Long short-term memory networks for proton dose calculation in highly heterogeneous tissues

Ahmad Neishabouri[a)]
Department of Medical Physics in Radiation Oncology, German Cancer Research Center - DKFZ, Im Neuenheimer Feld 280, D-69120, Heidelberg, Germany
Medical Faculty, University Heidelberg, Heidelberg, Germany
Heidelberg Institute for Radiation Oncology (HIRO), Heidelberg, Germany

Niklas Wahl
Department of Medical Physics in Radiation Oncology, German Cancer Research Center - DKFZ, Im Neuenheimer Feld 280, D-69120, Heidelberg, Germany
Heidelberg Institute for Radiation Oncology (HIRO), Heidelberg, Germany

Andrea Mairani
Heidelberg Ion-Beam Therapy Center (HIT), Im Neuenheimer Feld 450, D-69120, Heidelberg, Germany

Ullrich Köthe
Visual Learning Lab, Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Im Neuenheimer Feld 205, D-69120, Heidelberg, Germany

Mark Bangert
Department of Medical Physics in Radiation Oncology, German Cancer Research Center - DKFZ, Im Neuenheimer Feld 280, D-69120, Heidelberg, Germany
Heidelberg Institute for Radiation Oncology (HIRO), Heidelberg, Germany

(Received 17 June 2020; revised 9 November 2020; accepted for publication 20 November 2020; published 11 March 2021)

Purpose: To investigate the feasibility and accuracy of proton dose calculations with artificial neural networks (ANNs) in challenging three-dimensional (3D) anatomies.
Methods: A novel proton dose calculation approach was designed based on the application of a long short-term memory (LSTM) network. It processes the 3D geometry as a sequence of two-dimensional (2D) computed tomography slices and outputs a corresponding sequence of 2D slices that forms the 3D dose distribution. The general accuracy of the approach is investigated in comparison to Monte Carlo reference simulations and pencil beam dose calculations. We consider both artificial phantom geometries and clinically realistic lung cases for three different pencil beam energies.
Results: For artificial phantom cases, the trained LSTM model achieved a 98.57% γ-index pass rate ([1%, 3 mm]) in comparison to MC simulations for a pencil beam with initial energy 104.25 MeV. For a lung patient case, we observe pass rates of 98.56%, 97.74%, and 94.51% for an initial energy of 67.85, 104.25, and 134.68 MeV, respectively. Applying the LSTM dose calculation on patient cases that were fully excluded from the training process yields an average γ-index pass rate of 97.85%.
Conclusions: LSTM networks are well suited for proton dose calculation tasks. Further research, especially regarding model generalization and computational performance in comparison to established dose calculation methods, is warranted. © *2020 The Authors. Medical Physics published by Wiley Periodicals LLC on behalf of American Association of Physicists in Medicine.* [https://doi.org/10.1002/mp.14658]

Key words: deep learning, dose calculation, machine learning, proton therapy, radiation therapy, treatment planning

1. INTRODUCTION

The spatial calculation of the radiation dose within the patient's body is a central component of computer-aided treatment planning in the general radiotherapy chain. Thereby, accuracy is key — only a precise dose estimate enables a meaningful, patient-specific assessment of the treatment plan before the onset of therapy.[1–4]

At the same time, requirements regarding dose calculation speed keep rising.[5–7] This is most apparent in the context of real-time adaptive radiotherapy, which is pushed to become clinical reality by recent advances in image guidance technology and automatic segmentation tools.[8] Anatomical changes happening at an intrafractional time scale demand for dose calculation algorithms that update dose predictions concurrently.[9] This is especially relevant for proton therapy where anatomical changes have a more severe influence on the dose distribution than for photons. Moreover, massively repeated dose calculation for uncertainty quantification[10–14] (i.e., random dose samples and/or worst-case scenarios), and complex

simulations for biological effectiveness*[15,16] demand additional calculation load and therefore are still too time-consuming for widespread clinical application.

For particle therapy, the trade-off between dose calculation speed and accuracy is defined by pencil beam (PB) algorithms[17] on the one end and Monte Carlo (MC) algorithms on the other end. While pencil beam algorithms provide faster dose estimates, MC algorithms require a higher computational load.[18,19] At the same time, however, MC algorithms clearly outperform pencil beam algorithms regarding accuracy in complex geometries.[20–22]

In the field of machine learning, artificial neural network (ANN) models and their state-of-the-art deep learning models are currently making an impact at various stages in radiotherapy. This process is most notably in domains such as outcome prediction[23,24] and medical imaging for the purpose of image segmentation[25] and image reconstruction.[26,27] Academic studies investigating ANNs for dose calculation are limited, and they primarily investigate the feasibility of ANN models in photon therapy.[28–31] Furthermore, considerations are restricted on training a two-dimensional (2D)/three-dimensional (3D) model (e.g., U-Net[32]) for the final dose distribution of an optimized treatment plan — varying fluence modulations of the individual pencil beams are not considered. Consequently, such approaches cannot be applied for treatment planning where it is necessary to predict the dose for all contributing pencil beams individually during the optimization process.[33] Naturally, these limitations regarding fluence extend to the general beam setup, that is, the couch and gantry angle. To the best of our knowledge, only the approach by Liu et al.[34,35] focuses on the prediction of individual pencil beams, however, only considering a one-dimensional (1D) mapping of the activity distribution of the positron emitters to the 1D pencil beam dose distribution.

In this study, we introduce a novel dose calculation approach for proton therapy based on the application of long short-term memory (LSTM) networks,[36] in an attempt to mimic the physical characteristics of dose deposition for individual pencil beams. LSTM networks are an evolved version of the recurrent neural network (RNN)† class of ANNs. Unlike conventional feed-forward networks, they have a hidden inner state enabling efficient processing of sequences of data and effective propagation of information along the sequence.[37] Currently, LSTM networks are applied highly successful for time-series data, for example, stemming from speech or video.[38–41] We restrict this study to a minimal number of parameter dependencies, and establish an end-to-end model that predicts the dose distribution based on the input CT. Therefore, the 3D proton dose distribution of a pencil beam within the patient is understood as a sequence of 2D dose slices along the beam direction. This is the first work to exploit ANNs, and specifically LSTM networks, to perform 3D proton dose calculation.

2. MATERIALS AND METHODS

2.A. Problem parameterization

The elementary task underlying the dose calculation for an entire intensity-modulated proton therapy treatment is the calculation of the dose of a single proton pencil beam. Consequently, our study focuses on considerations for individual pencil beams. This reduction was chosen to study the fundamental characteristics of LSTM network-based dose calculations without averaging effects in treatment plans comprised of thousands of pencil beams which may conceal important aspects regarding the accuracy of the physical dose deposition.

Moreover, we consider MC simulations as the gold standard dose computation method and therefore as the *ground truth* for our learning problem. The MC dose calculations were carried out with the Topas (TOol for PArticle Simulation) wrapper[42] for Geant4,[43] and the PB dose calculation was performed by the open source matRad software toolkit[16].‡

In general, the 3D dose distribution of a single pencil beam within the patient body $\mathscr{D}$ is a function of the initial phase space (i.e., the initial position and momentum distribution) of the particles $\mathscr{P}$ and the 3D patient geometry $\mathscr{G}$.

$$\mathscr{D} = f(\mathscr{P}, \mathscr{G}) \tag{1}$$

In order to train a neural network for proton dose calculations, it is necessary to learn the mapping f from the 3D patient geometry $\mathscr{G}$ and the initial particle phase space $\mathscr{P}$ to the 3D dose distribution $\mathscr{D}$. To minimize the complexity of the training process for the neural network, we restrict the mapping to be learned for dose calculation to a single initial energy. This effectively reduces the space of possible dose calculation scenarios for the network and enables a denser sampling of the space of possible patient geometries and dose distributions. Additional energies can be addressed by separate learning problems.

The space of possible dose distributions can be further confined when switching from the patient coordinate system into the beam's eye view coordinate system. Here, the dose deposition is always oriented along the z'-axis, as shown in Fig. 1. The lateral extent and the particle range can be considered finite and is roughly known a priori for any given initial energy. Consequently, it is possible to perform a lateral and longitudinal clipping of the region of interest. In our case we use an isotropic resolution of 2 mm with $m = 15$ voxels in lateral direction and $l = 150$ voxels in longitudinal direction (for patient setup). With this parameterization, we deal with a supervised regression problem that maps the geometry input data $\mathscr{G}_i \in \mathbb{R}_+^{l \times m \times m}$ to real-valued dose output data $\mathscr{D}_i \in \mathbb{R}_+^{l \times m \times m}$.

Furthermore, we chose to perform the learning not on HU maps but on maps of the relative stopping power (RSP)§ which

*For ion therapy where the effective dose is not only dependent on the input patient geometry but also nonlinearly on other additional parameters, for example, on the linear energy transfer of the incoming particle beam and the tissue type of each voxel, among many others.

†The abbreviation RNN is used to refer to a class of ANN and also to the vanilla RNN architecture, interchangeably, across the paper.

‡http://www.matrad.org

§The RSP denotes the range loss in the geometry relative to a water phantom.

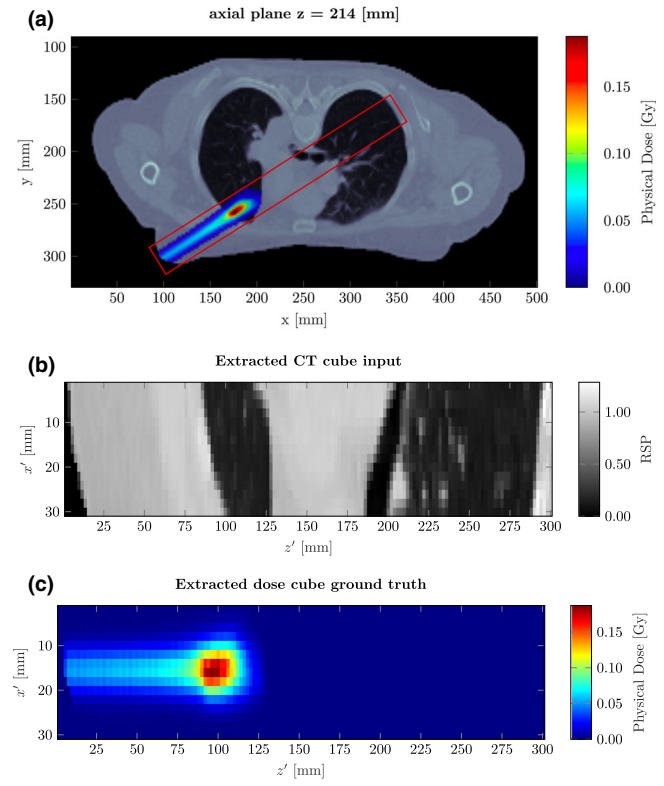

FIG 1. (a) Dose distribution of a single pencil beam with initial energy 104.25 MeV impinging from gantry angle 240° overlaying the patient computed tomography (CT). The clipping region is highlighted with a red box. (b) Respective CT slice and (c) dose distribution in beam's eye view coordinate system. [Color figure can be viewed at wileyonlinelibrary.com]

are also used for conventional pencil beam algorithms[16,20] for density scaling. We incorporate this conversion via HU look up tables yielding RSP values between 0 for vacuum and 2.5 for denser bone structures. The RSP values are in turn translated into the respective water density for MC simulations.

