

Doctoral thesis

Doctoral theses at NTNU, 2024:148

Misha Pieter Thijs Kaandorp

Deep Learning Diffusion Parameters from Magnetic Resonance Imaging

An Odyssey of Deep Learning IVIM
Parameter Estimation

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Medicine and Health Sciences
Department of Circulation and Medical Imaging



Norwegian University of
Science and Technology

Misha Pieter Thijs Kaandorp

Deep Learning Diffusion Parameters from Magnetic Resonance Imaging

An Odyssey of Deep Learning IVIM Parameter
Estimation

Thesis for the Degree of Philosophiae Doctor

Trondheim, April 2024

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Circulation and Medical Imaging



Norwegian University of
Science and Technology

NTNU

Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Medicine and Health Sciences

Department of Circulation and Medical Imaging

© Misha Pieter Thijs Kaandorp

ISBN 978-82-326-7890-7 (printed ver.)

ISBN 978-82-326-7889-1 (electronic ver.)

ISSN 1503-8181 (printed ver.)

ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2024:148

Printed by NTNU Grafisk senter

Sammendrag

Dyp læring av diffusjonsparametere fra magnetisk resonansavbildning

Kreft er en fremtredende årsak til for tidlig dødelighet over hele verden og kan ofte forebygges gjennom tidlig oppdagelse. Magnetisk resonansavbildning (MRI) er en viktig modalitet i etterforskningen av mange typer kreft, på grunn av dens sterke følsomhet for bløtvevskontrast. Utover denne evnen kan MR gi kvantitativ eller parametrisk informasjon om biofysiske vevsegenskaper og mikrostrukturelle prosesser, inkludert molekylær diffusjon og vevsperfusjon. Bruk av passende matematiske modeller for MR-data gjør det mulig å ekstrahere disse kvantitative egenskapene gjennom en prosess kjent som modelltilpasning.

I dag er den konvensjonelle MR-tilnærmingen for å vurdere vevsperfusjonsegenskaper for kreftdeteksjon og behandlingsovervåking dynamisk kontrastforsterket MR (DCE-MRI). Innhenting av DCE-MRI-data krever imidlertid injeksjon av et kontrastmiddel, typisk gadolinium. Denne tilnærmingen kan utgjøre kontraindikasjoner for personer med nyreinsuffisiens, i tillegg til de som er gravide eller ammer. Nyere funn indikerer også mulig avsetning av gadolinium i hjernen, noe kan vekke bekymringer om denne tilnærmingen.

Som et tillegg til DCE-MRI har diffusjonsvektet MR (DWI) vist seg å gi økt spesifisitet i kreftdiagnostikk. DWI bruker den tilfeldige bevegelsen til vannmolekyler, også kjent som diffusjon, for å generere bildekontrast. Ved å skaffe bilder ved flere diffusjonsvektinger (b-verdier), kan en diffusjonsmodell brukes på signalforfallet for hver voksel. Tilpasning av den klassiske mono-eksponentielle modellen til DWI-dataene gir den tilsynelatende diffusjonskoeffisienten (ADC), en metrikk som reflekterer vanddiffusiviteten i det observerte vevet. ADC har vist seg verdifull i tumordeteksjon, samt i karakterisering og vurdering av behandlingsrespons ved onkologisk avbildning.

I virkeligheten introduserer mikrosirkulasjonen av blod i kapillærnettverket en perfusjonskomponent i tillegg til diffusjon, noe som bidrar til signalforfall i DWI, hvilket har

potensial til å eliminere behovet for DCE-MRI. Bieksponentiell intravoksel inkoherent bevegelse (IVIM)-modellen tar hensyn til denne perfusjonskomponenten. IVIM-modellen gir i tillegg til diffusjonskoeffisienten også perfusjonsrelaterte parametere som den raskere pseudo-diffusjonskoeffisienten D^* (koblet til kapillærstrøm) og perfusjonsfraksjonen f . Tilpasning av IVIM-modellen til diffusjonsdata er imidlertid et utfordrende omvendt problem, først og fremst på grunn av små perfusjonsfraksjoner og det lave signal-til-støyforholdet (SNR) som er iboende til DWI. Dette fører til betydelige unøyaktigheter i perfusjonsrelaterte parametere ved bruk av konvensjonelle tilpasningsmetoder, og hindrer klinisk bruk av IVIM. Følgelig er alternative tilpasningsmetoder nødvendig for å lette den praktiske kliniske implementeringen av IVIM.

Innenfor dette dynamiske feltet for medisinsk bildebehandling har en transformativ revolusjon blitt utløst av den forbedrede datakraften som er opplevd det siste tiåret, og har gitt opphav til moderne metoder og algoritmer drevet av kunstig intelligens (AI) og dyp læring. Disse banebrytende teknikkene gjør det mulig for datamaskiner å oppdage kompliserte mønstre i store datasett. Som svar på disse fremskrittene hadde denne avhandlingen som mål å undersøke bruken av dyp læring for generering av klare, detaljerte bilder av perfusjonsbaserte biomarkører fra DWI, med hovedvekt på å forbedre parameterestimering for IVIM-modellen innen DWI. Innenfor konteksten av dyp læringsparameterestimering, utforsket denne avhandlingen ulike metoder, inkludert ulike læringsstrategier, undersøkelse av ulike trenings- og testdata, utforskning av nettverksarkitekturer og optimalisering av hyperparametere.

Denne avhandlingen er basert på tre artikler. I Paper I ble en tidligere uovervåket vokselvis dyp læring IVIM-tilpasningstilnærming forfinet ved å optimalisere ulike nettverkshyperparametere. Den optimaliserte tilnærmingen adresserte uventede korrelasjoner i den opprinnelige suboptimale uovervåkede tilnærmingen. Som et resultat viste den optimaliserte tilnærmingen forbedret nøyaktighet, uavhengighet og konsistens av både diffusjons- og perfusjonsrelaterte parametere. Dessuten, i simuleringer og in-vivo-data fra pasienter med kreft i bukspyttkjertelen, viste den optimaliserte tilnærmingen overlegen ytelse sammenlignet med toppmoderne IVIM-tilpasningsmetoder, der den viste de mest detaljerte, og betydelig mindre støyende parameterbildene. Spesielt utmerket den seg ved å oppdage de mest signifikante endringene i IVIM-parametre gjennom kjemoradioterapi.

Imidlertid avslørte påfølgende forskning mot Paper II at den optimaliserte, uovervåkede dyplæringstilnærmingen viste dårlig anatomisk generalisering når den ble brukt på hjernen. I mellomtiden viste annen forskning at veiledede dyplæringstilnærminger kan gi resultater treningsdata, noe som også krever etterforskning i sammenheng med IVIM-modellering. Det

er viktig å merke seg at disse overvåkede tilnærmingene er optimalisert ved å bruke parameterverdier som grunnsannhet. Imidlertid mangler estimater av jordsannhetsparametere ofte på grunn av den dårlige karakteren til signalanalyseproblemet i DWI. Følgelig blir disse grunnsannhetsparameterestimater syntetisk generert ved å simulere parameterestimater i henhold til en fordeling, slik som en enhetlig fordeling, eller alternativt utledet fra konvensjonelle estimater anvendt på de originale signaldataene.

