Medical Image Analysis 80 (2022) 102512

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Deep learning DCE-MRI parameter estimation: Application in pancreatic cancer

Tim Ottens[a,*], Sebastiano Barbieri[b], Matthew R. Orton[c], Remy Klaassen[d], Hanneke W.M. van Laarhoven[d], Hans Crezee[e], Aart J. Nederveen[a], Xiantong Zhen[f], Oliver J. Gurney-Champion[a]

[a] *Department of Radiology and Nuclear Medicine, Cancer Center Amsterdam, Amsterdam UMC, University of Amsterdam, HV Amsterdam 1081, the Netherlands*
[b] *Centre for Big Data Research in Health, UNSW, Sydney, Australia*
[c] *Department of Radiology, The Royal Marsden NHS Foundation Trust and The Institute for Cancer Research, Londen, United Kingdom*
[d] *Department of Medical Oncology, Cancer Center Amsterdam, Amsterdam UMC, University of Amsterdam, HV Amsterdam 1081, the Netherlands*
[e] *Department of Radiation Oncology, Cancer Center Amsterdam, Amsterdam UMC, University of Amsterdam, HV Amsterdam 1081, the Netherlands*
[f] *AIM Lab, University of Amsterdam, XH Amsterdam 1098, the Netherlands*

ARTICLE INFO

Article history:
Received 2 December 2021
Revised 4 May 2022
Accepted 6 June 2022
Available online 7 June 2022

MSC:
92C55
92B20

Keywords:
Dynamic contrast enhanced MRI
Magnetic resonance imaging
Convolutional neural network
Recurrent neural network
Unsupervised training
Pancreatic cancer

ABSTRACT

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is an MRI technique for quantifying perfusion that can be used in clinical applications for classification of tumours and other types of diseases. Conventionally, the non-linear least squares (NLLS) methods is used for tracer-kinetic modelling of DCE data. However, despite promising results, NLLS suffers from long processing times (minutes-hours) and noisy parameter maps due to the non-convexity of the cost function. In this work, we investigated physics-informed deep neural networks for estimating physiological parameters from DCE-MRI signal-curves. Three voxel-wise temporal frameworks (FCN, LSTM, GRU) and two spatio-temporal frameworks (CNN, U-Net) were investigated. The accuracy and precision of parameter estimation by the temporal frameworks were evaluated in simulations. All networks showed higher precision than the NLLS. Specifically, the GRU showed to decrease the random error on v_e by a factor of 4.8 with respect to the NLLS for noise (SD) of 1/20. The accuracy was better for the prediction of the v_e parameter in all networks compared to the NLLS. The GRU and LSTM worked with arbitrary acquisition lengths. The GRU was selected for in vivo evaluation and compared to the spatio-temporal frameworks in 28 patients with pancreatic cancer. All neural network approaches showed less noisy parameter maps than the NLLS. The GRU had better test-retest repeatability than the NLLS for all three parameters and was able to detect one additional patient with significant changes in DCE parameters post chemo-radiotherapy. Although the U-Net and CNN had even better test-retest characteristics than the GRU, and were able to detect even more responders, they also showed potential systematic errors in the parameter maps. Therefore, we advise using our GRU framework for analysing DCE data.

1. Introduction

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a promising MRI technique for quantifying perfusion. It is used in a wide range of clinical studies for noninvasive detection, characterization, and therapy monitoring of different diseases such as heart failure, renal rejection, and diverse types of tumours (Khalifa et al., 2014). DCE-MRI uses T1-weighted images that are measured dynamically before, during, and after bolus injection of a contrast agent to generate signal-intensity curves. These can be used to determine the change of the contrast concentration over time. Through the use of tracer-kinetic (TK) modelling, physiological parameters can then be estimated that characterize the vascular permeability and tissue perfusion per voxel (Sourbron and Buckley, 2013).

[*] Corresponding author.
E-mail addresses: t.ottens@amsterdamumc.nl (T. Ottens), s.barbieri@unsw.edu.au (S. Barbieri), matthew.orton@icr.ac.uk (M.R. Orton), r.klaassen@amsterdamumc.nl (R. Klaassen), h.vanlaarhoven@amsterdamumc.nl (H.W.M. van Laarhoven), h.crezee@amsterdamumc.nl (H. Crezee), a.j.nederveen@amsterdamumc.nl (A.J. Nederveen), zhenxt@gmail.com (X. Zhen), o.j.gurney-champion@amsterdamumc.nl (O.J. Gurney-Champion).

https://doi.org/10.1016/j.media.2022.102512

Different fitting methods have been applied to retrieve the underlying physiological parameters from the contrast concentration curves. One of the conventional methods for analysing these curves is the non-linear least squares (NLLS) approach (Ahearn et al., 2005; Buckley, 2002; Guo et al., 2018; Henderson et al., 2000; Murase, 2004). Despite its ability to estimate physiological parameters reasonably well, NLLS has some key limitations. The voxel-wise fitting approach combined with noisy concentration curves and non-convexity of the objective function results in large variance and bias in the parameter predictions, thereby causing poor repeatability of parameter estimates, which is essential for clinical applications of quantitative MRI techniques (Kurland et al., 2012; Rosenkrantz et al., 2015). Another limitation is that the method is computationally demanding and therefore results in long processing times for high-resolution MRI. More elaborate Bayesian estimation methods have shown to reduce the variability in local homogeneous regions, but come at the cost of even longer computation times (Kelm et al., 2009; Schmid et al., 2006; Orton et al., 2007).

Recently, deep neural networks have been explored to estimate the physiological parameters from DCE-MRI contrast concentration curves, due to their ability to generalize well over large datasets and perform inference significantly faster than the conventional methods (Ulas et al., 2019; Bliesener et al., 2020; Ketelkamp and Lingala, 2020; Zou et al., 2020). Ulas et al. (2019) proposed a convolutional neural network (CNN) which learns a mapping from 2D image time-series to physiological parameter maps. They were able to substantially reduce the processing time compared to the conventional methods. However, CNNs are not well suited for processing time-series with certain long-range temporal dependencies. Bliesener et al. (2020) proposes to combine the use of neural networks with Bayesian estimation to have the benefit of shortened computation time and uncertainty estimation from the posterior distribution over the physiological parameters. Both the concentration curve and the Arterial Input Function (AIF) are processed with 1D-convolutional networks, concatenated and put through a dense network to output the scale, rotation, and translation of the approximate posterior over the parameters. The major disadvantage of both proposed approaches is that the models can only be trained on data-specific total acquisition length, which varies per dataset. The effect it has on clinical applications is that the user would have to train the model on its own dataset, thereby limiting the amount of training data the model can use and restricting the usage to departments that are capable to train such neural networks.

One way of removing the dependency of the networks on the acquisition length is by deploying Recurrent Neural Networks (RNNs). Such networks are typically used for natural language and time series prediction tasks where the input data consists of arbitrary lengths. Zou et al. (2020) proposed using a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) which learns to predict the physiological parameters from contrast concentration curves per voxel. Another sequential model which has similar performance to the LSTM is the Gated Recurrent Unit (GRU) (Cho et al., 2014b). The GRU achieves similar performance to the LSTM with fewer parameters, and hence easier to train with fewer data. However, for long data sequences, performance of both LSTM and GRU deteriorates because important information from the start of the sequence is often not preserved at the last output of the network (Cho et al., 2014a).

Bahdanau et al. (2015) showed that using attention layers on the intermediate hidden states improved performance significantly. These layers regulate the information flow per hidden state to prevent information loss for long sequence data.

In this work, we propose a novel framework that utilizes a GRU network combined with attention layers to predict physiological parameters from DCE-MRI contrast concentration curves. The framework handles input with different acquisition lengths, shows to benefit from the addition of attention layers, and reduces computation time substantially compared to conventional methods. We compare the performance of our GRU with several other voxel-wise network approaches, including an LSTM, a fully connected network (FCN), and the conventional NLLS. Performance of the approaches was evaluated in simulated concentration curves, in which we determined the precision and accuracy. Subsequently, we compare the performance of the GRU and NLLS fit to a CNN and a U-Net in vivo, in 28 pancreatic cancer patients. We test the test-retest precision with Bland-Altman analysis and show in how many patients we can detect treatment effects with all four methods.

2. Materials and methods

The proposed networks learn a mapping from DCE-MRI contrast concentration curves to tracer-kinetic (TK) parameters. The TK model that is most commonly used and will also be applied here is the Extended Tofts-Kety (ETK) model (Tofts et al., 1999):

$$C_\mathrm{t}(t \mid \theta) = v_p C_\mathrm{p}(t + dt) + v_e k_{ep} \int_0^t C_\mathrm{p}(\tau + dt) e^{-k_{ep}(t-\tau)} d\tau \qquad (1)$$

with $\theta = \{v_p, v_e, k_{ep}, dt\}$, C_t is the equilibrium concentration of gadolinium in whole tissue, $C_p(t)$ captures the blood plasma contrast agent concentration, also known as the Arterial Input Function (AIF), dt is the voxel-specific arrival delay of the bolus, v_p the volume fraction of plasma in tissue, v_e represents the volume fraction of the extravascular extracellular space in tissue and k_{ep} represents the rate constant for efflux of gadolinium contrast back into plasma from the tissue extracellular space. K^{trans} is the volume transfer constant that influences the degree of contrast enhancement, which is equal to $k_{ep} \times v_e$.