2.B. Model architecture

The particle dose calculation problem exhibits a geometrical peculiarity that motivates a more specialized network architecture: dose deposition is almost exclusively taking place in a sequential *upstream-to-downstream* manner. That is, the highly energetic protons predominantly travel along one direction with moderate lateral scatter until they stop. This characteristic behavior allows for a representation of the 3D input and output as a sequence of 2D slices, as illustrated in Fig. 2.

Consequently, the dose calculation problem has strong similarities to conventional video analysis in terms of spatio-temporal features. In action recognition tasks for instance, models have to extract spatial features of objects within each frame, and temporal features to interpret the movement of those objects. Simulating the protons traverse through matter, and consequently their dose deposition, is very similar to this task. It is completely determined by the upstream geometry, that is, the geometry previously "seen" by the protons along their track through the patient. This implies causality from upstream to downstream within the 3D volume and it suggests a special role for regions in the input data that have high gradients in their RSP values (e.g., material interfaces to

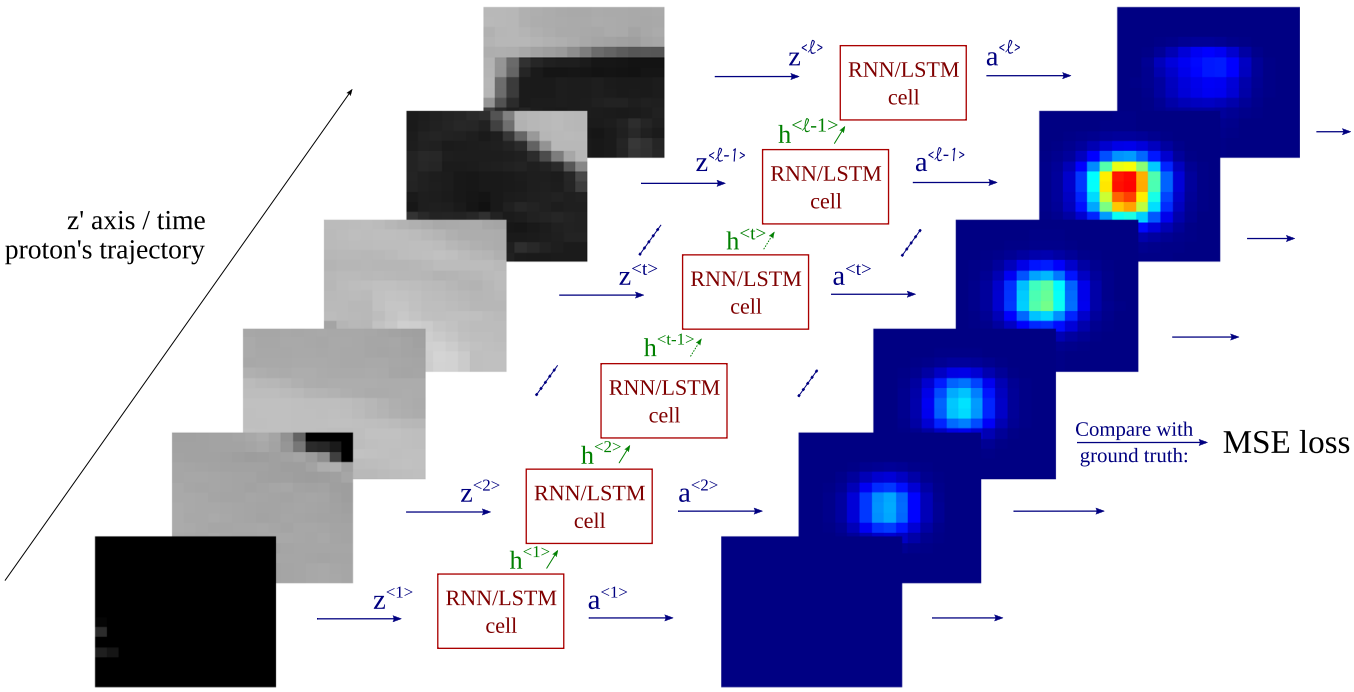

FIG 2. Sequential, spatio-temporal modeling of the proton dose calculation problem. Each $m \times m$ slice of the input is flattened into a 1D input array $z^{<t>}$. Each input array is then passed to the RNN/LSTM cell generating a hidden inner state $h^{<t>}$ and an output $a^{<t>}$. The hidden inner state is passed as an input information for subsequent slices (l slices in total), while the output is passed to a fully connected neural network back end to generate an $m \times m$ output slice. The output is then compared to the original ground truth by means of mean squared error loss. [Color figure can be viewed at wileyonlinelibrary.com]

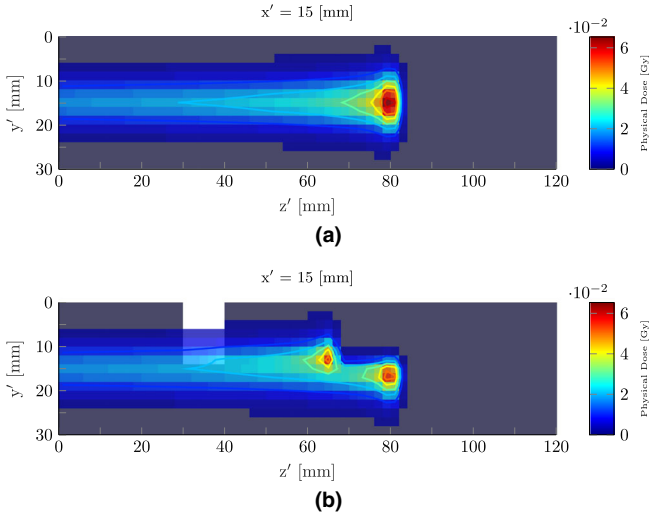

FIG 3. Effect of a heterogeneity on the shape of a pencil beam dose profile. The pencil beam is formed by 10^6 protons with an initial energy of 104.25 MeV passing through (a) water (1.0 RSP). (b) The cuboid heterogeneity (2.5 RSP) has 10 mm width in z' axis, 14 mm width in x' axis, and 1 mm distance to the center of the of proton beam. The effect of the cuboid mainly manifests in a bimodal Bragg peak region extending from ≈60 mm, that is, 20 mm after the heterogeneity. [Color figure can be viewed at wileyonlinelibrary.com]

bones with high RSP and cavities with low RSP). Thereby, the effect of each heterogeneity on the dose deposition is most pronounced at the end of the proton range as demonstrated in Fig. 3. Consequently, any model to simulate dose deposition for particles needs to extract spatio-temporal features and precisely propagate the impact of heterogeneities along the particle tracks.

Processing sequences with long dependencies requires a model capable of passing information through the series. RNNs with their hidden inner states, are capable of connecting many conventional one-input-to-one-output neural networks resulting in a model suitable to process many-input-to-many-output layouts. LSTM networks, an evolved version of simple RNNs, are capable of effectively transmitting relevant information through very long series thanks to their internal mechanism. Moreover, one directional LSTM models can fully adapt to the upstream-to-downstream propagation scheme of protons, eliminating dependencies between downstream to upstream resulting in a substantially reduced number of parameters for the model.

2.C. Model training

Many variants of the original LSTM[44–46] have been introduced so far, and this study is using the Pytorch[¶] implementation of this architecture. Training of the network was carried out with an Adam optimizer,[47] with a learning rate 10^{-5} and a mean squared error (MSE) loss function. The LSTM features one layer with 1000 neurons as internal layer, followed by a fully connected neural network for the back end. The back end network features one hidden layer with 100 neurons and an output layer with m^2 neurons to generate the slices. The activation function for the LSTM network and the fully connected backend network was chosen to be tanh and ReLU, respectively. The dose cubes were normalized to have values in 0 to 1 range, while we left the RSP input cubes of range 0 to 2.5 in tact. Empirically, we found no improvement in test loss after about 100 epochs, and after that overfitting of the training set has been observed. Training of the network takes 3 to 4 h for the patient dataset described in the next section, with a Geforce GTX 970 GPU.

[¶]https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html#torch.nn.LSTM

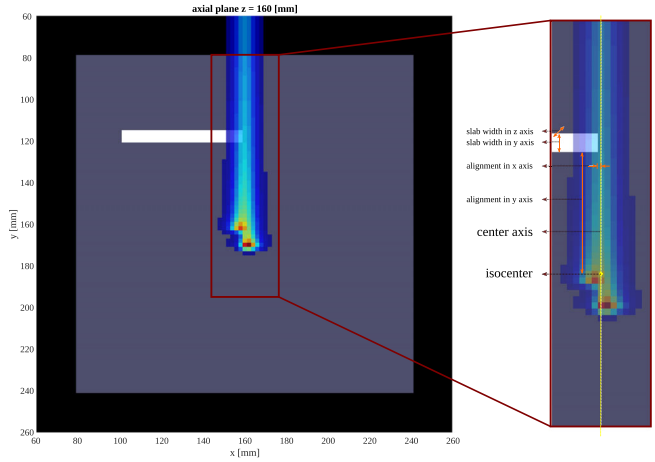

FIG 4. Phantom case setup; different geometric problems were generated by varying the slabs' dimensions in y and z axis, varying the alignment of the slab in x and y axis, and varying the density of both the water and the slab. [Color figure can be viewed at wileyonlinelibrary.com]

2.D. Data preparation and experiments design

2.D.1. Phantom cases

In order to study the performance of the proposed neural network dose calculation algorithms in an idealized setting, we first carried out simulations on phantom geometries featuring cuboid inhomogeneities of varying dimensions (2 to 14 mm in z' and x' axis) and densities (0.1 RSP to 2.5 RSP) placed randomly within a water phantom (0.8 RSP to 1.2 RSP), as shown in Fig. 4. For this task, 2500 phantom samples were generated with corresponding dose distributions from TOPAS MC simulations. As the physical problem of dose calculation exhibits a rotational symmetry around the beam axis, we can augment the dataset with CT and dose cubes that are rotated by 90° (to avoid interpolation and preserve the original MC simulations) resulting in a total of 10 000 training samples. Since neural networks are not invariant to input rotations, each of the added cubes would be considered as an unseen, informative training data that would effectively enhance the learning process.[48] From the 10 000 samples, 2000 samples were set aside as test set (Experiment I), 6000 samples were used as training set, and 2000 samples used as validation set (for hyper-parameter tuning). All the samples were simulated with $\sim 1.1 \times 10^6$ histories on

average, resulting in less than 1% statistical uncertainty. To normalize the dose cubes after MC simulation, we divide each cube by the corresponding fluence, and then divide the resulting cube by the integral dose[‖] of the respective cube. For reference comparison we include the dose distribution computed by a conventional pencil beam algorithm using the identical approach for normalization.