Basert på disse funnene, undersøkte Paper II virkningen av sentrale treningsfunksjoner, inkludert effekten av treningsdata og treningslengde, på både uovervåket og overvåket læring for vokselvis dyp læring IVIM parameterestimering. Funnene viste at utvidelse av trening utover «tidlig stopp» kunne adressere parameterkorrelasjoner og redusere feil, og tilby et alternativ til den uttømmende hyperparameteroptimaliseringen av Paper I. Imidlertid resulterte forlenget trening i økt følsomhet for støy for de uovervåkede estimatene, som lignet minste kvadraters estimering. Derimot viste overvåkede estimater en høyere presisjon, men også en merkbar skjevhet mot gjennomsnittet av treningsfordelingen som kan føre til potensielt villedende parameterbilder. Dette førte til konklusjonen at selv om vokselvis dyplæringsbasert modelltilpasning gir løfte for IVIM-parameterestimering, er en nøye evaluering av designvalg og deres innvirkning på tilpasningsytelse og skjevheter avgjørende.

Mens Paper I og Paper II grundig undersøkte vokselvis dyplæringsstilnæringer som behandler vokslar som uavhengige, er vevsmikromiljøet faktisk typisk lokalt homogent. I disse vevene endres ikke mikromiljøegenskaper tilfeldig mellom tilstøtende vokslar. Derfor bør utnyttelse av disse potensielle korrelasjonene mellom relevante signaler i nabovokslar forbedre modellparameter tilpasningen.

I Paper III omfattet undersøkelsen fire delstudier som undersøkte måter å inkorporere slik romlig informasjon for dyp læringsparameterestimering i biofysisk modellering, spesifikt brukt på IVIM-modellen i DWI. Denne studien avslørte at trening overvåket på romlig korrelerte syntetiske data i patcher effektivt utnytter romlig informasjon og fører til tilsynelatende redusert støyfølsomhet, i likhet med å ta signalgjennomsnittet. Dette resulterte i forbedret estimatørnøyaktighet og redusert iboende overvåket skjevhet. Spesielt ble ingen merkbare forbedringer observert for uovervåket læring. Dessuten overgikk oppmerksomhetsmodeller (transformatorer) konvolusjonsbaserte nettverk for dette formålet. Den nylig foreslåtte nabolagsoppmerksomheten tillot trening på større mottakelige felt enn konvensjonell selvoppmerksomhet, noe som førte til forbedret estimatorytelse. Disse forbedringene ble kvantitativt demonstrert i nye fraktal-støy-parameterbilder som viste en romlig korrelert

grunnsannhet. Kvalitative funn i in-vivo-hjerne-DWI-data var stort sett sammenlignbare med de kvantitative evalueringene i simuleringer. I tillegg demonstrerte studien ytterligere måter å forbedre metoden på ved å utnytte tilleggsinformasjon fra testsettet, inkludert underliggende romlig variasjon og underliggende parameterverdifordelinger. Disse lovende tilnærmingene har potensial til å bli utvidet til enhver biofysisk modell brukt på signaldata, og strekker seg utover omfanget av MR.

Som konklusjon har denne doktorgradsavhandlingen betydelig fremmet feltet dyplæringsdiffusjonsparameterestimering. Disse bidragene markerer et betydelig skritt mot den praktiske implementeringen av komplekse diffusjonssignalanalyseproblemer, som IVIM, i kliniske omgivelser.

Kandidat: Misha Pieter Thijs Kaandorp

Institutt: Institutt for sirkulasjon og bildediagnostikk

Veiledere: Peter T. While (hovedveileder), Frank Zijlstra (biveileder)

Finansiering: Norges forskningsråd

Summary

Deep learning diffusion parameters from magnetic resonance imaging

Cancer is a prominent cause of premature mortality worldwide and often preventable through early detection. Magnetic resonance imaging (MRI) is a vital modality in the investigation of many types of cancers, due to its strong sensitivity to soft tissue contrast. Beyond this capability, MRI can provide quantitative or parametric information about biophysical tissue properties and microstructural processes, including molecular diffusion and tissue perfusion. Utilizing appropriate mathematical models for MRI data enables the extraction of these quantitative characteristics through a process known as model fitting.

Today, the conventional MRI approach for assessing tissue perfusion characteristics for cancer detection and treatment monitoring is dynamic contrast-enhanced MRI (DCE-MRI). However, the acquisition of DCE-MRI data necessitates the injection of a contrast agent, typically gadolinium. This approach may pose contraindications for individuals with renal insufficiency, as well as those who are pregnant or breastfeeding. Additionally, recent findings indicating gadolinium deposition in the brain raise concerns about the desirability of this approach.

As an adjunct to DCE-MRI, diffusion-weighted MRI (DWI) has been shown to provide increased specificity in cancer diagnosis. DWI utilizes the random motion of water molecules, also known as diffusion, to generate image contrast. By acquiring images at multiple diffusion-weightings (b values), a diffusion model can be applied to the signal decay for each voxel. Fitting the classical mono-exponential model to the DWI data yields the apparent diffusion coefficient (ADC), a metric reflecting water diffusivity in the observed tissue. ADC has proven valuable in tumor detection, characterization, and assessing treatment response in oncological imaging.

In reality, the microcirculation of blood in the capillary network introduces a perfusion component alongside diffusion, contributing to signal decay in DWI, which has the potential to

obviate the need for DCE-MRI. The biexponential intravoxel incoherent motion (IVIM) model takes this perfusion component into account. The IVIM model provides besides the diffusion coefficient also perfusion-related parameters such as the faster pseudo-diffusion coefficient D^* (linked to capillary flow) and the perfusion fraction f . However, fitting the IVIM model to diffusion data is a challenging ill-posed inverse problem, primarily due to small perfusion fractions and the low signal-to-noise ratio (SNR) inherent to DWI. This leads to substantial inaccuracies in perfusion-related parameters when using conventional fitting approaches, impeding the clinical adoption of IVIM. Consequently, alternative fitting methods are necessary to facilitate the practical clinical implementation of IVIM.

Within this dynamic field of medical imaging, a transformative revolution has been sparked by the enhanced computing power experienced over the past decade, giving rise to modern methodologies and algorithms driven by artificial intelligence (AI) and deep learning. These cutting-edge techniques enable computers to discover complicated patterns in large data sets. In response to these advancements, this thesis aimed to investigate the application of deep learning for the generation of clear, detailed maps of perfusion-based biomarkers from DWI, with a primary emphasis on enhancing parameter estimation for the IVIM model within DWI. Within the context of deep learning parameter estimation, this thesis explored various methodologies, including diverse learning strategies, examination of various training and test data, exploration of network architectures, and optimization of hyperparameters.