For the purpose of this work, we used the model by Rata et al. (2015), which uses an analytical description of the AIF curve $C_p(t)$ that is characterized by eight parameters. This parameterized AIF allows for an analytical solution to the integrals used in the ETK model, which makes it computationally efficient (Orton et al., 2008). This analytical description also allows estimating onset time. We used the population-based AIF function obtained by Klaassen et al. (2018) and made it patient-specific by adapting it according to the patient's measured hematocrit (HCT).

2.1. TK parameter prediction framework

We implemented two types of frameworks that estimate the physiological parameters ($\tilde{\theta}$) from a concentration curve ($C(t)$): the temporal and spatio-temporal frameworks. The temporal frameworks predicted the physiological parameters based on the temporal information per voxel, while the spatio-temporal frameworks included spatial information through the use of convolutions. Since our synthetic dataset did not contain any correlation between voxels, only the temporal frameworks were evaluated on the synthetic dataset, after which both frameworks were evaluated on the patient dataset. The temporal frameworks that were tested consist of our proposed GRU network, an LSTM network, a fully connected network (FCN) and the conventional NLLS method. The spatio-temporal frameworks consisted of the convolutional neural network (CNN) with a local and global path proposed by Ulas et al. (2019) and a U-Net (Ronneberger et al., 2015), which has not been used for this purpose before, but has proven to perform well for many applications, such as biomedical image segmentation (Litjens et al., 2017).

All proposed networks are available on our GitHub at https://github.com/oliverchampion/DCENET and we encourage people to use and adapt them for their data.

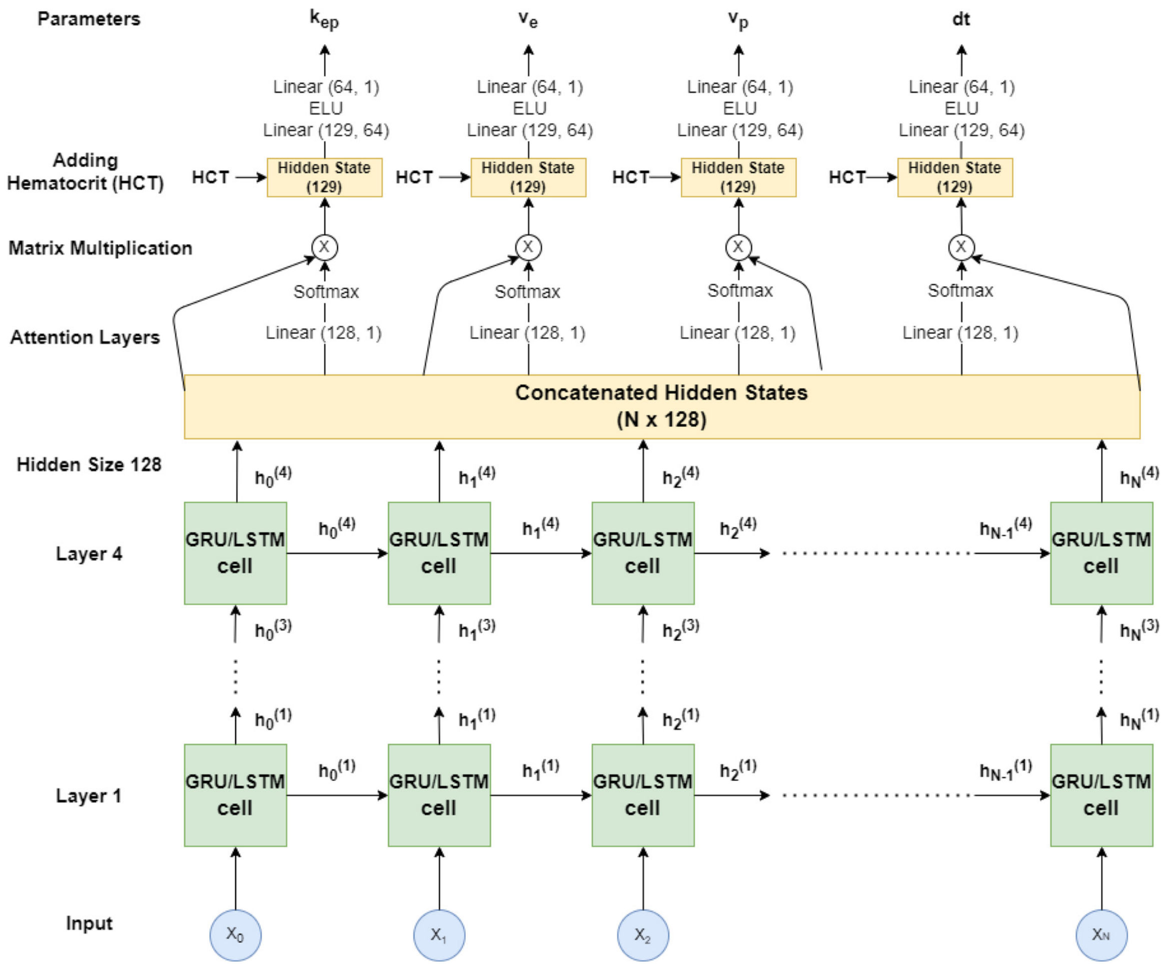

Fig. 1. GRU/LSTM architecture for predicting DCE-MRI parameters from one concentration curve with *N* data-points. Attention layers are added to include intermediate information from hidden states. These are implemented through a linear mapping from a hidden dimension of 128 to a score vector for every parameter from the concatenated hidden states. The scores are then normalized by the softmax function and matrix multiplied by the concatenated hidden states to output a hidden state for each parameter. These hidden states are used as input for a FCN with 2 layers that output one parameter for each voxel.

2.1.1. Temporal frameworks

GRU The proposed network architecture of the GRU with attention layers is illustrated in Fig. 1. The concentration curve with length *N* was processed by four stacked layers of GRU cells. At each cell, the hidden state was calculated and passed on to the next cell according to the following formulas (Cho et al., 2014b):

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\ \hat{h} &= tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h} \end{aligned} \tag{2}$$

where x_t is the input per time step, z_t the update gate, r_t the reset gate, $\hat{h}$ the candidate activation vector, h_t the output vector (hidden state), σ the sigmoid activation function, $\odot$ the Hadamard product and W, U and b the parameter specific weights and bias. The hidden size was heuristically chosen to be 128, since a larger hidden representation size showed little improvement in our experiments. After the last GRU layer, all the hidden states of the cells were concatenated. Attention layers were then used to calculate a score vector of sequence length *N* for every parameter that represented the importance of every hidden state for that specific parameter. These attention layers allow for long-range temporal dependencies to prevent information loss in long sequences. A softmax function was used to normalize the sum of the score vector to a value of 1, after which we did a matrix multiplication between the score and the concatenated hidden states of size 128. The HCT was added to the hidden state of every parameter. The hidden state of size 129 was then mapped with a linear layer, an ELU activation and another linear layer to one value for every parameter.

LSTM The LSTM network architecture was implemented in the same way as the GRU network, where the GRU cells were swapped with LSTM cells (Fig. 1). Zou et al. (2020) showed that the LSTM network architecture outperformed the CNN model from Ulas et al. (2019) while reducing the computation time by a factor of 90 compared to the conventional methods. Here, we extended their model by adding attention layers and feeding the network analytical AIF parameters, instead of the full AIF.

FCN The FCN architecture consisted of four hidden layers with 320, 160, 80 and 40 neurons and an ELU activation in between each layer. The HCT value were given as an additional input at the third layer of the network to emphasize its independence from the sequential input. The output was a linear mapping from the last 40 hidden neurons to the four predicted physiological parameters. A similar approach was described by Kaandorp et al. (2021) and Barbieri et al. (2019) to estimate intra-voxel incoherent motion parameters from diffusion-weighted MRI.

NLLS The NLLS method was used as current clinical standard reference. We implemented NLLS routine by Orton et al. (2008) and Rata et al. (2015), which is available from the OSIPI code collection at https://github.com/OSIPI/DCE-DSC-MRI_CodeCollection (`OG_MO_AUMC_ICR_RMH.DCE.curve_fit()`).

As initial guess, we gave the mean values of the parameters that were used in our synthetic dataset (Section 2.2.1), this resulted in $[k_{ep}, v_e, v_p] = [1, 0.35, 0.025]$. Since we defined the parameters in physiologically plausible ranges, we also assigned these same initial values to the NLLS method when predicting the parameter values of the patient dataset.

