radiotherapy with Grant Nos. 2016-03847, 2016-02529 and 2018-04353, by ‘Allmänna sjukhusets i Malmö Stiftelse för bekämpande av cancer’, by ‘Onkologiska klinikens Stiftelse för bekämpande av cancer’, and by ‘The Research Funds of Skåne University Hospital’. The authors also acknowledge the support given by the staff at Spectronic Medical AB throughout the study.

Ethical statement

This study was approved by the Swedish ethical review authority with numbers 2013-742, 2016-801 and 2020-00216.

ORCID iD

Christian Jamtheim Gustafsson  
https://orcid.org/0000-0003-2931-5615

References

Abadi M et al 2015 White paper: TensorFlow: large-scale machine learning on heterogeneous systems
Blomqvist L, Bäck A, Ceberg C, Enblad G, Frykholm G, Johansson M, Olsson L and Zackrisson B 2013 MR in radiotherapy – an important step towards personalized treatment? Report from the SSM’s Scientific Council on Ionizing Radiation Within Oncology SSM
Cox R W et al A (sort of) new image data format standard: NiFTI-1 10th Annual Meeting of Organisation of Human Brain Mapping (Budapest, Hungary, June 2004)
Dinis Fernandes C, Dinh C V, Steggerda M J, Ter Beek L C, Smolic M, van Buuren L D, Pos F J and van der Heide U A 2017 Prostate fiducial marker detection with the use of multi-parametric magnetic resonance imaging Phys. Imaging Radiat. Oncol. 1 14–20
Ghose S, Mitra J, Rivest-Henault D, Fazollahi A, Stanwell P, Pichler P, Sun J, Fripp J, Greer P B and Dowling J A 2016 MRI-alone radiation therapy planning for prostate cancer: automatic fiducial marker detection Med. 

Phys. Med. Biol. 65 (2020) 225011

C Janson Gustafsson et al

11

IOP Publishing

Phys. Med. Biol. 65 (2020) 225011

11
Tyagi N, Fontenla S, Zelefsky M, Chong-Ton M, Ostergren K, Shah N, Warner L, Kadii M, Mechalakos J and Hunt M 2017 Clinical workflow for MR-only simulation and planning in prostate Radiat. Oncol. 12 119
van Herk M 2004 Errors and margins in radiotherapy Semin. Radiat. Oncol. 14 52–64
van Herk M, Remeijer P, Rasch C and Lebesque JV 2000 The probability of correct target dosage: dose-population histograms for deriving treatment margins in radiotherapy Int. J. Radiat. Oncol. Biol. Phys. 47 1121–35
Wegener D, Zips D, Thorwarth D, Weiss J, Othman A E, Grosse U, Notohamiprodjo M, Nikolaou K and Muller A C 2019 Precision of T2 TSE MRI-CT-image fusions based on gold fiducials and repetitive T2 TSE MRI-MRI-fusions for adaptive IGRT of prostate cancer by using phantom and patient data Acta Oncol. 58 88–94
Wyatt J, Hedley S, Johnstone E, Speight R, Kelly C, Henry A, Short S, Murray L, Sebag-Montefiore D and McCallum H 2018 Evaluating the repeatability and set-up sensitivity of a large field of view distortion phantom and software for magnetic resonance-only radiotherapy Phys. Imaging Radiat. Oncol. 6 31–38
Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC and Gerig G 2006 User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability Neuroimage 31 1116–28