This thesis is based on three papers. In Paper I, a prior unsupervised voxelwise deep learning IVIM fitting approach was refined by optimizing various network hyperparameters. The optimized approach successfully addressed unexpected correlations in the original suboptimal unsupervised approach. As a result, the optimized approach exhibited improved accuracy, independence, and consistency of both diffusion and perfusion-related parameters. Moreover, in simulations and in vivo data from pancreatic cancer patients, the optimized approach demonstrated superior performance compared to state-of-the-art IVIM fitting methods, where it showed the most detailed and significantly less noisy parameter maps. Notably, it excelled in detecting the most significant changes in IVIM parameters throughout chemoradiotherapy.

However, subsequent research towards Paper II revealed that the optimized unsupervised deep learning approach exhibited poor anatomy generalization when applied to the brain. Meanwhile, other research demonstrated that supervised deep learning approaches may exhibit training data bias, which also warranted investigation in the context of IVIM modeling. It is important to note that these supervised approaches are optimized utilizing parameter values as

ground truth. However, ground truth parameters estimates are often lacking due to the ill-posed nature of the signal analysis problem in DWI. Consequently, these ground truth parameter estimates are synthetically generated by simulating parameter estimates according to a distribution, such as a uniform distribution, or alternatively, derived from conventional estimators applied to the original signal data.

Based on these findings, Paper II explored the impact of key training features, including the effect of training data and training length, on both unsupervised and supervised learning for voxelwise deep learning IVIM parameter estimation. The findings showed that extending training beyond early stopping could address parameter correlations and reduce errors, offering an alternative to the exhaustive hyperparameter optimization of Paper I. However, prolonged training resulted in increased sensitivity to noise for the unsupervised estimates, which resembled those obtained by least squares fitting. In contrast, supervised estimates displayed a higher precision, but also a notable bias towards the mean of the training distribution that could lead to potentially deceptive parameter maps. This led to the conclusion that while voxelwise deep-learning-based model fitting holds promise for IVIM parameter estimation, a careful evaluation of design choices and their impact on fitting performance and biases is essential.

While Paper I and Paper II extensively investigated voxelwise deep learning approaches that treat voxels as independent, it is essential to recognize that the tissue microenvironment is typically locally homogeneous. In these tissues, microenvironmental properties do not change randomly between adjacent voxels. Therefore, leveraging these potential correlations between relevant signals in neighboring voxels should enhance model-parameter fitting.

In Paper III, the investigation encompassed four sub-studies exploring means to incorporate such spatial information for deep learning parameter estimation in biophysical modeling, specifically applied to the IVIM model in DWI. This study revealed that training supervised on spatially-correlated synthetic data in patches effectively leverages spatial information with apparently reduced noise sensitivity, akin to signal averaging. This resulted in improved estimator accuracy and decreased inherent supervised bias. Notably, no noticeable improvements were observed for unsupervised learning. Moreover, attention models (transformers) outperformed convolution-based networks for this purpose. The recently proposed neighborhood-attention permitted training on larger receptive fields than conventional self-attention, leading to improved estimator performance. These improvements were quantitatively demonstrated in novel fractal-noise parameter maps that provided spatially-correlated ground truth. Qualitative *in vivo* findings in brain DWI data were broadly

comparable to the quantitative evaluations in simulations. Additionally, the study demonstrated further means to enhance the method by leveraging additional information from the test set, including underlying spatial variation and underlying parameter value distributions. These promising approaches have the potential to be extended to any biophysical model applied to signal data, extending beyond the scope of MRI.

In conclusion, this doctoral thesis has substantially advanced the field of deep learning diffusion parameter estimation. These contributions mark a substantial step towards the practical implementation of complex diffusion signal analysis problems, like IVIM, in clinical settings.

Symbols and Abbreviations

^{13}C	Carbon-13
^{15}N	Nitrogen-15
^1H	Hydrogen-1
1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
ADC	Apparent diffusion coefficient
AI	Artificial Intelligence
b	diffusion-weighting
B_0	Main magnetic field strength
B_1	Strength of temporarily applied radiofrequency pulse
CNN	Convolutional neural network
CRT	chemoradiotherapy
CV_{NET}	Coefficient of variation of the networks
D	Diffusion coefficient
D^*	Pseudo-diffusion coefficient
DCE	Dynamic contrast-enhanced
DNN	Deep neural network
DTI	Diffusion tensor imaging
DWI	Diffusion-weighted imaging
ETK	Extended Tofts-Kety
f	Perfusion fraction
FID	Free induction decay
G	Gradient strength
ISMRM	International Society for Magnetic Resonance in Medicine
IVIM	Intravoxel incoherent motion
$\text{IVIM-NET}_{\text{optim}}$	Optimized unsupervised approach proposed in Paper I
$\text{IVIM-NET}_{\text{orig}}$	Original unsupervised approach proposed by Barbieri et al. in 2020
K	Key
LSQ	Least squares
M_t	Non-zero net magnetization vector
M_{xy}	Non-zero net transverse (x - y) magnetization vector
M_z	Non-zero net longitudinal (z) magnetization vector
M_0	Non-zero net magnetization vector under thermal equilibrium
MLP	Multi-layer perceptron
MR	Magnetic resonance
MRI	Magnetic resonance imaging
MRM	Magnetic Resonance in Medicine
MSE	Mean squared error
NLP	Natural language processing
NMR	Nuclear magnetic resonance
NODDI	Neurite orientation dispersion and density imaging
NRMSE	Normalized root-mean-square error
ODD	Out-of-distribution
PDAC	Pancreatic ductal adenocarcinoma
PGSE	Pulsed gradient spin-echo
PNR	Parameter-to-noise ratio

Q	Query
RF	Radio frequency
RMSE	Root-mean-square error
ROI	Region of interest
ReduceLRonPlateau	Reduce learning rate on plateau
S_0	Diffusion signal at zero diffusion-weighting ($b = 0$)
$S(b)$	Diffusion signal at certain diffusion-weighting
SMT	Spherical mean technique
SNR	Signal-to-noise ratio
Swin	Hierarchical vision transformer using shifted windows
t	Time
T_1	Longitudinal spin-lattice
T_2	Transversal spin-spin
TE	Echo time
TR	Repetition time
V	Value
ViT	Vision transformer
WMTI	White matter tissue integrity
MDAPE	Median absolute percentage error
MDPE	Median percentage bias
MAPE	Mean absolute percentage error
wSD	Within-subject standard deviation
ω_0	Larmor frequency
θ	Set of parameters
γ	Gyromagnetic ratio
ρ	Spearman's correlation
δ	Duration of the diffusion gradient
Δ	Time between the two diffusion gradients

List of papers

Paper I

Improved unsupervised physics-informed deep learning for intravoxel incoherent motion modeling and evaluation in pancreatic cancer patients

Misha Pieter Thijs Kaandorp, Sebastiano Barbieri, Remy Klaassen, Hanneke W. M. van Laarhoven, Hans Crezee, Peter T. While, Aart J. Nederveen, Oliver J. Gurney-Champion

Magnetic Resonance in Medicine, 86(4), 2250-2265.