2.1.2. Spatio-temporal frameworks

CNN The CNN architecture with a local and global pathway was used by Ulas et al. (2019) to predict TK parameters directly from signal-intensity curves. The CNN took dynamically measured 2D slices ($X \times Y$ voxels) as input where the temporal dimension (N data-points) was given as input channels. Here, we implemented a similar 2D CNN structure with three convolutional layers for both pathways, ELU activation functions (instead of ReLU), a kernel size of 3×3 (instead of 4×4), padding of 1 and a dilation factor of 2, 4 and 8 for the global pathway. The output of both pathways were concatenated, resulting in an output of 320 per voxel, and given as input to a FCN with two hidden layers of sizes 160 and 80, ELU activation functions and an output layer that maps to the four physiological parameters.

U-Net The U-Net architecture (Ronneberger et al., 2015) has proven to perform well in biomedical image segmentation, and hence we adapted it for quantitative DCE parameter estimation. Our 2D U-Net took dynamically measured 2D slices as input where the temporal dimension was given as input channels. It had a depth of four with at each depth two convolutional layers, batch normalization and ReLU activation functions. On the contracting path, the temporal dimension was doubled at every depth and the spatial size of the image was reduced by two through 2D max pooling. The expansive path concatenated the information from the contracting path and expanded the image size by two through 2D transposed convolutions. The final output of 160 per voxel was flattened and concatenated with the HCT, after which we used a FCN with one hidden layer of size 80 and an output layer to predict the four physiological parameters per voxel.

2.1.3. Parameter constraints

To ensure that the predicted extended Tofts parameter values lay in physiologically plausible ranges, the parameters were constrained by:

$$\tilde{\theta} = \theta_{min} + \sigma(\tilde{\theta}_{out}) * (\theta_{max} - \theta_{min}) \quad (3)$$

where $\tilde{\theta}$ is the final parameter prediction, $\tilde{\theta}_{out}$ is the output value of the network, σ is the sigmoid activation function, $\theta_{\min}$ is the lower bound of the parameter and $\theta_{\max}$ the upper bound. The bounds for each parameter were: $k_{ep} \in [1e^{-8}, 3]$, $v_e \in [1e^{-8}, 1]$ and $v_p \in [1e^{-8}, 0.1]$. The $1e^{-8}$ was taken as the lower bound to prevent numerical errors in the ETK model.

2.1.4. Loss function

Since ground truth parameters are not easily obtained from in vivo data, a supervised learning approach can only try and mimic fit results from alternative algorithms in vivo, which limits its potential to outperform conventional fitting. Instead, we adopted an unsupervised learning approach based on a physics-informed loss. We used the mean squared error (MSE) loss between a predicted concentration curve $\tilde{C}_t(t \mid \tilde{\theta})$ and the measured $C_t(t)$. This unsupervised loss leads to a $\tilde{\theta}$ that closely describes the data and hence is close to the ground truth. Since Kettelkamp and Lingala (2020) showed that including the AIF for high-resolution data (<5s acquisition time) improved their results, we personalized our AIF by adjusting our population-based AIF for the patient-specific HCT by scaling the curve by $1/(1 - HCT)$ when calculating $\tilde{C}_t(t \mid \tilde{\theta})$.

2.2. Data preparation

Two datasets were used in our experiments, a synthetic dataset and a retrospectively acquired patient dataset. The synthetic dataset contained simulated concentration curves that were created from known ETK parameters. Knowing the ground truth parameters enabled us to evaluate the performance of the tested methods. Note that these ground truth parameters were not used during training. The patient dataset was used to evaluate the performance between the proposed approaches in a clinical setting.

2.2.1. Synthetic data

We simulated a total of 500,000 signal curves using Eq. (1). Data points were simulated at 160 frames with a temporal resolution of 1.75 seconds per frame. The physiological parameter values were randomly distributed in the following ranges: $k_{ep} \in [0, 2]$; $v_e \in [0.01, 0.7]$; $v_p \in [0.001, 0.05]$. The bolus arrival was varied between the 26th and 32nd dynamic scan and the HCT was varied between 0.3 and 0.6. These ranges depict physiologically realistic values, such that the simulated concentration curves closely resemble the ones from *in vivo* data.

Since the real patient data is inherently noisy, we added Gaussian noise to the simulated concentration curves to make the synthetic data closely resemble the patient data. We added Gaussian noise with standard deviation (SD) $1/n$, where n ranged from 5 to 100 using geometric progression.

To evaluate the networks, we created a test set with 10,000 curves for noise values n=5, 7, 10, 15, 20, 25, 40 and 100. Furthermore, a second test set at noise n=20 was created in which we simulated 10,000 curves per HCT level, ranging from 0.3 to 0.6 in steps of 0.05. This second test set was analysed by the GRU with correct HCT, but also by a GRU which did not receive HCT information and a GRU which received wrong (random) HCT values, to test whether the network used the HCT correctly. A third test set at noise n=20 was created with 10,000 simulations per acquisition length, for which we varied the acquisition length from 80 to 160, in steps of 10. This test set was analysed by the GRU, LSTM and NLLS using no adaptations. For the FCN, zero padding was used to extend the sequences to a length of 160.

2.2.2. Patient data

To test how the network performed in vivo, we applied our network to 28 patients with locally advanced or metastatic pancreatic ductal adenocarcinoma (PDAC) from two studies (NCT01995240; NCT01989000) that were performed at the Amsterdam UMC (Klaassen et al., 2018; 2020). PDAC is a highly aggressive lethal malignancy and is the most prevalent type of pancreatic neoplasm, which accounts for more than 90% of pancreatic cancer cases. One of the studies consisted of 16 patients with repeated MRI scans without any treatment in-between (average 4.5 days apart; range 1–8). The other study consisted of 12 patients that were treated with a complete course of neoadjuvant chemoradiation therapy (CRT) between both scans as part of the PREOPANC trial (Versteijne et al., 2020). Both studies were approved by the local ethics committee and all patients gave written informed consent before participating.

MRI scans were performed with a 3 T system (Ingenia, Philips, Best, The Netherlands). Before obtaining DCE, a baseline T1 map was scanned using an ultrafast gradient echo Look Locker sequence. DCE data were obtained using a dynamic 3D ultrafast gradient echo sequence for 280 s at 1.75 s per frame (Table 1). After the 10th dynamic scan, 0.1 mmol/kg of 1.0 mmol/mL gadobutrol (Gadovist, Bayer Healthcare, Leverkusen, Germany) was administered at a rate of 5 mL/s followed by a 15 mL saline flush at the same rate. HCT levels were measured at the time of the scan. Individual AIFs were averaged over all patients (Klaassen et al., 2018)

Table 1
Details of MRI sequence parameters.

	DCE FFE	T1 look-locker
FOV (RL × AP) (mm^2)	400 × 400	400 × 350
Acquisition matrix	160 × 160	132 × 116
Slice thickness (mm)	2.5 (5.0 non-interpolated)	5.7 (11.4 non-interpolated)
Slices	30	13
TR/TE1 (ms)	3.2/2.0	3.5/1.6
TI1/TI (ms)	–	19/85
FA (°)	20	8
SENSE (RL/AP)	3.6/1.5	3/1.3
Scan time (total) (s)	1.75 (280)	24
Resp. compensation	Post-processing	1 breath hold
Fat saturation	None	None

Abbreviations: FFE: Fast Field Echo, FOV: Field of View, RL: Right Left, AP: Anterior Posterior,
TR: Repetition Time, TE: Echo Time, TI: Inversion Time, FA: Flip Angle.

and a fit was performed to parameterize the resulting AIF curve. Only 9 of the CRT patients had repeated scans after CRT, so only these 9 patients were evaluated.

First, the DCE images were registered using a 4D principal component analysis-based group-wise image registration in Elastix (Huizinga et al., 2016). The baseline T1 (T_{10}) obtained by look-locker were used to calculate effective $T_1(t)$ over time (t) from the signal curves:

$$T_1(t) = \frac{TR}{\log\left(\frac{M_0\sin(\alpha)-S(t)\cos(\alpha)}{M_0\sin(\alpha)-S(t)}\right)} \tag{4}$$

where TR is the repetition time, α the flip angle and $S(t)$ the DCE-MRI signal. M_0 is the fully relaxed magnetization, which is obtained by setting the left-hand side of Eq. (4) to T_{10} and then solving for M_0 for $S(t < t_{bolus})$.

Then, we use $T_1(t)$ to calculate the concentration curve:

$$C(t) = \frac{1}{r_1}\cdot\left(\frac{1}{T_1(t)} - \frac{1}{T_{10}}\right) \tag{5}$$

where r_1 is the relaxivity of the contrast agent (5L mmol^{-1} s^{-1}).

2.3. Experiments

2.3.1. Network training

The datasets were split into a training/validation set with a ratio of 0.9/0.1. This ratio was applied on the total amount of voxels in the synthetic dataset for the temporal frameworks and the total amount of slices in the patient dataset for the spatiotemporal frameworks. The temporal frameworks were trained with a batch size of 256 and an ADAM optimizer (Kingma and Ba, 2015) An adaptive learning rate was used that started at 1e^{-3} and was reduced by a factor of 10 for every three consecutive non-improvements over the validation loss. Training was done for a total duration of 50 epochs, where each epoch consisted of 1000 iterations.