2.D.2. Patient cases

In order to study the performance of the proposed neural network dose calculation algorithms for real-world patient cases, we further considered dose calculation tasks on lung patient cases[ǀ] exhibiting highly pronounced inhomogeneities between normal tissue, lung tissue, and bony anatomy (rib cage & spine). For this task, 1000 lung case samples were generated with corresponding dose distributions from TOPAS MC simulations. All samples stem from the same patient. Different geometric problems could be extracted from one patient by sampling the beam orientation in 5° steps from 0° to 335° in combination with isocenter position samples in 10 mm shifts spanning the lung along the z axis, as shown in Fig. 5. All the samples were simulated with 2.5×10^6 histories on average, ensuring a statistical uncertainty between 1% and 2%. The total number of samples was raised to 4000 samples by augmenting rotated replicas (90° angles) of both the input and the output cubes, as described in the previous section. Once more, we choose to have a 60%–20%–20% split ratio scheme, leading to 2400 samples for training set, 800 samples for validation set, and 800 samples for test set (Experiment II). The original CT was downsampled to an isotropic 2 mm resolution.

In order to assess the generalization of the LSTM dose calculation engine to previously unseen patients, that is, data from other patients that were not considered during training, the performance of the network was evaluated on five additional lung cancer patients. For each patient, 200 pencil beams with randomly selected gantry angles and isocenter shifts were prepared, and the deposited dose was calculated using MC calculations (Experiment III).

For a meaningful application within a clinical treatment planning system where pencil beams of numerous energies are incorporated, it is crucial for the model to generalize to additional energies. Therefore, we prepared a dataset with three distinct energies, that is, low-range (68.33 MeV), mid-range (104.25 MeV), and high-range (134.22 MeV) proton pencil beams (Experiment IV). 3000 samples were randomly selected with an identical approach as performed in Section 2.D.2, allocating ∼1000 samples for each energy. For the high-range dataset, the cubes were extended to $l = 200$ voxels in longitudinal direction. The total number of samples was raised to 12 000 by means of data augmentation and with identical split ratio scheme as described before.

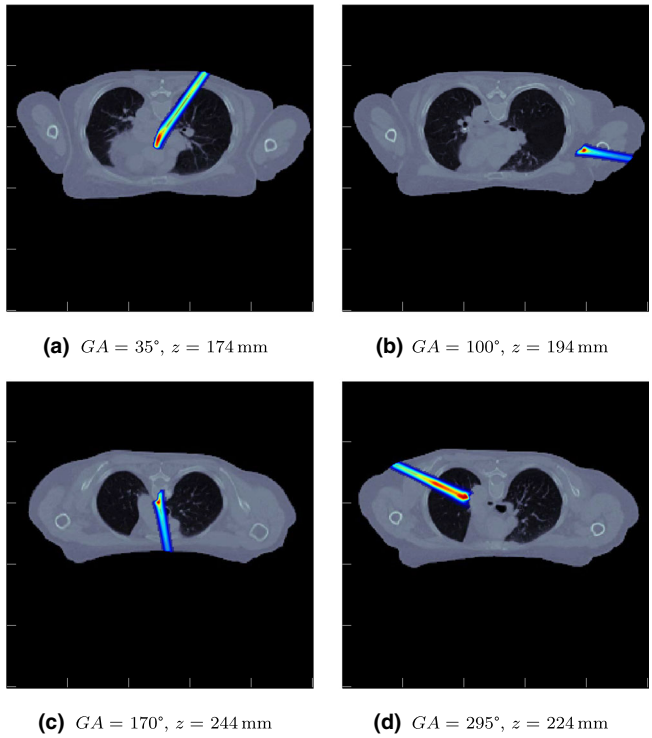

FIG 5. Lung case setup; generating different geometric problems for preparing training dataset by varying the gantry angles (GA) and shifting the isocenter along the z axis. [Color figure can be viewed at wileyonlinelibrary.com]

2.E. Metrics

In order to compare 3D dose distributions, a γ-analysis[49] was performed with a 1% dose difference and 3 mm distance-to-agreement criterion ([1%, 3 mm]) for both the phantom case and the patient case. Taking into consideration the MC statistical uncertainty of 1%, the chosen 1% dose difference criterion will ensure a strict comparison of the dose cubes. Any lower dose difference criterion would analyze the characteristics of MC simulation noise. Meanwhile, the 3 mm distance-to-agreement will ensure a neighborhood search of at least one voxel in each direction, given the 2 mm resolution of the dose cubes in both the phantom case and the patient case. We utilize gamma index distributions in our figures to illustrate the performance of estimated dose distributions in comparison to MC simulations, locally, in response to heterogeneities in the track of pencil beams. At the same time, we incorporate the γ-index pass rate, mean absolute error (MAE), and the MSE (defined over the entire range) to reduce the discrepancy of two 3D dose distributions to a single number which facilitates the large-scale comparisons needed for our study working with several thousand training, validation, and test samples.

[‖]The sum of all the voxels of the Hadamard product of the input cube and the respective dose cube.

[ǀ]The lung patients have been treated with photon IMRT at Heidelberg University Hospital. All patients consented to the anonymous use of their data for research purposes as part of their treatment agreement with Heidelberg University Hospital.

TABLE I. γ-index analysis ([1%, 3 mm]) comparing the two trained network models and a PB algorithm to the MC simulation in the phantom case (Experiment I).

	Mean (%)	SD (%)	Min (%)	Max (%)
RNN	97.57	1.38	91.17	99.31
LSTM	98.57	0.84	92.37	99.81
PB	97.83	0.86	88.53	98.94

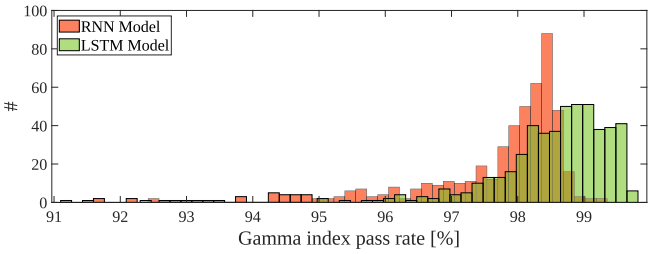

FIG 6. Comparison between the RNN and LSTM model γ-index pass rate distribution over all test cases in the phantom case (Experiment I). [Color figure can be viewed at wileyonlinelibrary.com]

3. RESULTS

3.A. Phantom cases

The prepared dataset for the water box phantom was used for training of the simple RNN and the LSTM network. The performance of the two networks was evaluated dosimetrically for the test set. Table I presents the outcome of the γ-analysis comparing the estimated dose from the two networks and a PB algorithm with the ground truth MC calculations.

While both networks seem generally suited for dose calculation with mean pass rates >97.57% (Fig. 6), the LSTM network outperforms the RNN by 1.0 percentage point. We have observed that differences between the LSTM network and RNN mainly originate from cases with pronounced heterogeneities as shown in Fig. 7. In this example, the LSTM model demonstrates an evident improvement in comparison to the RNN model, which fails to predict the bimodal Bragg peak behind the density interface resulting in approximately 1.3 percentage point increase in overall γ-index pass rate. Figure 7 additionally includes the PB algorithm performance and the corresponding γ-analysis on the representative test sample. In particular, we want to point out the distinct cut in the dose profiles behind interfaces,[20] in contrast to the bimodal Bragg peak in the ground truth MC and the estimated LSTM dose profile. This behavior is further illustrated by plotting the 1D profile of the center voxel array of the representative test sample cube along the z' axis in Fig. 8.

Since all the methods result in a comparable mean γ-index pass rate, we report the MAE and the MSE between the