Paper II

Deep learning intravoxel incoherent motion modeling: Exploring the impact of training features and learning strategies

Misha Pieter Thijs Kaandorp, Frank Zijlstra, Christian Federau, Peter T. While

Magnetic Resonance in Medicine, 90(1), 312-328.

Paper III

Incorporating spatial information in deep learning parameter estimation with application to the intravoxel incoherent motion model in diffusion-weighted MRI

Misha Pieter Thijs Kaandorp, Frank Zijlstra, Davood Karimi, Ali Gholipour, Peter T. While

Submitted to Medical Image Analysis (Dec. 2023)

List of conference presentations

1. (*submitted*) **Kaandorp**, Misha P. T., Zijlstra, F., Karimi, D., Gholipour, A., While, P. T. Neighborhood-attention models for incorporating spatial information in deep learning parameter estimation applied to IVIM. *Submitted (Nov. 2023) to the 2024 Annual Meeting of the ISMRM*, Singapore.
2. (*submitted*) Kaushik, S., Zijlstra, F., **Kaandorp**, Misha P. T., While, P. T. Deep-learning reconstruction of under-sampled readout-segmented echo-planar diffusion-weighted images at multiple b values. *Submitted (Nov. 2023) to the 2024 Annual Meeting of the ISMRM*, Singapore.
3. (*acceptance as oral*) **Kaandorp**, Misha P. T., Zijlstra, F., Karimi, D., Gholipour, A., While, P. T. Incorporating spatial information in deep learning parameter estimation applied to IVIM. *Submitted (Dec. 2023) to the 2024 ISMRM Workshop on Moving Forward with Intravoxel Incoherent Motion Modeling for Diffusion-Weighted MRI: An Attempt at Consensus*, Erlangen, Germany.
4. (*acceptance as poster*) Kaushik, S., Zijlstra, F., **Kaandorp**, Misha P. T., While, P. T. Accelerating readout-segmented EPI for IVIM using deep learning. *Submitted (Dec. 2023) to the 2024 ISMRM Workshop on Moving Forward with Intravoxel Incoherent Motion Modeling for Diffusion-Weighted MRI: An Attempt at Consensus*, Erlangen, Germany.
5. (*poster*) **Kaandorp**, Misha P. T., Zijlstra, F., Karimi, D., Gholipour, A., While, P. T. (2023). Attention models for incorporating spatial information in deep learning parameter estimation applied to intravoxel incoherent motion. *Proceedings of the Eur. Soc. for Magnetic Resonance in Medicine and Biology*, Basel, Switzerland.
6. (*power pitch & poster*) **Kaandorp**, Misha P. T., Zijlstra, F., Karimi, D., Gholipour, A., While, P. T. (2023). Incorporating spatial information in deep learning parameter estimation using transformer neural networks applied to intravoxel incoherent motion. *Autumn Research School in Artificial Intelligence Methods in Medical Imaging*, Sommarøy, Norway.

7. (oral) **Kaandorp**, Misha P. T., Zijlstra, F., While, P. T. (2023). Synthetic data shows the potential of unsupervised and supervised learning for incorporating spatial information in IVIM fitting. *Proceedings of the Int. Soc. for Magnetic Resonance in Medicine*, Toronto, Canada.
8. (power pitch & poster) **Kaandorp**, Misha P. T., Zijlstra, F., While, P. T. (2022). The impact of deep learning training features & learning strategies for IVIM fitting. *Proceedings of ISMRM Workshop on Diffusion MRI: From Research to Clinic*, Amsterdam, The Netherlands.
9. (power pitch & poster) **Kaandorp**, Misha P. T., Zijlstra, F., While, P. T. (2022). The impact of training features and strategies for deep learning IVIM fitting. *Microstructure imaging meets Machine Learning*, London, United Kingdom.
10. (poster) **Kaandorp**, Misha P. T., Zijlstra, F., de Almeida Martins, J. P., While, P. T. (2022). The impact of learning rate, network size, and training time on unsupervised deep learning for intravoxel incoherent motion (IVIM) model fitting. *Proceedings of the Int. Soc. for Magnetic Resonance in Medicine*, London, United Kingdom.
11. (poster) **Kaandorp**, Misha P. T., Zijlstra, F., While, P. T. (2022). The impact of training features and strategies for deep learning IVIM fitting. *Proceedings of the Int. Soc. for Magnetic Resonance in Medicine: Benelux Chapter*, Maastricht, The Netherlands.
12. (oral) **Kaandorp**, Misha P. T., Barbieri, S., Klaassen, R., van 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. *Proceedings of the Int. Soc. for Magnetic Resonance in Medicine*, Montréal, QC Canada, Online meeting.
13. (oral) **Kaandorp**, Misha P. T., Barbieri, S., Klaassen, R., van 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. *Proceedings of the Int. Soc. for Magnetic Resonance in Medicine: Benelux Chapter*, Online meeting.

Contents

1. Introduction	19
1.1 Introduction to cancer	19
1.2 Magnetic resonance imaging (MRI)	21
1.2.1 Fundamental principles of nuclear magnetic resonance (NMR)	21
1.2.2 Signal generation and relaxation	23
1.2.2.1 Longitudinal spin-lattice (T_1) relaxation	23
1.2.2.2 Transversal spin-spin (T_2) relaxation	24
1.2.2.3 Free induction decay (FID)	25
1.2.3 Spatial encoding and image formation	26
1.2.4 MRI pulse sequences	28
1.2.4.1 Spin-echo pulse sequence	28
1.2.4.2 Gradient-echo sequence	29
1.2.4.3 T_1 - and T_2 -weighted images	30
1.2.5 Quantitative MRI	31
1.2.5.1 Dynamic contrast-enhanced MRI (DCE-MRI)	31
1.3 Diffusion-weighted imaging (DWI)	34
1.3.1 Brownian motion and free diffusion	34
1.3.2 The ‘basic’ diffusion sequence	34
1.3.3 The apparent diffusion coefficient (ADC)	36
1.3.4 The mono-exponential ADC Signal Model	37
1.3.5 Advanced diffusion models	38
1.4 Intravoxel incoherent motion (IVIM)	40
1.4.1 The IVIM model	40
1.4.2 Fitting methods for IVIM	42
1.4.2.1 One-step nonlinear least-squares	42
1.4.2.2 Two-step nonlinear least-squares or segmented fitting	42
1.4.2.3 Bayesian interference	43
1.5 Artificial Intelligence	45
1.5.1 Basic principles of deep neural networks	45
1.5.2 Types of network architectures	48
1.5.2.1 Multi-layer perceptron (MLP)	49
1.5.2.2 Convolutional neural networks (CNNs)	49
1.5.2.3 Transformers	50