In patient data, the networks were trained only on the 12 CRT patients, but evaluated on both patient sets. The concentration curves that had a mean signal value lower than $1e^{-3}$ were assumed to be background voxels with no contrast enhancement and thus discarded from the patient dataset.

The spatio-temporal frameworks were trained with the same hyperparameters as the temporal frameworks except for the batch size, which was set to 8 images that each consisted of a resolution of 160 × 160. Furthermore, background voxels were not discarded for the spatio-temporal frameworks because this would change the resolution per image and complicate the training procedure.

Deep ensembling was used to address the uncertainty and stochasticity of the deep learning frameworks (Ganaie et al., 2021). Every model was trained five times and the median value of every parameter prediction was used on the validation set.

All frameworks were implemented in Python 3.9.5 using Pytorch 1.9.0 (Paszke et al., 2019) and executed on a single Tesla V100 GPU.

2.3.2. Evaluation metrics

For the synthetic dataset, we calculated the random and systematic error between the predicted parameters and the ground truth as measures for precision and accuracy, respectively. We defined the random error as the standard deviation of the difference between the predicted and ground truth parameter values, and the systematic error as the mean of the difference. These errors are plotted as function of noise (Fig. 2), HCT (Fig. 3) and acquisition length (Fig. 4).

Furthermore, both the LSTM and GRU have attention layers that calculated scores for every hidden state of the time points in the concentration curve. These scores determine how much of the hidden state was used for the final prediction of one specific parameter. We plotted these scores to better understand what the network focusses on.

The NLLS (reference) and the best performing network (GRU) from the simulations were evaluated *In vivo*, alongside the spatio-temporal networks. Due to a lack of ground truth parameter values, we cannot determine random and systematic errors in the patient dataset. Instead, we evaluate the performance of the network based on the structure similarity (SSIM) and normalized root-mean-square error (nRMSE) between the reconstructed concentration curve and the original concentration curve. The nRMSE was defined as follows:

$$nRMSE = \sqrt{\sum_{t=1}^{T}\frac{(y_t - x_t)^2}{\bar{x_t}}} \tag{6}$$

where x_t represents time point t of the original concentration curve, y_t represents time point t of the predicted concentration curve and $\bar{x_t}$ the average value of x at time step t over the whole dataset.

The SSIM metric compares the perceived quality of a predicted imaged to a reference image and is defined as follows:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{7}$$

where x and y are the two images, μ_i is the mean of image i, σ_i is the variance of image i and c_1 and c_2 are parameters for stabilizing a weak denominator based on the range of pixel-values, which are set to $(0.01)^2$ and $(0.03)^2$ respectively. Since the patient dataset consists of image time series, the SSIM score is calculated

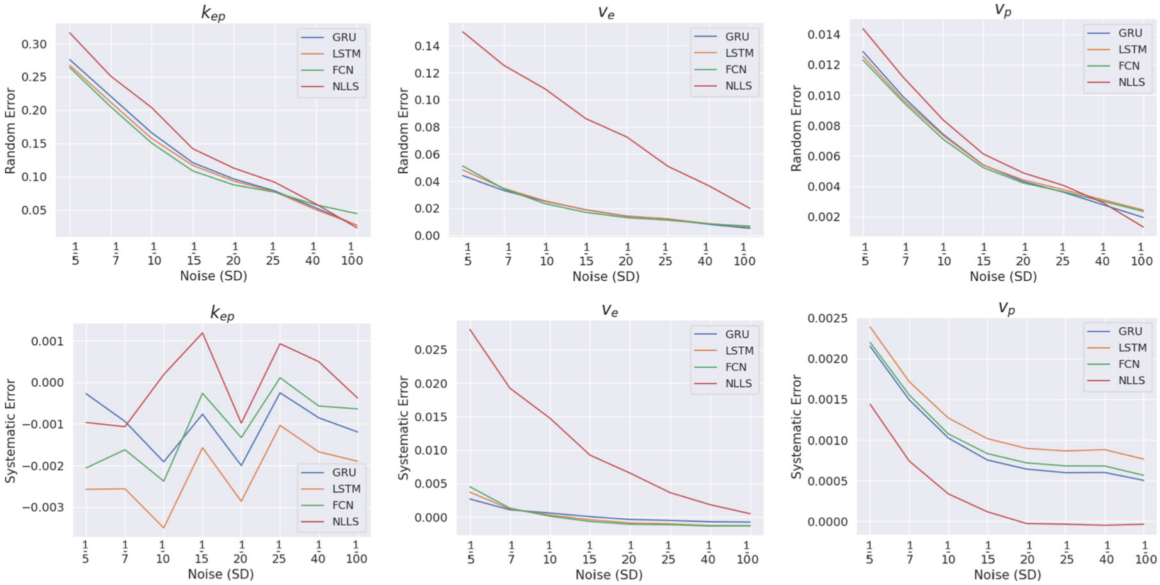

Fig. 2. Random error (top row) and systematic error (bottom row) between the network predictions and the physiological ground truth parameters. Errors are plotted as function of different noise levels. Abbreviations: GRU: Gated Recurrent Unit; FCN: Fully Connected Network; LSTM: Long Short-Term-Memory; NLLS: Non-Linear Least Squares.

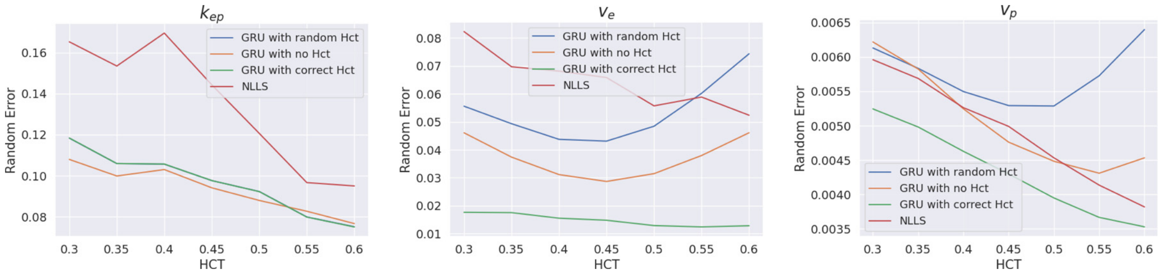

Fig. 3. Measured random error on GRU with the correct, same and without added HCT value and the NLLS. *GRU with correct HCT* received the correct HCT value, *GRU with no HCT* received no HCT value during training and evaluation and *GRU with random HCT* received correct HCT during training, but random HCT values during evaluation, in the same range as simulated in the dataset. In the graph for the k_{ep} parameter, the results of the GRU with correct HCT overlaps with the GRU with random HCT, thereby only showing the former network.

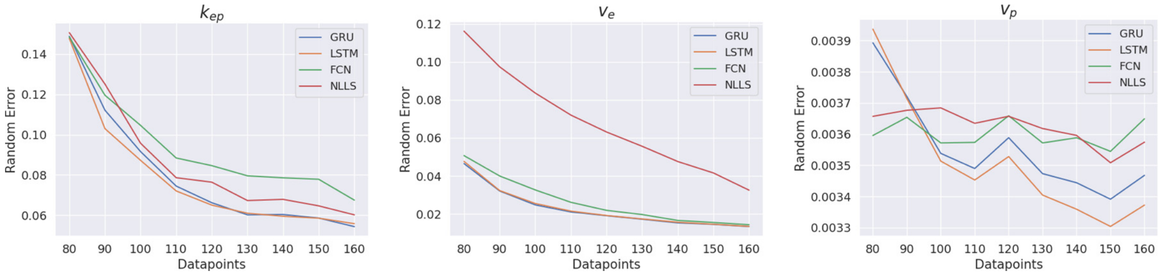

Fig. 4. Measured random error on GRU, LSTM, FCN and NLLS with different acquisition lengths ranging from 80 to 160 and added noise SD of 1/20.

on each time interval and divided by the total acquisition length of the image set.

To determine the clinical value of the network, we investigated whether we could detect changes of parameter values in the tumour of the patients that underwent neo-adjuvant CRT. First, we determined the precision (test-retest repeatability) of our method in the patients with repeated baseline scans using Bland-Altman analysis. This allowed us to determine a 95% confidence interval. Then, we compared the changes observed in the patients receiving the CRT to this 95% confidence interval and classified the changes which were larger than this confidence interval as significant changes. We can then compare the repeatability of all approaches and the potency of each method to detect significant changes in patients receiving treatment.