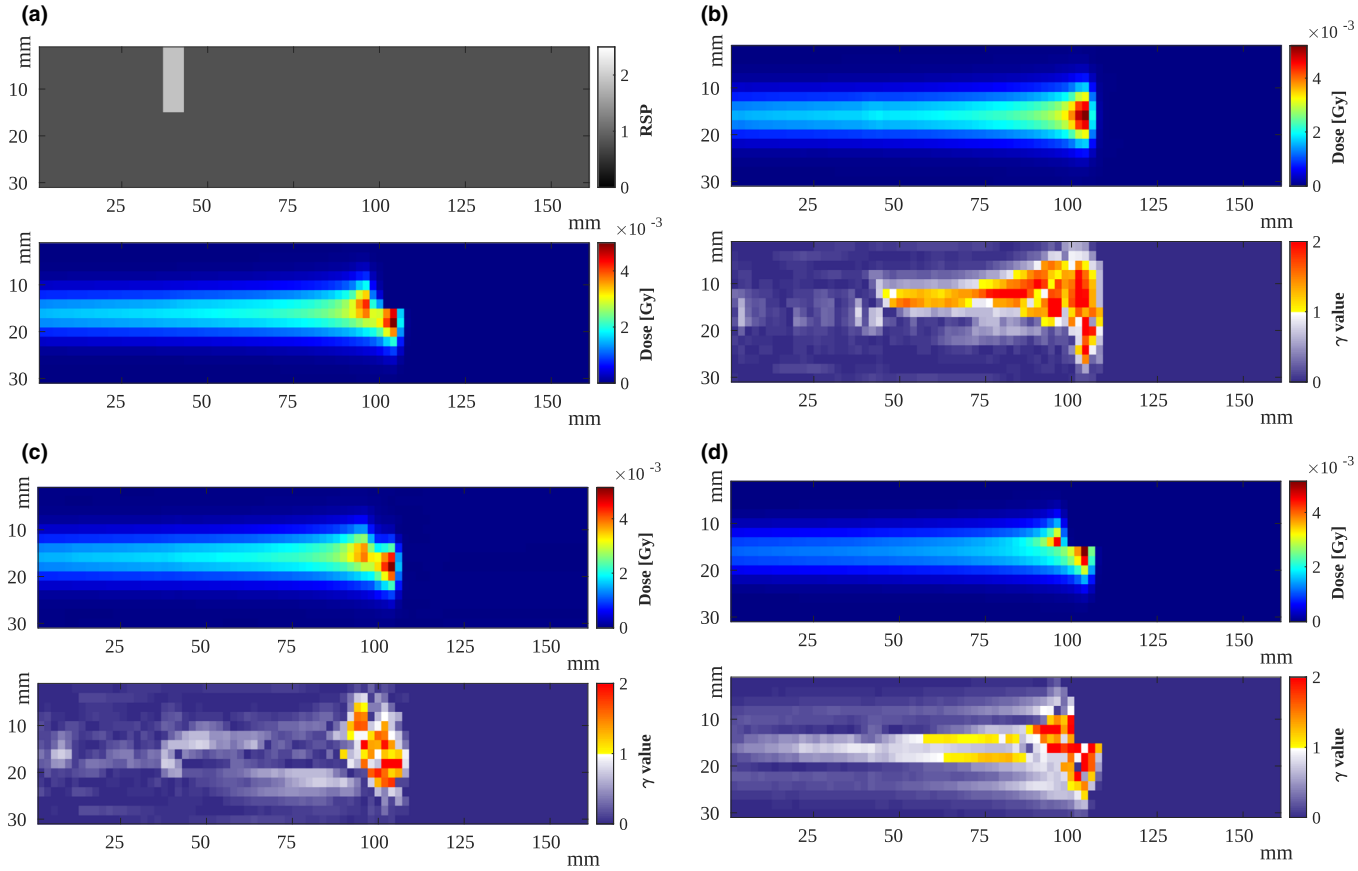

FIG 7. Performance comparison of (b) the RNN network, (c) the LSTM network, and (d) the PB algorithm with (a) ground truth MC calculation for a representative sample (104.25 MeV, with a 6 mm width slab and 1.9 RSP, γ-analysis criterion = [1%, 3 mm]). [Color figure can be viewed at wileyonlinelibrary.com]

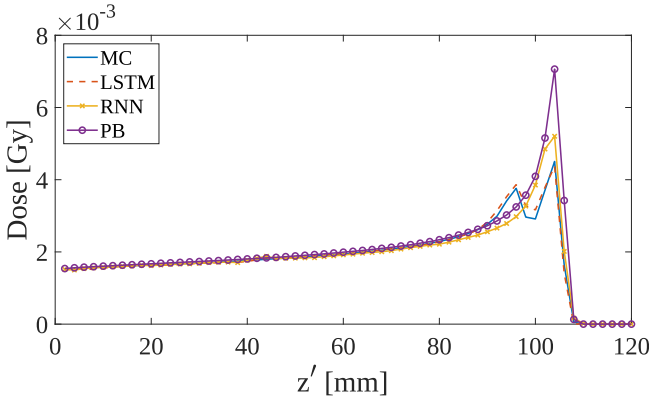

FIG 8. 1D dose profile of the center voxel array for all the models performed on the test sample corresponding to Fig. 7. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE II. MAE and MSE between the network models and the MC simulation in the phantom case (Experiment I).

	LSTM	RNN	PB
MAE (Gy)	3.3×10^{-3}	6.1×10^{-3}	3.8×10^{-3}
MSE (Gy^2)	4.4×10^{-4}	1.6×10^{-3}	6.7×10^{-4}

TABLE III. γ-index analysis ([1%, 3 mm]), MAE, and MSE of the LSTM model and PB algorithm compared to MC calculations for the lung patient case (Experiment II).

γ-analysis	Mean (%)	SD (%)	Min (%)	Max (%)
LSTM	98.50	1.00	93.93	99.82
PB	99.15	1.26	92.16	99.93
	MAE (Gy)		MSE (Gy^2)	
LSTM	6.9×10^{-3}		6.8×10^{-4}	
PB	4.7×10^{-3}		1.5×10^{-3}	

generated dose cubes and the ground truth MC simulations for the entire test set in Table II.

3.B. Patient cases

We further trained the LSTM network on the lung patient dataset. Table III summarizes the outcome of the comparison with ground truth MC simulation (γ-analysis, MAE, and MSE) for the set-aside test set.

Figure 9 shows the performance of the trained network and the PB algorithm on a representative test sample. In particular, we want to point out the capability of the trained network to deal with oblique gantry angles where voxels with vanishing density prior to entering the patient are successfully recognized and not confused with low-density lung voxels lying within the patient. Furthermore, the LSTM network correctly predicts a smeared out Bragg peak without a distinct maximum at the end of the particles' range which is due to low-density lung tissue at the location of the Bragg peak. Also the irregular shape of the distal fall-off which originates from inhomogeneities in the pencil beam track is qualitatively predicted by the network dose calculation algorithm. Additional samples are showcased in Appendix.

In order to localize the gamma index criteria violations, we divide each pencil beam in four equidistant quarter regions in their longitudinal range, and count the number of failed voxels in each region. We report the distribution of the failing voxels depending on their location along the longitudinal axis in percentage for the entire test set in Fig. 10.

Approximately 70% of the failing voxels for the PB algorithm occur in the fourth quarter where the Bragg peak is located. This failure to predict complex, potentially bimodal, Bragg shapes behind density interfaces is characteristic for PB algorithms. For the LSTM network, only ∼15% of the failed voxels fall in the Bragg peak region, and failed voxels are relatively distributed equally along the protons range.

Furthermore, we report the result of the two designed experiments (outlined in Section 2.D.2) to examine the generalization of the trained network to (a) other unseen patients and (b) other distinct energies. Table IV reports the result of incorporating the trained network to carry out dose estimation for the unseen patients dataset. Table V summarized the results for three separate networks that have been trained for 67.85, 104.25, 134.68 MeV initial energy. Lastly, we also showcase two representative challenging samples from the low-range and high-range dataset in Fig. 11.

3.C. Run-times

Table VI lists the average run-times for estimating the dose for the five above-mentioned patients for a single pencil beam, for MC, PB, and LSTM dose calculation. The MC simulations were performed with Topas on a calculation node with 28 virtual CPUs on an Openstack[**] cluster. For the trained network, the run times were measured for two systems with different GPUs. Depending on the facilitated hardware, we measure average run-times of 6 to 23 ms for the LSTM approach. Note that these run-times included the time required to send the input CT cube for each pencil beam from CPU to GPU and vice versa for the yielded dose cube. However, in applications such as adaptive radiotherapy which requires repetitive online dose estimations, the input CT cubes can be prepared and sent to the GPU in advance. Consequently, the only relevant run-times would be the network feed forward, that is, matrix multiplication operations run-times, reported to be 1.5 to 2.5 ms for the two facilitated hardware stacks. The average Topas run-time was 1160 s, performed with 2.5×10^6 histories on average.

4. DISCUSSION

In this paper, we have demonstrated the general feasibility of proton dose calculation based on an LSTM neural

[**]https://www.openstack.org/

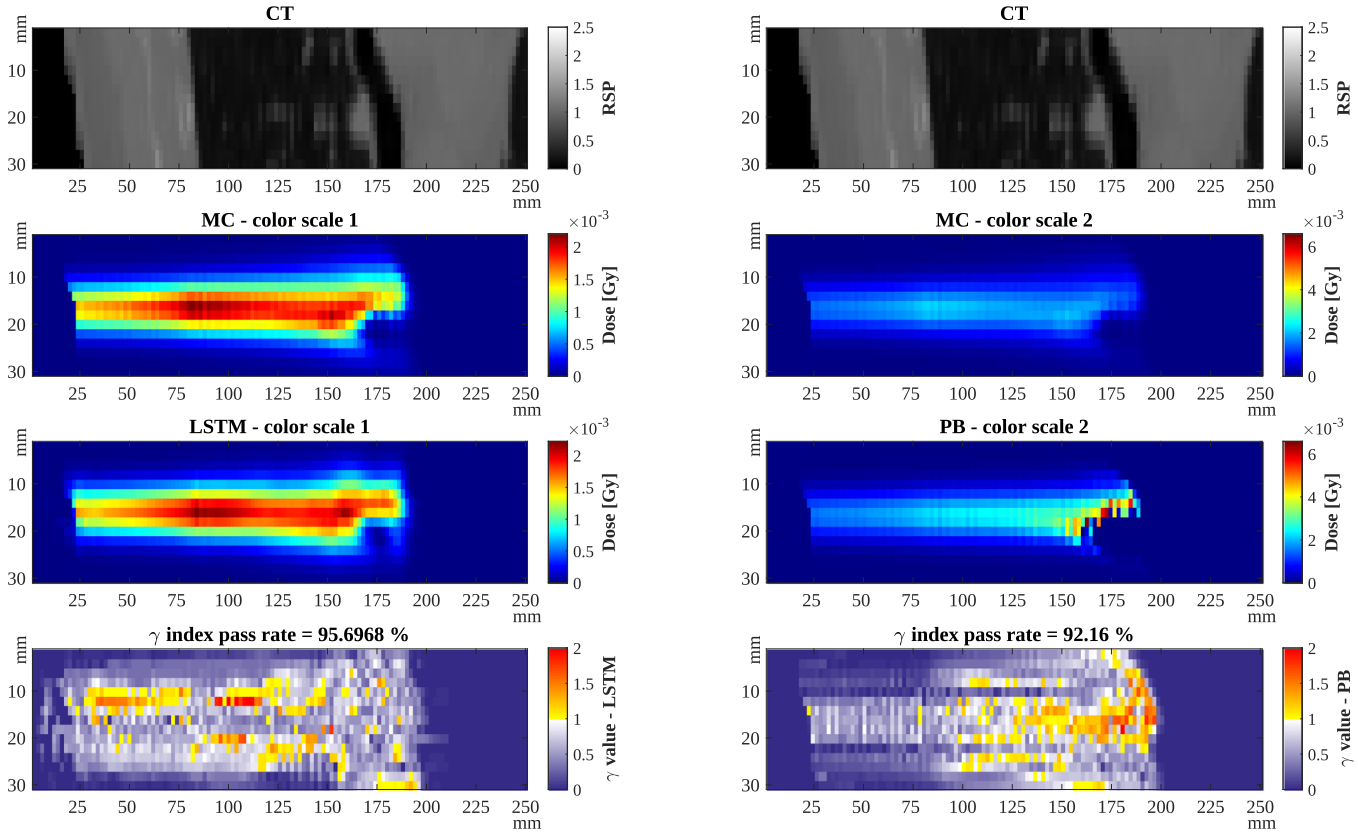

FIG 9. Dose estimation result for a sample test data (104.25 MeV) for LSTM network (left) and PB algorithm (right). Starting from top is the input patient CT, the ground truth MC dose distribution, the estimated dose by the LSTM network (left) and PB algorithm (right), and the corresponding γ-index map ([1%, 3 mm]). Due to the apparent difference in the range of the estimated dose values, the reference MC is plotted in two distinct color scale, depending on the dose range of the compared cube. The cubes are clipped in longitudinal axis to show 125 voxels (250 mm) instead of the original 150 voxels (300 mm), for better visualization. [Color figure can be viewed at wileyonlinelibrary.com]