1.5.3	Types of learning strategies	54
1.5.3.1	Supervised learning.....	54
1.5.3.2	Unsupervised learning	55
1.5.3.3	Semi-supervised learning.....	56
1.5.4	Hyperparameters.....	56
1.5.4.1	Training hyperparameters: learning rate, optimizer, batch size	56
1.5.4.2	Network hyperparameters: hidden units, hidden layers, activation function	57
1.5.4.3	Regularization: batch normalization, dropout, early stopping	57
2.	Objectives	60
3.	Summary of papers	62
3.1.	Paper I.....	63
3.2.	Paper II.....	67
3.3.	Paper III	72
4.	Discussion	83
4.1	Learning strategies for deep learning parameter estimation.....	84
4.2	Training and test data for deep learning parameter estimation	86
4.3	Network architectures for deep learning parameter estimation.....	90
4.4	Hyperparameters for deep learning parameter estimation.....	93
4.5	Communication, exploitation, and reproducibility.....	97
4.6	Research ethics.....	98
5.	Conclusion	101
6.	References	105

Chapter 1

Introduction

1.1 Introduction to cancer

Cancer comprises a collection of diseases characterized by irregular cell proliferation, capable of infiltrating or metastasizing to distant body regions. Due to genetic and epigenetic alterations (mutations) during cell proliferation, abnormal cells emerge and form a mass called a tumor, disrupting the body homeostasis (1). The molecular and cellular features which ensures the survival, proliferation and metastasis of cancer cells are referred to as the hallmarks of cancer in ‘The hallmarks of cancer’ by Hanahan and Weinberg (2–4). Currently (4), the core hallmarks comprise: evading growth suppressors, avoiding immune destruction, enabling replicative immortality, tumor promoting inflammation, activating invasion and metastasis, inducing or accessing vasculature, genome instability and mutation, resisting cell death, deregulating cellular metabolism, and sustaining proliferative signaling. In this latest elaboration of this concept (4), Hanahan proposed two additional emerging hallmarks: ‘unlocking phenotypic plasticity’ and ‘non-mutational epigenetic reprogramming’, along with two additional enabling characteristics: ‘polymorphic microbiomes’ and ‘senescent cells’. Apart from the cancer cells themselves, tumors contribute to a complex tumor microenvironment consisting of its own network of blood vessels, diverse cell types, extracellular matrix components, and signaling molecules. This intricate milieu promotes tumor functionality and ensuring its survival, and can vary substantially across distinct tumor types (5).

Cancer is a leading cause of premature death worldwide. In the year 2020, there were 9.9 million recorded cancer-related deaths, and an estimated 19.3 million new cancer cases emerged (6), with this incidence steadily rising every year. The worldwide cancer burden is projected to reach 28.4 million cases by the year 2040, marking a substantial 47% increase when compared to 2020, with mortality rates anticipated to remain relatively consistent (6).

Despite the intricate nature of cancer and the rising global incidence rates, mortality rates have shown a decline in recent years within highly developed countries in Europe and America (7,8). This positive trend can be attributed to advancements in screening, early detection, diagnosis, and treatment monitoring.

1.2 Magnetic resonance imaging (MRI)

Magnetic resonance imaging (MRI) is a powerful non-invasive imaging modality that produces detailed visualizations of human anatomy, along with insights into functional and physiological processes, all without subjecting individuals to ionizing radiation. MRI is typically employed for examining soft tissue conditions, due to its strong sensitivity to various tissue characteristics. These attributes establish MRI as an indispensable tool for both screening and diagnostic purposes. Furthermore, MRI offers many possibilities for treatment monitoring of various medical conditions, including cancer.

This section outlines the fundamental principles of nuclear magnetic resonance (NMR) and MRI, the generation of MRI signal and relaxation mechanisms, and the principles of spatial encoding and image formation. Additionally, it provides a brief introduction to quantitative MRI techniques, with specific mention of dynamic contrast-enhanced MRI (DCE-MRI). The content provided in this section is primarily based on the book 'MRI in practice' by C. Westbrook and J. Talbot (9), unless otherwise specified.

1.2.1 Fundamental principles of nuclear magnetic resonance (NMR)

MRI relies on the foundational principles of nuclear magnetic resonance (NMR) first described by Felix Bloch and Edward Purcell in 1946 (10,11). They observed that after applying a strong external magnetic field, the energy of atomic nuclei can be increased when absorbing radio waves with the same frequency in a magnetic field. Subsequently, as the atomic nuclei return to their initial energy state, they emit radio waves that can be detected. The frequency of the emitted radio waves is a specific characteristic of atomic nuclei. These groundbreaking findings earned Bloch and Purcell the Nobel Prize in Physics in 1952. In the 1970s, Peter Mansfield (12) and Paul Lauterbur (13) made crucial achievements in applying the principles of NMR to create two-dimensional (2D) images by introducing gradients in the magnetic field, which served as the basis for MRI. Both Mansfield and Lauterbur were honored with the Nobel Prize in 2003 for providing the foundation for the development of magnetic resonance into a valuable clinical imaging technique, known today as MRI.

The fundamental principles of NMR and MRI are rooted in the quantum mechanical characteristics of atomic nuclei, which exhibit angular momentum or spinning motion, known as spin. Atomic nuclei are composed of protons and neutrons that have a spin of $\frac{1}{2}$ or $-\frac{1}{2}$, with their spins oriented in opposite directions. Nuclei with an odd number of protons and neutrons cause a non-zero net spin, resulting in a magnetic moment. These are known as magnetic

resonance (MR) active nuclei that can be utilized in MRI. Important MR-active nuclei are hydrogen-1 (^1H), Carbon-13 (^{13}C), and Nitrogen-15 (^{15}N). ^1H atoms possess only a single proton and are widely represented in biological tissues. Therefore, ^1H is particularly relevant in MRI of the human body.

In the absence of an applied magnetic field, the spins of atomic nuclei are randomly oriented, as depicted in Figure 1.1A. When these atomic nuclei are placed within an MR scanner featuring a strong magnetic field B_0 , typically measured in Tesla (1.5T, 3T, 7T), the spins start to align either parallel or anti-parallel to the B_0 , as illustrated in Figure 1.1B. According to the thermal equilibrium theory, an increased number of magnetic moments will align parallel to B_0 . In this context, the net magnetization M_t will have an equilibrium M_0 aligned with B_0 . The direction of B_0 is generally referred to as the longitudinal direction or z -axis, while the plane perpendicular to it is termed the transverse (x - y) plane.