3. Results

3.1. Synthetic data

Fig. 2 displays the performance of all temporal frameworks on the parameter prediction task. The neural networks had higher precision (lower random error) than the NLLS method for all parameters. Furthermore, they were substantially better at predicting v_e for all noise levels. However, as the amount of noise decreased,

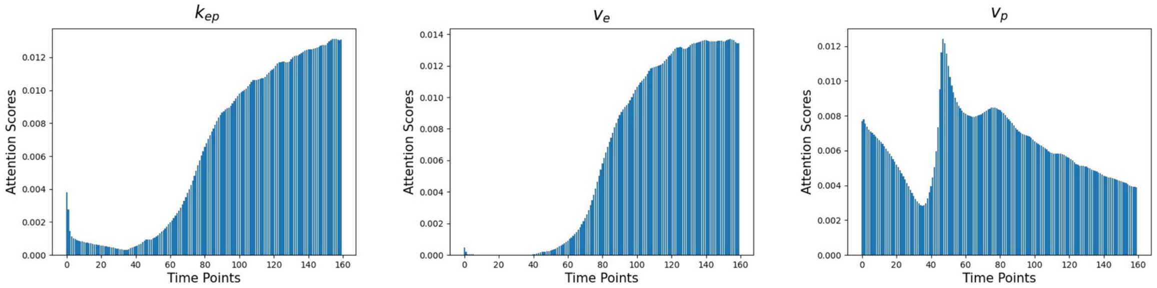

Fig. 5. Scores of the hidden states per time point extracted from the attention layers of the GRU. The scores indicate the amount of hidden information passed from that datapoint to the final prediction of the parameters. For this experiment, bolus arrival was set to the 32th time point (56 seconds after start of measurement).

the NLLS method performed better in predicting k_{ep} and v_p. No substantial differences were observed between networks.

There was a slightly larger bias in the neural networks for v_p (systematic error) than for NLLS. This roughly translated to a 3% bias term error relative to the mean v_p value in the synthetic dataset, while the other parameters had a bias term error lower than 0.3%. These biases were considerably smaller than the random error. The systematic error for the LSTM network on the prediction of v_p was larger than for the other frameworks. The systematic errors on k_{ep} were close to 0 for all approaches and not affected by the difference in noise. Furthermore, all neural networks achieved a lower systematic error than the NLLS method on the prediction of v_e. This correlates with the relatively high random error of the NLLS on the prediction of v_e.

The GRU is able to correctly take into account the HCT, and requires the right HCT value to make a good prediction, as shown in Fig. 3. Looking at the figures, we see that the improvement of adding the correct HCT value shows to be substantial for v_e and v_p. However, we also observe a decrease in performance in predicting k_{ep} with respect to the framework that is trained without the additional HCT value.

Sequential models, like the LSTM and GRU, can receive different input sizes as opposed to the FCN architectures. For the FCN architecture, zero padding must be used on the data before giving it as input to the network. In Fig. 4 performance is shown for the GRU, LSTM, FCN and NLLS on varying acquisition lengths with an added noise SD value of 1/20, where the input is given directly to the GRU, LSTM and NLLS, and zero padding is used to extend the sequence to 160 datapoints for the FCN. It shows that the GRU and LSTM perform better than the NLLS and FCN in predicting all parameters with a small exception for the v_p parameter on an acquisition length less than 100 datapoints.

The attention layer scores (Fig. 5) highlight that k_{ep} and v_e focus on the last part of the concentration curve, considerably later than the bolus arrival (around 32nd time point), while v_p focuses more on the bolus arrival itself. All networks have some focus at the beginning of the sequence, before bolus arrival, where concentration is zero.

Performance with the GRU and LSTM was generally superior to the FCN, and in particular, the FCN performed noticeably worse when evaluating data of different lengths. As there was no clear benefit of the LSTM over the GRU, we chose to evaluate the slightly simpler GRU in vivo.

3.2. Patient data

3.2.1. Quantitative analysis

The difference in SSIM and nRMSE scores between the implemented frameworks was small (Table 2). The GRU achieved marginally higher SSIM and lower nRMSE than the NLLS method. The spatio-temporal frameworks achieved marginally lower SSIM and a marginally higher nRMSE than the NLLS and GRU method. The time per slice of all networks was substantially shorter in inference than the NLLS.

3.2.2. Qualitative analysis

Overall, the NLLS method and GRU showed quite similar parameter maps with the GRU giving less noisy predictions, particularly for v_e (Fig. 6). These results agree with the simulations. The CNN showed noisy predictions on the edges of the abdominal region for the k_{ep} parameter and in the v_p maps, there was an increase in the liver and decrease in the left kidney, when compared to the NLLS and GRU (orange and brown ovals). The U-Net showed smooth k_{ep} plots but also a lack of detail in comparison to the other methods. Fig. 7 shows a close-up of the segmented PDAC of the same patient.

3.2.3. Bland-Altman plots

The GRU had better (lower) 95% confidence interval (CI_{95}) for all parameters than the NLLS (Fig. 8; Table 2). On the other hand, treatment effects were very similar, allowing the GRU to identify one additional significant change in k_{ep} for a borderline responder in the NLLS (Fig. 8, light blue). For CNN and U-Net, the CI_{95} is even lower for k_{ep} and v_e, with U-Net being best. This allows U-Net to now classify 5 patients as having a significant change in k_{ep} as opposed to 2 for NLLS and 4 as opposed to 2 for v_e. In contrast, the CI_{95} of v_p is highest for the U-Net, allowing it to classify 1 fewer significant change in v_p.

4. Discussion

We successfully implemented several neural network based frameworks for analysing DCE data. Using simulations, we show that our temporal frameworks output more precise parameter estimates for all parameters, and more accurate for k_{ep} and v_e, than conventional NLLS. From these temporal frameworks, we selected the GRU as best, as it was capable to deal with arbitrary datalength and had fewer trainable parameters than the LSTM, although LSTM performed similarly well. In vivo, we showed that our networks resulted in less noisy parameter maps compared to NLLS and were at least 90 times faster. Generally, the neural networks had less day-to-day variation than the NLLS, particularly the U-Net. This allowed us to detect more treatment induced changes than when using NLLS. However, the spatio-temporal frameworks did show potential biases in the v_p and had worse nRMSE and SSIM than the GRU. Therefore, we suggest using the GRU.

Although others have trained neural networks for DCE fitting (Ulas et al., 2019; Zou et al., 2020; Bliesener et al., 2020; Kettelkamp and Lingala, 2020), their focus was on being faster than NLLS. Here, we focus not only on being faster, but also on outperforming the NLLS. We are therefore the only study that did extensive simulations to characterize the accuracy and precision of networks for fitting DCE and show and quantify their advantages. Furthermore, we assess their performance quantitatively in vivo and

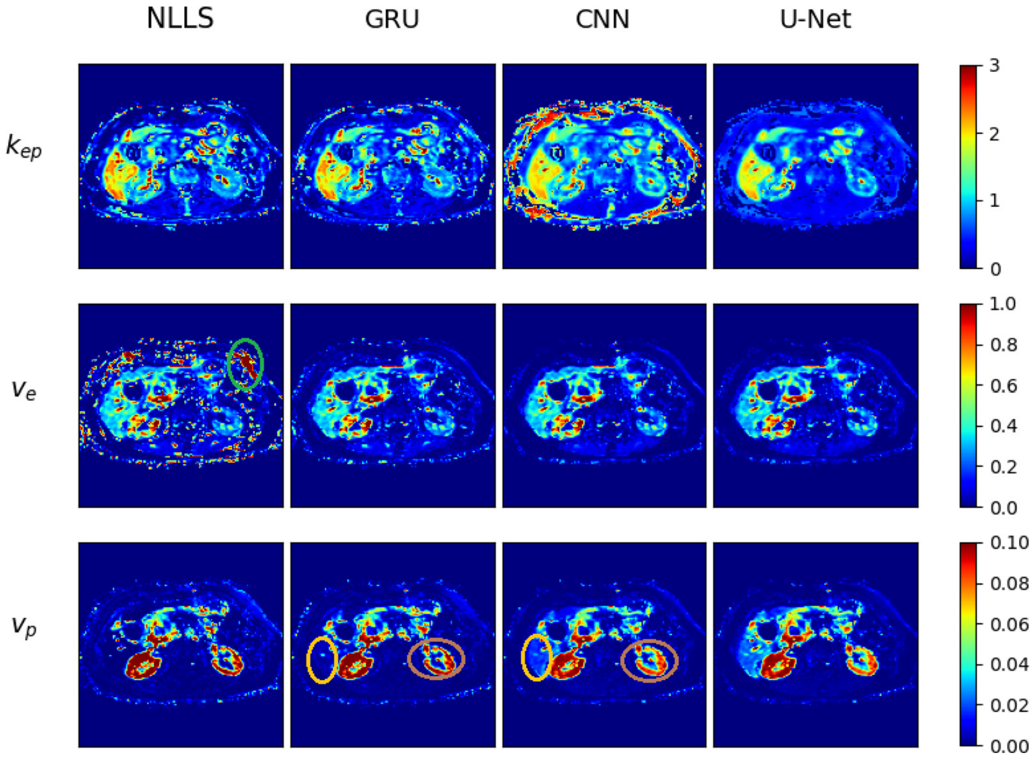

Fig. 6. Parameter maps of a 57-year-old female patient with PDAC. Maps were obtained with the NLLS, GRU, CNN and U-Net. The GRU shows less noisy predictions on the v_e parameter than the NLLS method (green oval). The CNN shows noisy predictions on the border cases of the k_{ep} parameter while the U-Net shows a smooth parameter map but a lack of detail compared to the other methods. The CNN and U-Net also predict higher v_p values for the liver region (orange oval) and lower v_p values in the left kidney (brown oval) compared to the NLLS and GRU.