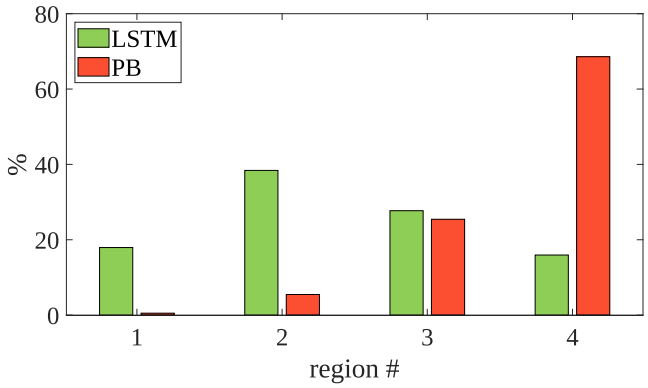

FIG 10. Distribution of the voxels which failed the γ-analysis criteria according to their location in longitudinal axes. Regions are equidistant quarters along the range of the pencil beam. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE IV. γ-index analysis on five different lung cancer patients ([1%, 3 mm]). The network has been trained on patient 0 (Experiment III). Patients 4 and 5 have very low RSP values in lung (further discussed in Section 4).

	Mean (%)	SD (%)	Min (%)	Max (%)
Patient 0	98.50	1.00	93.93	99.82
Patient 1	98.27	0.97	94.66	99.65
Patient 2	98.35	1.30	94.35	99.78
Patient 3	98.45	1.10	94.51	99.60
Patient 4	96.71	3.01	81.66	99.61
Patient 5	97.47	1.87	87.82	99.61

network. The LSTM network correctly models the proton dose deposition characteristics in the entrance, in the Bragg peak, and in the distal fall-off region — also in heterogeneous geometries.

In comparison to RNN networks, LSTM networks proved particularly suited for this task, especially in heterogeneous geometries. Using phantom and lung patient cases, we have observed very good agreement for individual pencil beams with an initial energy of 104.25 MeV at run-times of 6 to 23 ms per pencil beam. In our phantom case study that specifically focused on challenging heterogeneities, the LSTM network shows an improved performance in comparison to the conventional PB algorithm, in every incorporated metric. This is also evident in Fig. 7 where the PB algorithm fails to predict a smooth bi-modal Bragg peak behind the interface.

In the patient case study, however, we experience a slight set back in mean γ-index pass rate in comparison to the PB algorithm. While having a lower training set size can be one cause of such set back, the other reason can also be the

TABLE V. γ-index analysis on datasets with three distinct energies proton pencil beams.

Energy (MeV)	Mean (%)	SD (%)	Min (%)	Max (%)
67.85	98.56	1.30	95.35	99.79
104.25[a]	97.74	1.48	92.57	99.74
134.68	94.51	2.99	85.37	99.02

[a]The results for this energy are different to what reported earlier since this dataset is generated again for this analysis and the samples are randomly chosen.

TABLE VI. Run-time comparison of the MC calculation vs LSTM predictions and PB algorithm. Run times reported in parenthesis consider purely the network feed forward time consumption and do not count the time required to send each input/output from CPU to GPU and vice versa.

	MC[a]	LSTM[b]	LSTM[c]	PB[d]
Average run time (s)	1159.5	0.023 (0.0025)	0.006 (0.0015)	1.025

[a]Computational node, 28 VCPUs, 64 Gb RAM.
[b]Intel Core i7-6700 3.4 GHz - Nvidia GTX 970 - 64 Gb RAM.
[c]Intel Xeon W-2135 3.7 GHz - Nvidia Quadro RTX 6000 - 64 Gb RAM.
[d]matRad software toolkit: Intel Xeon W-2135 3.7 GHz - 64 Gb RAM.

interpolation effects experienced in the process of extracting cubes in the desired format[††] (Fig. 1). Figure 10 illustrates this effect by detailing where (relative to the depth dose curve) voxels fail the gamma criterion for the different dose calculation approaches, with approximately 85 percent located in the entrance region for the dose estimated by the LSTM network. Discrepancies in the entrance region may ultimately affect the final treatment plan less in comparison to the Bragg peak region where pencil beams of different energies and gantry angles accumulate to construct the spread out Bragg peak in tumor regions. Moreover, we observed that the PB algorithm substantially overestimates the peak dose (see Figs. 8 and 9). However, the γ-index pass rate is blind regarding the extent with which a voxel fails the γ-criterion. These effects are reliably detected by the MSE which considers larger deviations with a higher (quadratic) weighting. Table III further supports this argument where for the LSTM network, MSE reported to be ~50% lower while MAE is reported to be ~30% higher in relative difference with PB algorithm.

Based on the approach to study dose calculation accuracy for an individual energy, we were able to show the generalization of our algorithm to patient cases that were not considered during LSTM training. While the γ-index pass rates for patients 1 to 3 was >98%, the γ-index pass rates for patients 4 and 5 ranged between 96% and 97%. This slight decline was attributed to very low RSP values in lung which could not be discriminated against air volumes penetrated by the beam before entering the patient. This phenomenon originated from beam angles in the training set where the beam enters and exits the patient arms before impinging on the chest [see Figs. 12(a) and 15(d)]. While this issue could be easily avoided by including only clinically feasible beam orientations during training, we decided to have them included as challenging test scenario for the network.

In order to implement dose calculation for an entire treatment plan, however, additional networks need to be trained for different energies. Alternatively, and conceptually more appealing, it may be possible to train a network that is able to generalize also over different initial energies. In this study, we were able to verify the generalization of the network for additional energies using the former method of independent training. The energies were chosen to represent three proton ranges from low-range to high-range energies. We experienced deteriorated result for the high-range energy datasets in comparison with the other two dataset, primarily because of two different scenarios. (a) Pronounced air cavities led to overshoot cases where the Bragg peak was located behind (outside) the patient. (b) The Bragg peak reached the second lung of the patient after completely penetrating the first lung. Again, we argue that both scenarios are clinically not relevant and may not occur when improving the selection of the geometries based on clinical considerations. In order to provide a complete characterization of the proposed method within the paper, however, we explicitly decided to have these edge cases included in the analysis.

The run-times given in Section 3.C. for the networks, the PB algorithm, and the MC simulations have to be interpreted with care. The pencil beam algorithm used by matRad, for example, relies on a Matlab implementation that does not use explicit parallelism. We are aware that substantially faster implementations exist in clinical practice and in research.[50] Also, the run time benefits of several orders of magnitude over MC simulations as shown in Table VI will not manifest in the same way for clinical treatment plans comprised of several thousands of pencil beams. Here, MC simulations can save substantially because the geometry will only be initialized once for the entire simulation. Furthermore, it will be possible to reduce the number of histories per pencil beam to achieve sufficient statistical certainty over the entire treatment plan for a simple dose recalculation. For the computation of a dose influence matrix which is needed for dose optimization, however, the MC run-time reductions will be more moderate. Such a speedup over MC dose calculation at comparable accuracy, is key for online re-planning and motivates further work in this direction. Efforts regarding run-time optimization are further supported through constant advances in dedicated deep-learning hardware, and more prominently by leveraging the embarrassingly parallel nature of the problem. Finally, as previously indicated, the transfer times between CPU and GPU will only be necessary once per patient for LSTM network dose calculations. In our case this made up 75% of the run time for the faster GPU hardware.

Our study concentrated on the ability to estimate dose in heterogeneous geometries, and no effort was made in improving the model efficiency. Various model compression techniques, for example, pruning, quantization, and tensor decomposition methods (achieving low-rank structures in the weight matrices),[51–53] may substantially lower the number of

[††]This effect does not occur in phantom study since the impinging pencil beams only enter the water phantom in nonoblique angles.