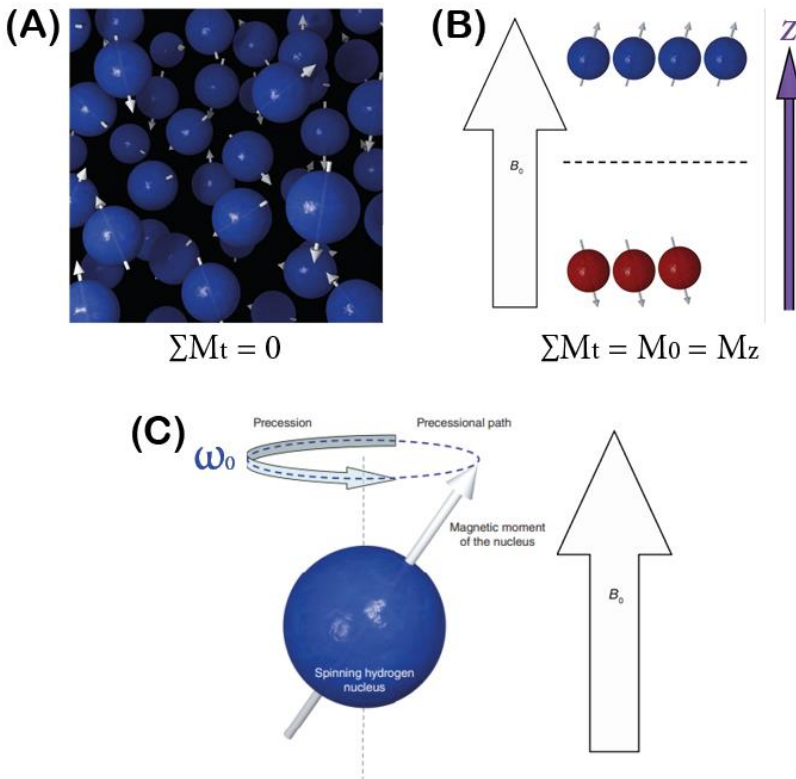


Figure 1.1: (A) Random alignment of the magnetic moments of the atomic nuclei when no external field B_0 is applied. (B) In the presence of B_0 , the magnetic moments align parallel or antiparallel to the magnetic field. In accordance with the thermal equilibrium theory, an elevated quantity of magnetic moments will orient themselves in parallel to B_0 . (C) Spin precession around B_0 (z -axis) at the Larmor frequency ω_0 . This figure is adapted and modified from (9) with permission of John Wiley and Sons.

With the application of the magnetic field, the spins align and initiate rotation within a cone centered around the z -axis with a net longitudinal magnetization component M_z . This spinning rotation is known as spin precession (Figure 1.1C). The rate of rotation is termed the Larmor frequency ω_0 and is directly proportional to the strength of the magnetic field, as described by the following equation:

$$\omega_0 = \gamma B_0, \quad (1)$$

where γ is the gyromagnetic ratio and depends on the nucleus. The γ of ^1H is 42.58 MHz/T.

1.2.2 Signal generation and relaxation

In the MR scanner, the process of generating a signal, which is later transformed into an image, begins with disrupting the equilibrium M_0 state. This disruption is achieved by delivering a radio frequency (RF) excitation pulse that is in resonance with the Larmor frequency. This perturbation leads to an absorption of energy by the protons, causing M_t to deviate from its alignment with B_0 . Following the RF excitation pulse, the magnetization vector will be at an angle relative to the longitudinal z -axis, which is referred to as the flip angle. A common excitation pulse is the 90° RF excitation pulse, aligning M_t with the transverse (x - y) plane, resulting in the complete loss of M_z and the establishment of the net transverse magnetization component M_{xy} . This 90° RF excitation pulse produces the highest MR signal. Another frequently employed RF pulse is the 180° RF pulse, often termed the 'inversion pulse,' as it completely reverses the longitudinal component. It is common to use both 90° and 180° RF pulses together in pulse sequences to create diverse tissue contrasts, as discussed in Section 1.2.4.

After the application of the RF pulse, the excited spins will try to return to their original energy state, aligning with B_0 , through a precession movement. This process is termed relaxation and occurs by two distinctive mechanisms operating simultaneously: longitudinal spin-lattice (T_1) and transversal spin-spin (T_2) relaxation.

1.2.2.1 Longitudinal spin-lattice (T_1) relaxation

T_1 relaxation is the result of the dissipation of energy exchange between protons and their surrounding lattice molecules. This energy dissipation leads to the restoration of the longitudinal component M_z to its equilibrium. The recovery process is expressed by an exponential function:

$$M_z(t) = M_0 \left(1 - e^{-\frac{t}{T_1}} \right), \quad (2)$$

where T_1 represents the T_1 recovery time, which is the time it takes for M_z to recover by 63% within a tissue. The general shape of T_1 relaxation is illustrated in Figure 1.2A.

1.2.2.2 Transversal spin-spin (T_2) relaxation

T_2 relaxation occurs due to the loss of M_{xy} , resulting from the dephasing of spins, where spins give their energy from the excitation pulse to neighboring spins. This decay is expressed by:

$$M_{xy}(t) = M_0 e^{-\frac{t}{T_2}}, \quad (3)$$

where t is time and T_2 represents the T_2 decay time, which is the time it takes for M_{xy} to dephase by 63%, leaving only 37% of the initial magnetization in the tissue. The general shape of T_2 relaxation is illustrated in Figure 1.2B. Both T_1 and T_2 times are intrinsic contrast parameters that are specific to the type of tissue, see Table 1.

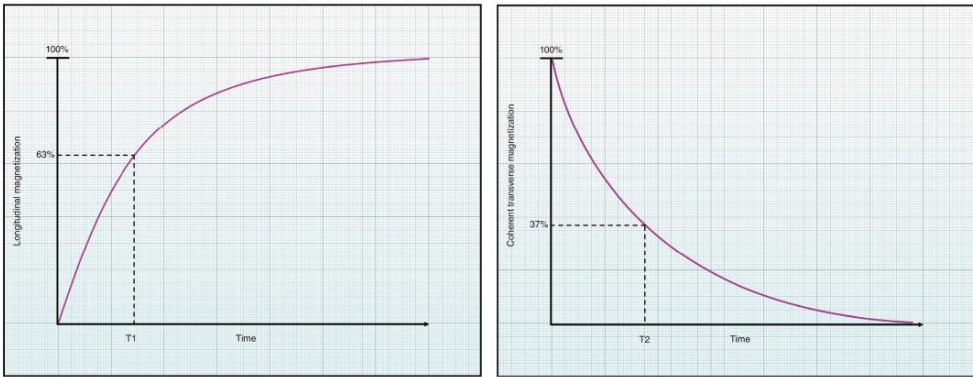


Figure 1.2: (left) T_1 relaxation curve. The dotted line represents the T_1 recovery time where 63% of the net longitudinal magnetization component M_z has been recovered. (right) T_2 relaxation curve. The dotted line represents the T_2 decay time where 37% of the net transverse magnetization M_{xy} has been recovered. This figure is adapted and modified from (9) with permission of John Wiley and Sons.