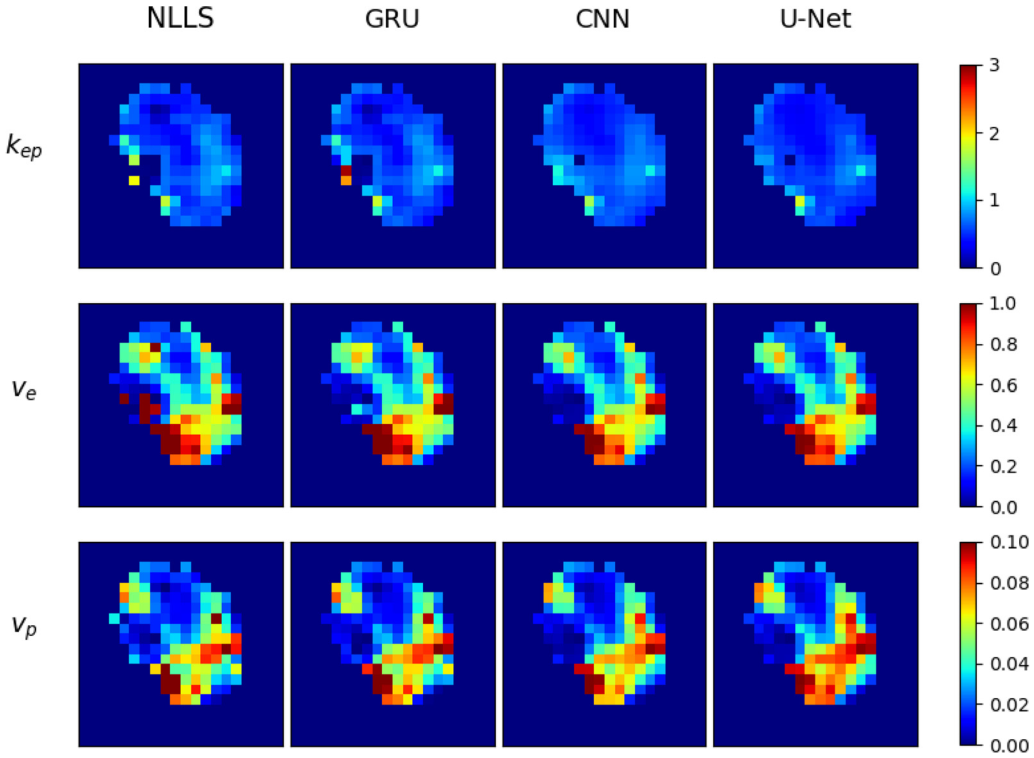

Fig. 7. Parameter maps of the segmented PDAC in the abdominal region obtained by the models: NLLS, GRU, CNN and U-Net. The methods show similar plots for all parameters with slight deviations in the intensity of some voxels.

Table 2
Evaluation of the NLLS method and the different implemented frameworks used on our patient data.

Network	SSIM	nRMSE	CI_{95} (k_{ep}, v_e, v_p)	CV (k_{ep}, v_e, v_p)	Time per slice (s)	Type
NLLS	0.805	0.312	(1.53e^{-1}, 2.62e^{-1}, 7.84e^{-3})	(0.22, 0.31, 0.46)	61.71	Temporal
GRU	0.807	0.312	(1.27e^{-1}, 2.25e^{-1}, 6.96e^{-3})	(0.19, 0.31, 0.44)	0.68	Temporal
CNN	0.800	0.323	(1.20e^{-1}, 1.80e^{-1}, 9.59e^{-3})	(0.17, 0.29, 0.41)	0.43	Spatio-temporal
U-Net	0.800	0.323	(0.74e^{-1}, 1.62e^{-1}, 10.10e^{-3})	(0.14, 0.29, 0.40)	0.33	Spatio-temporal

Abbreviations: SSIM: structural similarity index measure, nRMSE: normalized root-mean-square error, CI_{95}: 95% confidence interval, CV: coefficient of variation

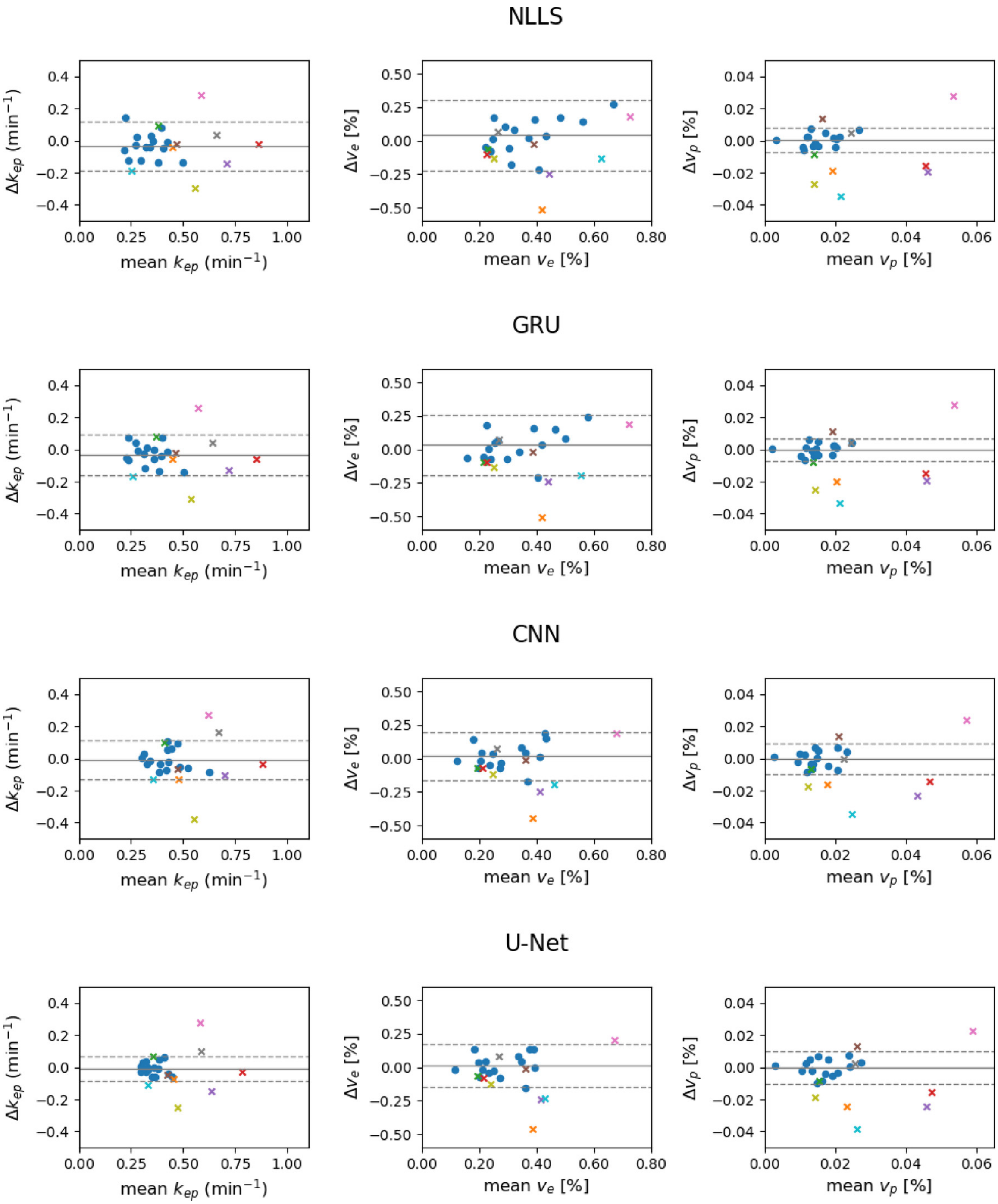

Fig. 8. Bland-Altman plots of the neural network approaches to fitting of the physiological parameters of DCE-MRI concentration curves. The difference in parameter value from the PDAC between repeated sessions is plotted against the mean over both sessions (blue dots). Furthermore, the difference in parameter value from the PDAC between pre and post chemoradiotherapy treatment is plotted as function of the mean (coloured crosses). The dotted lines indicate the 95% confidence intervals for typical day-to-day changes whereas the line solid shows the bias between both visits. Plots are shown for NLLS (top), GRU (second row), CNN (third row) and U-Net (bottom row) for k_{ep} (left column), v_e (middle column) and v_p (right column).

show their measurements are more stable and better at detecting treatment effects.