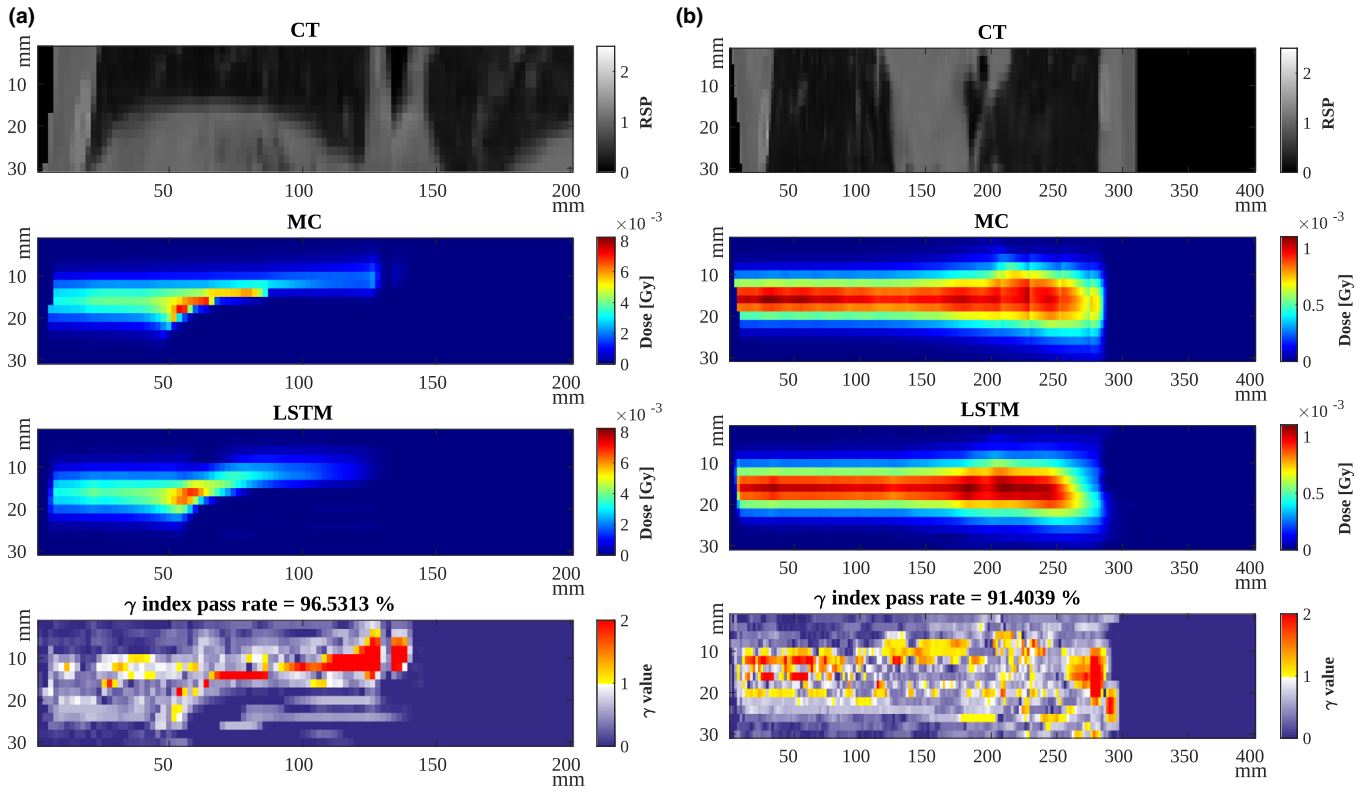

FIG 11. Dose estimation results for two representative samples from (a) the low-range (67.85 MeV) dataset and (b) high-range (134.68 MeV) dataset. The sub-figures follow the layout outlined in Fig. 9. Note that the longitudinal axis range is different in each sub-figure. [Color figure can be viewed at wileyonlinelibrary.com]

parameters in fully connected layers.[54,55] The efficiency of the model can be further enhanced through fine-tuning of the model architecture. This study parameterized the size in longitudinal direction as a fixed hyper-parameter (parameter *l*, see Section 2.A.). While the range of mono-energetic protons is more or less fixed in a homogeneous geometry, it can vary substantially when they travel through wide cavities such as the lung. This issue coerced us to train the model with very long sequences, to encompass all the potential pencil beam ranges. However, the LSTM models can be designed in what is referred to as *sequence to sequence learning*, which can accept a variable length input and outcome with a variable length output, incorporated effectively in machine translation problems.[34,56] Utilization of such a model can restrict the number of matrix multiplication operations accustomed to the plan, resulting in even faster estimations. In a different approach, one could also incorporate autoencoders[57] as a front-end to the model, compressing the input CT to a latent feature space, leading to a reduction in number of input parameters.

We intend to explore this approach in many aspects in future studies. We see possible applications in photon dose calculation as well as in heavier ions (Carbon, Oxygen, Helium) dose calculation in an attempt to facilitate uncertainty quantification (and therefore robust and 4D planning in anatomies like lung) and direct biological dose predictions.

5. CONCLUSION

In this paper, we have investigated the role of two different neural network architectures for proton dose calculation, that is, an RNN and an LSTM network. For individual pencil beams on varying heterogeneous phantom geometries, the average γ-index pass rate ([1%, 3 mm]) was 97.6 % for the RNN and 98.6% for the LSTM network compared to MC reference simulations. The LSTM network was further evaluated on a highly heterogeneous lung case where we observed an average γ-index pass rate of 98.5% ([1%, 3 mm]). Average LSTM network run-times ranged between 6 to 23 ms. The generalization of the model was further verified by testing for five unseen patients and three distinct proton energies, achieving >94.5% γ-index pass rates.

Our results indicate that LSTM networks are well suited for particle therapy dose calculation tasks. Further research, especially regarding model generalization and computational performance in comparison to established dose calculation methods, is warranted.

ACKNOWLEDGMENTS

The authors thank Dr. Lucas Burigo at the German Cancer Research Center for providing the TOPAS MC interface for matRad. Open Access funding enabled and organized by ProjektDEAL.

CONFLICT OF INTEREST

The authors have no relevant conflict of interest to disclose.

[a)]Author to whom correspondence should be addressed. Electronic mail: a.neishabouri@dkfz-heidelberg.de.

REFERENCES

1. Newhauser W, Koch N, Hummel S, Ziegler M, Titt U. Monte Carlo simulations of a nozzle for the treatment of ocular tumours with high-energy proton beams. *Phys Med Biol.* 2005;50:5229–5249.
2. Newhauser W, Fontenot J, Zheng Y, et al. Monte Carlo simulations for configuring and testing an analytical proton dose calculation algorithm. *Phys Med Biol.* 2007;52:4569–4584.
3. Bauer J, Sommerer F, Mairani A, et al. Integration and evaluation of automated Monte Carlo simulations in the clinical practice of scanned proton and carbon ion beam therapy. *Phys Med Biol.* 2014;59: 4635–4659.
4. Mein S, Kopp B, Tessonnier T, et al. Dosimetric validation of Monte Carlo and analytical dose engines with raster-scanning 1H, 4He, 12C, and 16O ion-beams using an anthropomorphic phantom. *Phys Med.* 2019;64:123–131.
5. Mein S, Choi K, Kopp B, et al. Fast robust dose calculation on GPU for high-precision 1H, 4He, 12C and 16O ion therapy: the FRoG platform. *Sci Rep*. 2018;8:1–12.
6. Jia X, Schümann J, Paganetti H, Jiang SB. GPU-based fast Monte Carlo dose calculation for proton therapy. *Physics in Medicine and Biology.* 2012;57:7783–7797. https://doi.org/10.1088/0031-9155/57/23/7783
7. Wang Y, Mazur TR, Park JC, Yang D, Mutic S, Li H. Development of a fast Monte Carlo dose calculation system for online adaptive radiation therapy quality assurance. *Phys Med Biol Inst Phys Eng Med*. 2017;62:4970–4990.
8. Boldrini L, Bibault J-E, Masciocchi C, Shen Y, Bittner M-I. Deep learning: a review for the radiation oncologist. *Front Oncol*. 2019;9:977.
9. Boldrini L, Cusumano D, Cellini F, Azario L, Mattiucci GC, Valentini V. Online adaptive magnetic resonance guided radiotherapy for pancreatic cancer: state of the art, pearls and pitfalls. *Radiat Oncol*. 2019;14:71.
10. Unkelbach J, Bortfeld T, Martin BC, Soukup M. Reducing the sensitivity of IMPT treatment plans to setup errors and range uncertainties via probabilistic treatment planning. *Med Phys.* 2008;36:149–163.
11. Kraan AC, van de Water S, Teguh DN, et al. Dose uncertainties in IMPT for oropharyngeal cancer in the presence of anatomical, range, and setup errors. *Int J Radiat Oncol Biol Phys*. 2013;87:888–896.
12. Park PC, Cheung JP, Zhu XR, et al. Statistical assessment of proton treatment plans under setup and range uncertainties. *Intl J Radiat Oncol Biol Phys*. 2013;86:1007–1013.
13. Bangert M, Hennig P, Oelfke U. Analytical probabilistic modeling for radiation therapy treatment planning. *Phys Med Biol.* 2013;58: 5401–5419.
14. Wahl N, Hennig P, Wieser HP, Bangert M. Efficiency of analytical and sampling-based uncertainty propagation in intensity-modulated proton therapy. *Phys Med Biol.* 2017;62:5790–5807.
15. Mairani A, Böhlen TT, Schiavi A, et al. A Monte Carlo-based treatment planning tool for proton therapy. *Phys Med Biol.* 2013;58:2471–2490.
16. Wieser HP, Hennig P, Wahl N, Bangert M. Analytical probabilistic modeling of {RBE}-weighted dose for ion therapy. *Phys Med Biol.* 2017;62:8959–8982.
17. Hong L, Goitein M, Bucciolini M, et al. A pencil beam algorithm for proton dose calculations. *Phys Med Biol.* 1996;41:1305–1330.
18. Fippel M, Soukup M. A Monte Carlo dose calculation algorithm for proton therapy. *Med Phys.* 2004;31:2263–2273.
19. Szymanowski H, Oelfke U. Two-dimensional pencil beam scaling: an improved proton dose algorithm for heterogeneous media. Technical Report; 2002.
20. Schaffner B, Pedroni E, Lomax A. Dose calculation models for proton treatment planning using a dynamic beam delivery system: an attempt to include density heterogeneity effects in the analytical dose calculation, Technical Report; 1999.
21. Soukup M, Fippel M, Alber M. A pencil beam algorithm for intensity modulated proton therapy derived from Monte Carlo simulations. *Phys Med Biol.* 2005;50:5089–5104.
22. Taylor PA, Kry SF, Followill DS. Pencil beam algorithms are unsuitable for proton dose calculations in lung. *Int J Radiat Oncol Biol Phys.* 2017;99:750–756.
23. Gulliford SL, Webb S, Rowbottom CG, Corne DW, Dearnaley DP. Use of artificial neural networks to predict biological outcomes for patients receiving radical radiotherapy of the prostate. *Radiother Oncol.* 2004;71:3–12.
24. Gabryś HS, Buettner F, Sterzing F, Hauswald H, Bangert M. Design and selection of machine learning methods using radiomics and dosiomics for normal tissue complication probability modeling of Xerostomia. *Front Oncol*. 2018;8:35.
25. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys.* 2017;44:547–557.
26. Chen S, Qin A, Zhou D, Yan D. U-net-generated synthetic CT images for magnetic resonance imaging-only prostate intensity-modulated radiation therapy treatment planning. *Med Phys.* 2018;45:5659–5665.
27. Nie D, Cao X, Gao Y, Wang L, Shen D. Estimating CT image from MRI data using 3D fully convolutional networks. In: *Deep Learning and Data Labeling for Medical Applications*. Berlin: Springer; 2016:170–178.
28. Nguyen D, Long T, Jia X, et al. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci Rep*. 2019;9:1–10.
29. Kontaxis C, Bol GH, Lagendijk JJW, Raaymakers BW. DeepDose: towards a fast dose calculation engine for radiation therapy using deep learning. *Phys Med Biol.* 2020;65:75013.
30. Mahmood R, Babier A, McNiven A, Diamant A, Chan TCY. Automated treatment planning in radiation therapy using generative adversarial networks. In: *Proceedings of the 3rd Machine Learning for Healthcare Conference*, in PMLR 85; 2018:484–499.
31. Kearney V, Chan JW, Haaf S, Descovich M, Solberg TD. DoseNet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks. *Phys Med Biol.* 2018;63:235022.
32. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing; 2015:234–241.
33. Wu C, Nguyen D, Xing Y, et al. Improving proton dose calculation accuracy by using deep learning. Technical Report.
34. Liu C, Li Z, Hu W, Xing L, Peng H. Range and dose verification in proton therapy using proton-induced positron emitters and recurrent neural networks (RNNs). *Phys Med Biol.* 2019;64:175009.
35. Hu Z, Li G, Zhang X, Ye K, Lu J, Peng H. A machine learning framework with anatomical prior for online dose verification using positron emitters and PET in proton therapy. *Phys Med Biol.* 2020;65:185003.
36. Hochreiter S, Urgen Schmidhuber JJ. Long short-term memory. Technical Report 8; 1997.
37. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14. Cambridge, MA, USA: MIT Press; 2014:3104–3112.
38. Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14; 2014:II–1764–II–1772, JMLR.org.
39. Donahue J, Hendricks LA, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description. Technical Report.
40. Ng JY-H, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short dnippets: deep networks for video classification; 2015.
41. Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset; 2017.
42. Perl J, Shin J, Schümann J, Faddegon B, Paganetti H. TOPAS: an innovative proton Monte Carlo platform for research and clinical applications. *Med Phys.* 2012;39:6818–6837.