Table 1: Typical T_1 recovery time and T_2 decay times at 1T. This table is adapted and modified from (9) with permission of John Wiley and Sons.

Tissue	T_1 recovery time (ms)	T_2 decay time (ms)
Water	2500	2500
Fat	200	100
CSF	2000	300
White matter	500	100

1.2.2.3 Free induction decay (FID)

Upon excitation, the M_{xy} induces a current voltage in the receiving coil of the MR scanner. This current voltage is the signal that will later be converted into an image. The T_2 decay leads to a decrease in this current voltage, known as free induction decay (FID), see Figure 1.3. The FID is caused by loss of precession of M_{xy} as it spirals up from the transverse x - y plane to its original alignment with B_0 . In an ideal scenario, only spin-to-spin interactions affect T_2 decay, but in reality, local field inhomogeneities in B_0 affect the original T_2 decay, accelerating the decay process. Therefore, the FID is affected by both spin-to-spin T_2 decay and T_2 decay from local field inhomogeneities (T_2'). This combined decay is expressed as T_2^* :

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2'}, \quad (4)$$

which is effectively evaluated through the FID curve from the MRI scanner.

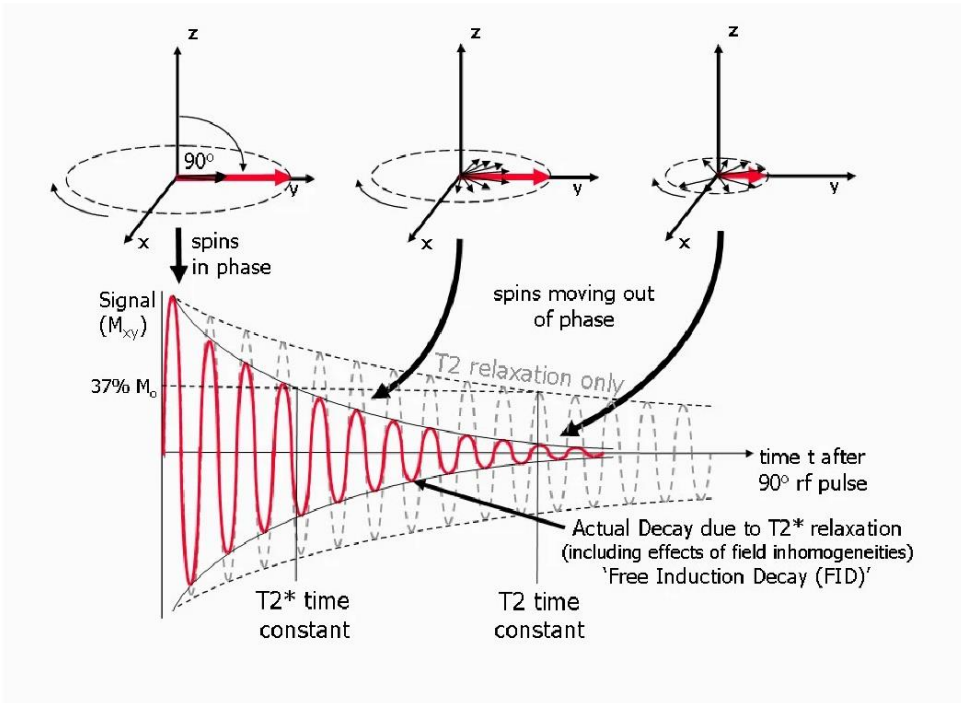


Figure 1.3: The free induction decay (FID). After the application of a 90° excitation pulse, the M_{xy} has a maximum amplitude as all the spins rotate in phase in this plane. During the relaxation phase, the amplitude of M_{xy} decays as the magnetic moments move out of phase with each other. The resulting signal decay is the FID. The FID is characterized by both T_2 decay and additional dephasing caused by local field inhomogeneities (T_2'), resulting in an overall T_2^* signal decay that is the signal decay of the FID. This figure is adapted and modified from (14), CC-BY-2.0.

1.2.3 Spatial encoding and image formation

In the MR scanner, the RF excitation pulse induces resonance in the nuclei resulting in the generation of transverse magnetization M_{xy} . The transverse magnetization induces a signal in the receiver coils that is also positioned in the transverse x - y plane. This signal takes the form of an alternating voltage with a frequency matching the Larmor frequency, regardless of the origin of the signal. Since all magnetic moments precess at the same Larmor frequency within a uniform magnetic field, the signal also oscillates at this frequency. Consequently, the signal lacks spatial information and must undergo spatial encoding to create an image.

Signal localization is achieved through the application of gradients in gradient coils. These gradients are applied in a linearly sloped manner to control the spatial inhomogeneity, resulting in a sloped magnetic field strength in the MR scanner. As denoted from eq. (1), the Larmor frequency ω_0 is directly proportional to B_0 . So, variations in the B_0 also alter the precessional frequency. In order to acquire a 2D image, three gradients are applied to spatially encode the signal: the slice selection gradient, frequency-encoding gradient and phase-encoding gradient. These gradients are illustrated in Figure 1.4A for a spin-echo pulse sequence (a series of RF pulses; see next subsection 1.2.4). Each of these gradients is applied along a different plane or axis (x , y or z).

The slice selection gradient gradually alters the precessional frequencies of the magnetic moments within the selected scan plane, usually the z -axis. A specific slice is selectively excited by applying an RF excitation pulse with a frequency matching the precessional frequencies of magnetic moments within that slice. The thickness of the slice is determined by the bandwidth (frequency range) of the slice selection gradient and the strength of the gradient. A thinner slice necessitates a steeper slope or a narrower transmit bandwidth, which enhances spatial resolution. The slice-rephasing lobe (not shown in Figure 1.4) typically follows the slice-selection gradient to address phase dispersion occurring simultaneously with the application of the main slice-select gradient. The absence of this rephasing lobe would result in intravoxel phase dispersion, resulting in a loss of signal.

After the spins are excited for a given slice, the position of the signal along both axes of the slice needs to be encoded (usually the x - y plane). Before any other gradient is applied, all precessional frequencies are synchronized and in phase with each other. By applying the phase-encoding gradient, the precessional frequencies will change along the axis to which it is applied. As the precessional frequency changes, also the phase along their precessional path changes. Once the phase-encoding gradient is turned off, the magnetic moments exhibit different phase

shift along the direction the phase-encoding is applied, while the precessional frequencies of the magnetic moments along its perpendicular axis remain unchanged. Subsequently, by applying the frequency-encoding gradient along the axis perpendicular to the phase-encoding gradient, the precessional frequencies change along the axis the frequency-encoding gradient is applied. Typically, the longest axis of the anatomy performs frequency-encoding, usually termed the x -axis, while the shorter axis conducts phase encoding, often called the y -axis.