As reported in other papers (Ulas et al., 2019; Zou et al., 2020; Bliesener et al., 2020; Kettelkamp and Lingala, 2020), the computation time per slice was substantially lower for our networks than for the NLLS method. It is worth highlighting that this is the case for inference, while the training times can be considerably longer than NLLS. This further highlights the advantage of a general network, like GRU, that once properly trained can be used for any acquisition length, without retraining (Fig. 4).

All previously published networks, trained to predict DCE, were trained with (partially) supervised losses. Our training was fully unsupervised, with a physics-informed loss. This allowed us to train on real data, without the need for obtaining reference values. We believe that, at least in part because of this, our networks were able to outperform NLLS, whereas supervised training would have resulted in the network mimicking the NLLS. Interestingly, the networks are trained to the same loss as the NLLS is trying to minimize. However, such a loss-landscape (as function of ETK parameters) can be complex, and the NLLS can potentially get stuck in local minima, or find a global minimum using an unlikely parameter combination. The neural network, on the other hand, needs to define a set of weights that find the minimal loss for all DCE curves at once. It seems that, as a result of this, it returns more accurate and precise ETK parameters. Potentially, the network learns some Bayesian prior. But also, it seems to be better at finding global minima in the loss, which is reflected by the lower nRMSE (Table 2).

We implemented several aspects to make our network more general. Different DCE-MRI scans can contain different total acquisition lengths, therefore a model that could handle these varying sequence lengths would be beneficial. The LSTM and GRU were specifically tested because of their ability to handle such varying acquisition lengths and also showed that the they outperform the FCN and NLLS on predicting all physiological parameters for most of the tested acquisition lengths (Fig. 4). Furthermore, patients have varying AIFs. Fig. 3 highlights that the HCT is taken into account properly in our networks and that it has a large effect on the prediction of v_e and v_p, with quite a substantial improvement. Interestingly, there was a decrease in accuracy for the k_{ep} parameter in comparison with a network with no additional HCT value. Also, there is no difference in error between passing the correct HCT and a random HCT for the k_{ep} parameter. HCT only influences the height of C_p, but not the shape. With Eq. 1, HCT could hence be seen as a constant with which C_p is multiplied, and when moved outside the integral can be accounted for with only v_e and v_p. Hence, the network has no need for HCT in the k_{ep} branch, explaining why the network is agnostic to HCT for predicting k_{ep}. Furthermore, all parameter predictions of the GRU that sees the correct HCT show that the random error is inversely related with the HCT value. The C_p relates to HCT by $C_p = AIF_{pop}/(1 - HCT)$, in which AIF_{pop} is our population-based AIF. A high HCT-value hence results in higher C_p values, which consequently results in higher overall signal (Eq. (1)), explaining the inverse relation between random error and HCT.

We are the first to introduce attention layers for predicting ETK parameters. These layers also allow us to ”see" what the network is focussing on, and, in the era of explainable AI, ensure ourselves it is performing well (Fig. 5). Our network is looking at the right location in the curve for the different predictions. k_{ep} and v_e have a large impact once concentration leaks in and out of the extra vascular extracellular space, which is best visible towards the end of the concentration curve, as reflected in their attention scores. The first pass of the bolus is most informative to v_p, as at that point no exchange has occurred yet. Interesting is the score for the first few time points where there should be no contrast enhancement. The physiological parameters have no influence on the part before bolus arrival and thus little attention should be given to those time points. Possibly, the network is trying to estimate the typical noise of the data from these time points.

One potential pitfall of using deep neural networks for parameter prediction is that they make poor predictions when data is outside their training regime. Particularly, Gyori et al. (2021) show that certain pathologies can be obscured when they are underrepresented in the training data, with supervised learning. We saw no clear evidence for this happening in our temporal network. We believe that, due to the unsupervised loss, our network may be better at generalizing, as it directly learns the relation between the data, predicted ETK parameters and contrast concentration curve.

The nRMSE is a direct measure from how close the fitted curve lays to the data, and closely resembles the loss used in the networks and NLLS. Hence, a lower nRMSE would reflect a more optimal solution found by the algorithm, in terms of minimizing its loss. The GRU had an equal nRMSE compared to the NLLS, suggesting it is at least as good at determining the global minima as the NLLS (Table 2). The spatio-temporal networks had higher nRMSE than the GRU and NLLS. A possible explanation for the difference in performance is the training data. The temporal frameworks predicted the physiological parameters per voxel, whereas the spatio-temporal frameworks included information from the neighbouring voxels. This also means that background voxels, which take up a substantial amount of an image, need to be included in the spatio-temporal frameworks, whereas in temporal frameworks they can be discarded. The influence on the training procedure could be observed when all voxels were given as input to the temporal framework, and as a result, performance decreased.

The qualitative results in Fig. 6 showed that the neural networks output less noisy parameter maps than NLLS. However, the spatio-temporal networks differ in their predictions of the intensity of the v_p parameter map around the liver and left kidney, whereas the GRU shows similar mean intensity as the NLLS parameter maps. As we extensively tested GRU in simulations and show no substantial bias, it is expected to give accurate results in vivo. However, we had no systematic way of comparing the U-Net to ground truth data. We believe this bias may occur due to the training data being too few. In particular, as convolutional networks, the spatio-temporal frameworks had substantially more trainable parameters while the number of training samples decreased as training was done slice-wise, not voxel-wise. Potentially, increasing the training data could further improve the spatio-temporal frameworks' performance.

The GRU performed marginally better than the NLLS method in the Bland-Altman analysis, with reduced day-to-day variation, quantified as CI_{95}, and consequently 1 additional patient with significant changes. The CNN and, particularly, the U-Net were substantially better at detecting individuals with significant change in ETK parameter after treatment, with even lower CI_{95} for k_{ep} and v_e. The CV values are in line with the values from the confidence interval but show less deviation from the NLLS values. This is because the CV depends on the mean value of the parameter estimate which is estimated lower by the CNN and U-Net compared to the NLLS.

On the other hand, the U-Net and CNN performed worse when solely looking at nRMSE and SSIM. Potentially, including spatial awareness allows spatio-temporal frameworks to balance data consistency and spatial consistency, sometimes resulting in ETK parameter estimates that do not closest resemble the data curve for some voxels, but better fit the spatial trend. Therefore, voxel-wise methods like NLLS and GRU achieve the lowest nRMSE scores but are more subjected to the inherit noise per voxel.

Without ground truth data for the spatio-temporal frameworks, it is hard to know whether they are reliable. So despite their great performance in the Bland-Altman analysis, we would advice to proceed using the GRU for the time being, due to the biases found. To better assess the performance improvement of spatio-temporal frameworks, a next step would be to create 3D synthetic data with a certain correlation in between the voxels. This dataset can then be evaluated in the same way as is done for the temporal frameworks. Furthermore, the spatio-temporal frameworks were trained with background voxels, which showed to decrease performance when also used in the temporal frameworks. To resolve this issue, another approach would be desirable that could circumvent using these background voxels during training.

A limitation of our work is that we have set out to make a generalized network, but we have only tested it for a given HCT-corrected population-based AIF. For our in vivo dataset, we had shown in previous work (Klaassen et al., 2018) that repeatability was best with a HCT-corrected population-based AIF and hence that was what we aimed for. However, if the network can adapt for the HCT by adding it to the hidden layer after the GRUs, then additional AIF parameters can be added there too. In fact, in preliminary simulations (not shown), we have confirmed that the network works for arbitrary AIFs by adding the AIF parameters from Rata et al. (2015) to the hidden state of our GRU network.

Although our tests on the synthetic dataset showed that generally our neural network based approaches are favourable over the conventional NLLS method, there were some points at which the NLLS scored better. The systematic error in the neural networks for the prediction of the v_p parameter is larger than for the NLLS method. Potentially, the v_p was estimated slightly poorer due to the way the unsupervised loss was defined. v_p is predominantly determined by the first pass of contrast agent, which is a fast process. Hence, it only effects a limited number of time-frames in the data and consequently, also the loss. Besides, v_p is generally low and hence often only has a little effect on the curve. Hence, v_p has a limited effect for only a limited number of timeframes, resulting in a limited contribution to the loss function. Therefore, the loss function forces the network to focus on optimizing the other parameters. Potentially, tweaking the loss to focus on this first pass could help improve the v_p prediction. Another possible solution could be to vary the parameters one by one, showing the difference more clearly, instead of training the network on parameter values that vary simultaneously.

5. Conclusion

In this work, we presented various frameworks that analyse DCE-MRI concentration curves and output robust and reliable extended Tofts-Kety parameter estimates. Our implemented framework, combined with the GRU and attention layers, is our method of choice, since we tested it thoroughly on synthetic data and patient data, where it outperforms the NLLS method at typical in vivo noise, at substantially shorter computation times.

Declaration of Competing Interest

None.

Acknowledgments

The KWF Dutch Cancer Society supported this work under Grants No. KWF-UVA 2017-10873 and KWF-UVA 2013-5932.