43. Agostinelli S, et al. GEANT4: a simulation toolkit. *Nucl Instrum Meth A*. 2003;506:250–303.
44. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw*. 2005;18:602–610.
45. Gers FA, Schmidhuber JA, Cummins FA. Learning to forget: continual prediction with LSTM. *Neural Comput.* 2000;12:2451–2471.
46. Gers FA, Schmidhuber J. Recurrent nets that time and count. In: *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium 3, Vol. 3; 2000:189–194.
47. Kingma DP, Ba J. Adam: a method for stochastic optimization; 2014.
48. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6:60.
49. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. Technical Report; 1998.
50. Jia X, Ziegenhein P, Jiang SB. GPU-based high-performance computing for radiation therapy. *Phys Med Biol.* 2014;59:R151–R182.
51. Grachev AM, Ignatov DI, Savchenko AV. Compression of recurrent neural networks for efficient language modeling. *Appl Soft Comput*. 2019;79:354–362.
52. Yang Y, Krompass D, Tresp V. Tensor-train recurrent neural networks for video classification. In: Precup D, Teh YW, eds. *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia; 2017:3891–3900, PMLR.
53. Ye J, Wang L, Li G, et al. Learning compact recurrent neural networks with block-term tensor decomposition; 2017.
54. Yang Z, Moczulski M, Denil M, et al. Deep fried convnets. In: *The IEEE International Conference on Computer Vision (ICCV)*; 2015.
55. Han S, Mao H, Dally WJ. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding; 2015.
56. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, Vol. 27. New York, NY: Curran Associates, Inc.; 2014:3104–3112.
57. Baldi P. Autoencoders, unsupervised learning, and deep architectures. In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*; 2012:37–49.

APPENDIX

In this section, we illustrate the performance of the trained network on additional challenging test samples, in which the inputs exhibit noticeable heterogeneities. Figures 12 and 13 showcase samples from the test dataset of the original trained patient (patient 0), while Figs. 14 and 15 showcase the result of incorporating the trained network on other unseen patients (patients 3 and 5, respectively).

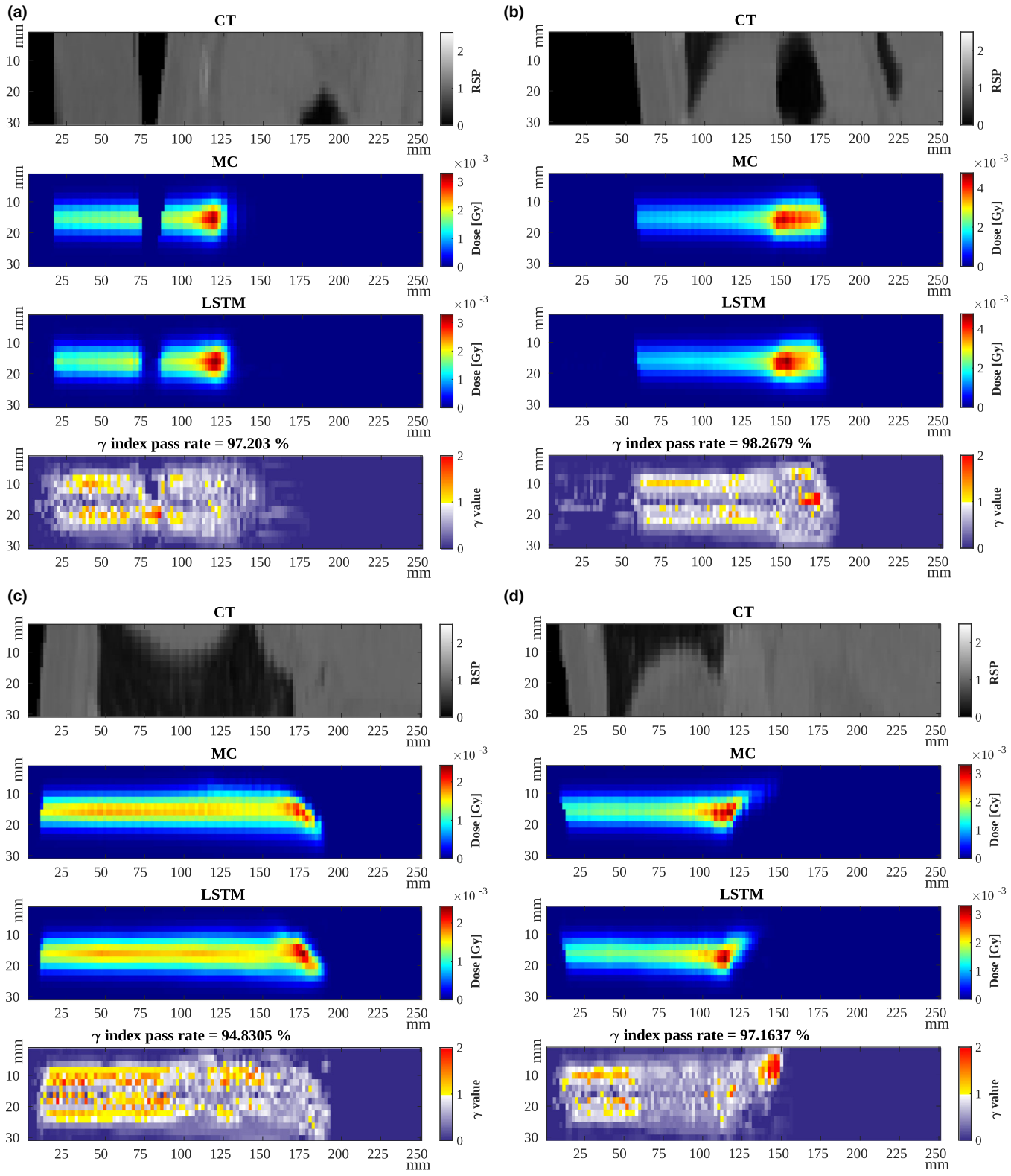

FIG 12. Dose estimation results for four test data from **patient 0** ($E = 104.25$ MeV, γ-index criteria $= [1\%, 3$ mm]). The sub-figures follow the layout outlined in Fig. 9. Note that sample (a) has an air gap between the patient's arm and chest (discussed in Section 4). [Color figure can be viewed at wileyonlinelibrary.com]