References

Ahearn, T., Staff, R., Redpath, T., Semple, S., 2005. The use of the levenberg-marquardt curve-fitting algorithm in pharmacokinetic modelling of dce-mri data. Phys. Med. Biol. 50 9, N85–92.

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. CoRR. abs/1409.0473

Barbieri, S., Gurney-Champion, O.J., Klaassen, R., Thoeny, H.C., 2019. Deep learning how to fit an intravoxel incoherent motion model to diffusion weighted mri. Magn. Reson. Med. 83, 312–321.

Bliesener, Y., Acharya, J., Nayak, K., 2020. Efficient DCE-MRI parameter and uncertainty estimation using a neural network. IEEE Trans. Med. Imaging 39, 1712–1723.

Buckley, D., 2002. Uncertainty in the analysis of tracer kinetics using dynamic contrastenhanced t1weighted MRI. Magn. Reson. Med. 47.

Cho, K., Merrienboer, B. V., Bahdanau, D., Bengio, Y., 2014a. On the properties of neural machine translation: encoder-decoder approaches. ArXiv abs/1409.1259.

Cho, K., Merrienboer, B. V., aglar Gülehre, Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014b. Learning phrase representations using RNN encoder decoder for statistical machine translation. ArXiv abs/1406.1078.

Ganaie, M. A., Hu, M., Tanveer, M., Suganthan, P. N., 2021. Ensemble deep learning: a review. ArXiv abs/2104.02395.

Guo, Y., Lingala, S.G., Bliesener, Y., Lebel, R.M., Zhu, Y., Nayak, K., 2018. Joint arterial input function and tracer kinetic parameter estimation from undersampled dynamic contrast enhanced MRI using a model consistency constraint. Magn. Reson. Med. 79, 2804–2815.

Gyori, N.G., Palombo, M., Clark, C.A., Zhang, H., Alexander, D.C., 2021. Training data distribution significantly impacts the estimation of tissue microstructure with machine learning. Magn. Reson. Med..

Henderson, E., Sykes, J., Drost, D., Weinmann, H.J., Rutt, B., Lee, T.Y., 2000. Simultaneous MRI measurement of blood flow, blood volume, and capillary permeability in mammary tumors using two different contrast agents. J. Magn. Reson. Imaging 12, 991–1003.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Huizinga, W., Poot, D.H.J., Guyader, J.-M., Klaassen, R., Coolen, B.F., van Kranenburg, M., van Geuns, R.J., Uitterdijk, A., Polfliet, M., Vandemeulebroucke, J., Leemans, A., Niessen, W.J., Klein, S., 2016. PCA-based groupwise image registration for quantitative MRI. Med. Image Anal. 29, 65–78.

Kaandorp, M.P., Barbieri, S., Klaassen, R., Laarhoven, H.W., Crezee, H., While, P.T., Nederveen, A.J., Gurney-Champion, O.J., 2021. Improved unsupervised physics informed deep learning for intravoxel incoherent motion modeling and evaluation in pancreatic cancer patients. Magn. Reson. Med. 86, 2250–2265.

Kelm, B., Menze, B., Nix, O., Zechmann, C., Hamprecht, F., 2009. Estimating kinetic parameter maps from dynamic contrast-enhanced MRI using spatial prior knowledge. IEEE Trans. Med. Imaging 28, 1534–1547.

Kettelkamp, J., Lingala, S.G., 2020. Arterial input function and tracer kinetic model-driven network for rapid inference of kinetic maps in dynamic contrast-enhanced MRI (AIF-TK-net). In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1450–1453.

Khalifa, F., Soliman, A., El-Baz, A., El-Ghar, M.E.A., El-Diasty, T., Gimel'farb, G., Ouseph, R., Dwyer, A., 2014. Models and methods for analyzing DCE-MRI: a review. Med. Phys. 41 12, 124301.

Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. CoRR. abs/1412.6980

Klaassen, R., Gurney-Champion, O., Wilmink, J., Besselink, M., Engelbrecht, M., Stoker, J., Nederveen, A., van Laarhoven, H.V., 2018. Repeatability and correlations of dynamic contrast enhanced and t2* MRI in patients with advanced pancreatic ductal adenocarcinoma. Magn. Reson. Imaging 50, 1–9.

Klaassen, R., Steins, A., Gurney-Champion, O.J., Bijlsma, M.F., tienhoven, G.V., Engelbrecht, M.R.W., van Eijck, C.H.J., Suker, M., Wilmink, J.W., Besselink, M.G., Busch, O.R.C., de Boer, O.J., van de Vijver, M.J., Hooijer, G.K.J., Verheij, J., Stoker, J., Nederveen, A.J., van Laarhoven, H.W., 2020. Pathological validation and prognostic potential of quantitative MRI in the characterization of pancreas cancer: preliminary experience. Mol. Oncol. 14, 2176–2189.

Kurland, B., Gerstner, E., Mountz, J., Schwartz, L., Ryan, C., Graham, M.M., Buatti, J., Fennessy, F., Eikman, E., Kumar, V., Forster, K., Wahl, R., Lieberman, F., 2012. Promise and pitfalls of quantitative imaging in oncology clinical trials. Magn. Reson. Imaging 30 9, 1301–1312.

Litjens, G.J.S., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Murase, K., 2004. Efficient method for calculating kinetic parameters using t1weighted dynamic contrastenhanced magnetic resonance imaging. Magn. Reson. Med. 51, 858–862.

Orton, M.G., Collins, D., Walker-Samuel, S., dArcy, J., Hawkes, D., Atkinson, D., Leach, M., 2007. Bayesian estimation of pharmacokinetic parameters for dce-mri with a robust treatment of enhancement onset time. Phys Med Biol 52 9, 2393–2408.

Orton, M.R., darcy, J.A., Walker-Samuel, S., Hawkes, D.J., Atkinson, D., Collins, D.J., Leach, M.O., 2008. Computationally efficient vascular input function models for quantitative kinetic modelling using DCE-MRI. Phys. Med. Biol. 53 5, 1225–1239.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: an imperative style, high-performance deep learning library. NeurIPS.

Rata, M., Collins, D., Darcy, J., Messiou, C., Tunariu, N., deSouza, N., Young, H., Leach, M., Orton, M., 2015. Assessment of repeatability and treatment response in early phase clinical trials using DCE-MRI: comparison of parametric analysis using mr- and ct-derived arterial input functions. Eur. Radiol. 26, 1991–1998.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. MICCAI.

Rosenkrantz, A., Mendiratta-Lala, M., Bartholmai, B., Ganeshan, D., Abramson, R., Burton, K., Yu, J.-P.J., Scalzetti, E., Yankeelov, T., Subramaniam, R., Lenchik, L., 2015. Clinical utility of quantitative imaging. Acad. Radiol 22 1, 33–49.

Schmid, V., Whitcher, B., Padhani, A., Taylor, N.J., Yang, G., 2006. Bayesian methods for pharmacokinetic models in dynamic contrast-enhanced magnetic resonance imaging. IEEE Trans. Med. Imaging 25, 1627–1636.

Sourbron, S., Buckley, D., 2013. Classic models for dynamic contrastenhanced mri. NMR Biomed. 26, 1004–1027.

Tofts, P., Brix, G., Buckley, D., Evelhoch, J., Henderson, E., Knopp, M., Larsson, H., Lee, T.Y., Mayr, N., Parker, G., Port, R., Taylor, J., Weisskoff, R., 1999. Estimating kinetic parameters from dynamic contrastenhanced t1weighted MRI of a diffusable tracer: standardized quantities and symbols. J. Magn. Reson. Imaging 10, 223–232.

Ulas, C., Das, D., Thrippleton, M., Hernández, M.V.V., Armitage, P., Makin, S.D., Wardlaw, J., Menze, B., 2019. Convolutional neural networks for direct inference of pharmacokinetic parameters: application to stroke dynamic contrast-enhanced mri. Front. Neurol. 9, 1147.

Versteijne, E., Suker, M., Groothuis, K.B.C., Akkermans-Vogelaar, J.M., Besselink, M.G., Bonsing, B.A., Buijsen, J., Busch, O.R.C., Creemers, G.-J.M., Dam, R.M.V., Eskens, F.A., Festen, S., de Groot, J.W.B., Koerkamp, B.G., de Hingh, I.H.J.T., Homs, M.Y.V., van Hooft, J.E., Kerver, E.D., Luelmo, S.A.C., Neelis, K.J., Nuyttens, J.J., Paardekooper, G., Patijn, G.A., van der Sangen, M.J., de Vos-Geelen, J., Wilmink, J.W., Zwinderman, A.H., Punt, C.J.A., van Eijck, C.H.J., tienhoven, G.V., 2020. Preoperative chemoradiotherapy versus immediate surgery for resectable and borderline resectable pancreatic cancer: results of the dutch randomized phase III preopanc trial. J. Clin. Oncol. 38, 1763–1773.

Zou, J., Balter, J., Cao, Y., 2020. Estimation of pharmacokinetic parameters from DCE-MRIby extracting long and short time-dependent features using an LSTM network. Med. Phys. 47, 3447–3457.