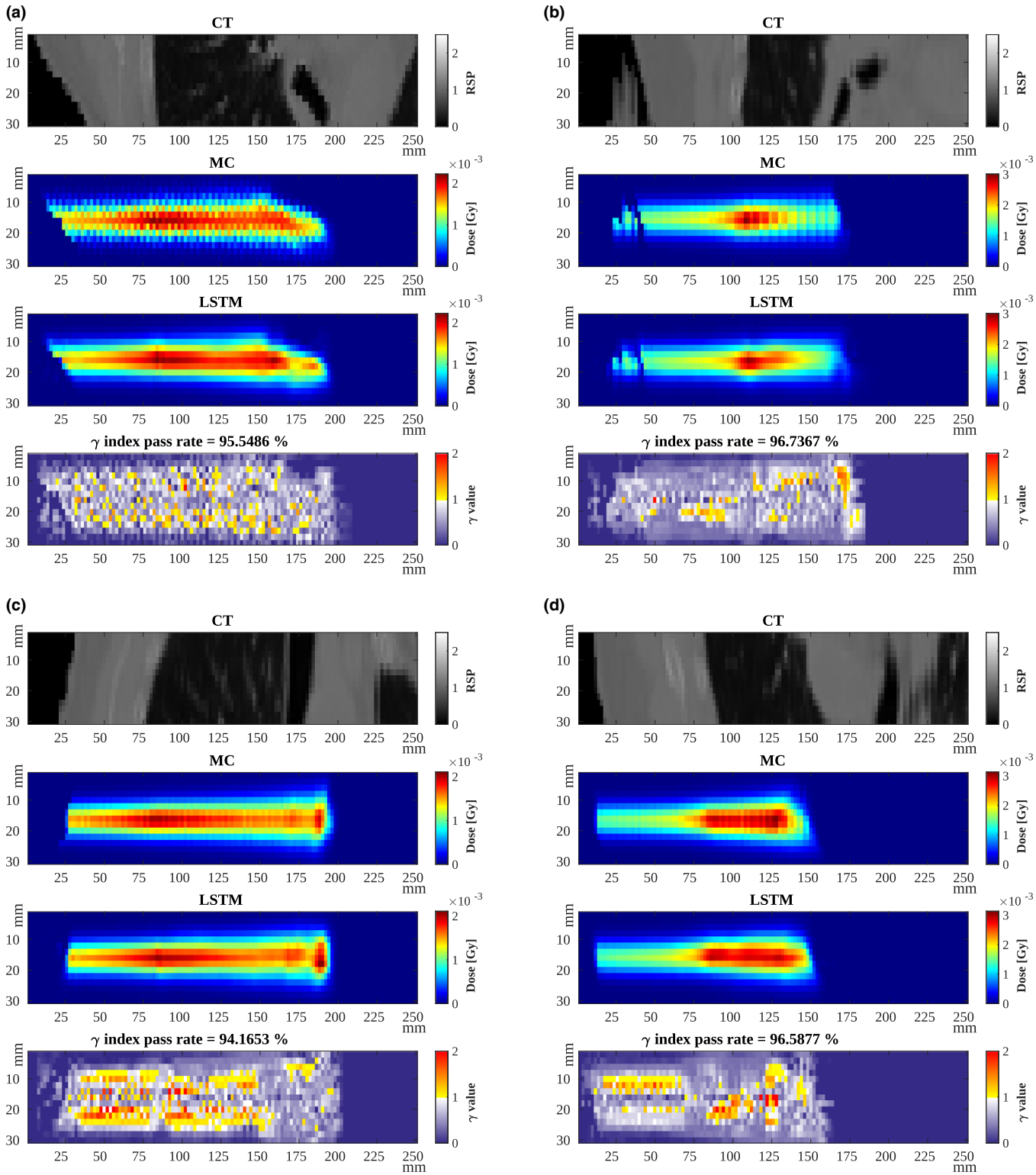

FIG 13. Dose estimation results for four test data from **patient 0** ($E = 104.25$ MeV, γ-index criteria $=$ [1%, 3 mm]). The sub-figures follow the layout outlined in Fig. 9. Note that the aliasing effect in sample (a) is due to the cube extraction interpolation of oblique gantry angles. [Color figure can be viewed at wileyonline library.com]

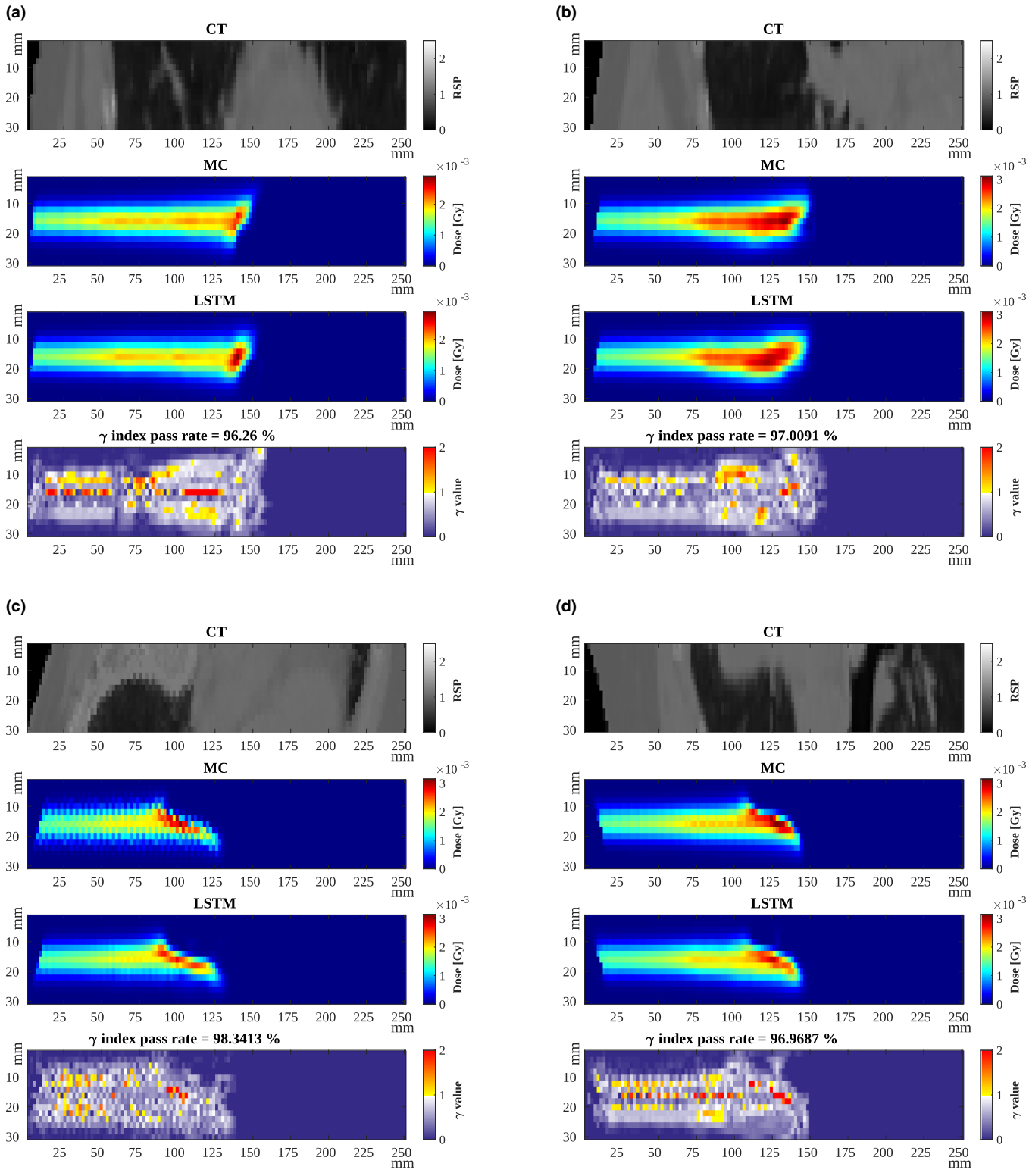

FIG 14. Dose estimation results for four test data from **patient 3** (E=104.25 MeV, γ-index criteria=[1%, 3 mm]). The sub-figures follow the layout outlined in Fig. 9. [Color figure can be viewed at wileyonlinelibrary.com]

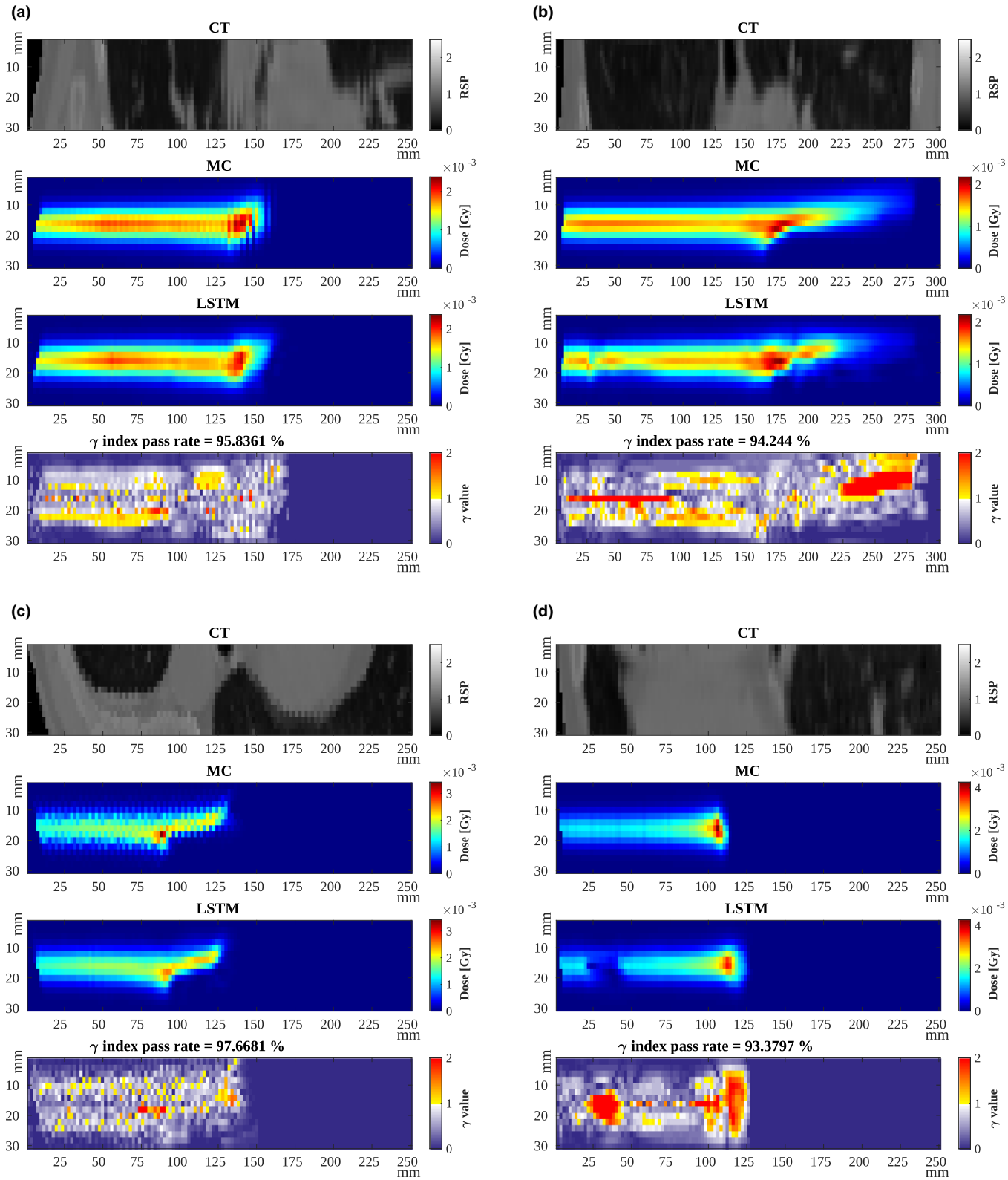

FIG 15. Dose estimation results for four test data from **patient 5** (E=104.25 MeV, γ-index criteria=[1%, 3 mm]). The Subfigurs follow the layout outlined in Fig. 9. Note that the network fails to distinguish between the lung and the air cavity in sample (d) due to the very low RSP value of the lung. [Color figure can be viewed at wileyonlinelibrary.com]