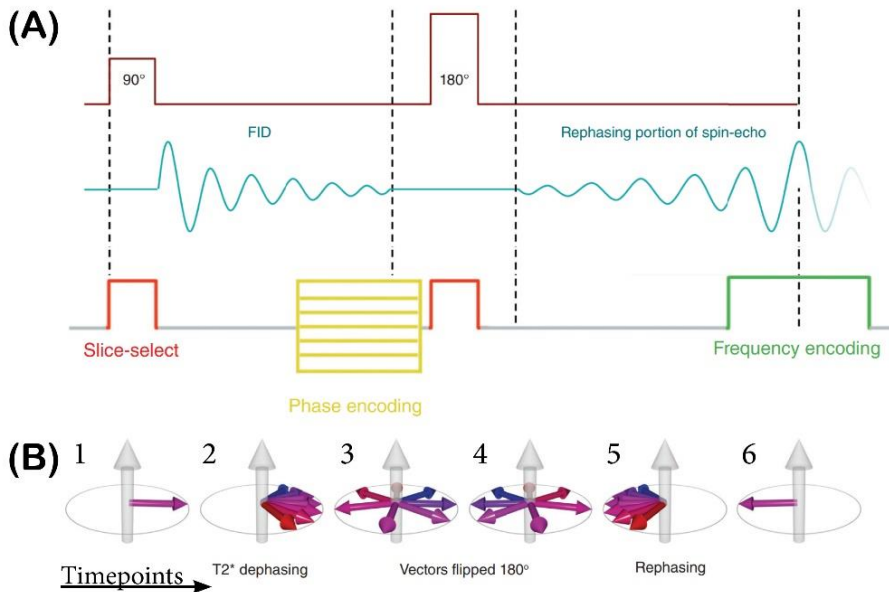


Figure 1.4: (A) Basic spin-echo pulse sequence containing an excitation pulse (90°) and a refocusing pulse (180°) to produce the echo. Corresponding illustrations of the FID and spatially encoding gradients (slice-gradient, phase encoding, and frequency encoding) are also shown. (B) An illustration of the magnetic moments during a spin-echo sequence as described in Section 1.2.4.1 This figure is adapted and modified from (9) with permission of John Wiley and Sons.

Through the application of the phase-encoding gradient and frequency-encoding gradient, the resultant signal received by the receiver coils comprises numerous frequencies that correspond to the spatial location of each pixel within the slice. To acquire a high-resolution image, multiple signals are acquired with slight variations in the amplitude of the phase-encoding gradient throughout the pulse sequence (illustrated by the yellow ladder in Figure 1.4A), determining the degree of phase shift. Typically, in a pulse sequence, the amplitude of the frequency-encoding gradients remains consistent each time it is applied. Subsequently, the generated signals are sampled in the frequency domain, termed k -space (see Figure 1.5). Each horizontal line in k -space represents a different signal containing a different initial phase shift.

Applying the inverse Fourier transformation of the entire k-space produces the MR image (Figure 1.5). Note that every data point within the image encompasses both phase and frequency information derived from the entire slice at a specific moment within the sampling window. This process is repeated for each slice to generate a three-dimensional (3D) image from a set of 2D slices. Each 3D pixel represents a small volume element within the body, denoted as a voxel.

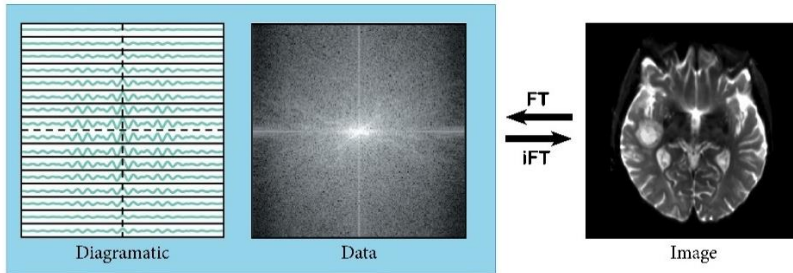


Figure 1.5: An illustration of k-space of a singular slice with an illustrative representation of a brain tumor patient. (*left*) As illustrated in the diagrammatic representation of k-space, the frequency encoding data are positioned along the horizontal axis of k-space, and the data from the phase encoding are positioned along the vertical axis of k-space perpendicular to the frequency axis. (*right*) The application of the 2D inverse (fast) Fourier transform to k-space results in the generation of the final image. This figure is adapted and modified from (9) with permission of John Wiley and Sons.

1.2.4 MRI pulse sequences

MRI pulse sequences comprise a series of sequenced events involving RF pulses and gradients, designed to control the contrast in an image. These pulse sequences are methods to rephase the magnetic moments, which produces a measurable signal, often referred to as an ‘echo’. There are many types of pulse-sequences, but the spin-echo pulse sequence and the gradient echo pulse sequence are considered the fundamental MRI pulse sequences, and often serve as the basis for other MRI pulse sequences.

1.2.4.1 Spin-echo pulse sequence

The spin-echo sequence consists of two RF pulses, a 90° excitation pulse followed by a 180° refocusing pulse, as illustrated in Figure 1.4A. An illustration of the magnetic moments in a pulse sequence is provided in Figure 1.4B. In this sequence, the 90° excitation pulse is used to fully flip the M_t into the transverse plane, as illustrated in timepoint 1 of Figure 1.4B. This induces a voltage in the receiver coil. When the excitation pulse is switched off, the M_t will try to realign with B_0 according to the FID. The T_2^* dephasing due to B_0 field inhomogeneities occurs almost immediately, where some magnetic moments are dephasing faster (leading edge;

red in timepoint 2 of Figure 1.4B) than other slower magnetic moments (trailing edge; *blue* in timepoint 2 of Figure 1.4B). Subsequently, the 180° refocusing pulse is used to compensate for this dephasing. This 180° refocusing pulse flips the individual magnetic moments 180° , as such that the faster dephasing magnetic moments that were the leading edge now form the trailing edge (still *red* in timepoint 5 of Figure 1.4B). Conversely, the slower dephasing magnetic moments that were the trailing edge now form the leading edge (still *blue* in timepoint 5 of Figure 1.4B). Because the precession of the magnetic moments is the same, the trailing edge catches up with the leading edge. Then, when the magnetic vectors have fully rephased in the transverse plane (timepoint 6 of Figure 1.4B), the receiver coils receive maximum voltage, and a specific 'spin-echo' after a specific echo time (TE) occurs. The 180° refocusing pulse is exactly done at half TE (TE/2). Within a pulse sequence, the time interval between two consecutive similar RF pulses is referred to as the repetition time (TR). The TE and TR are specific pulse sequence parameters of a scan protocol that determines the image contrast (see T₁- and T₂-weighted images; Section 1.2.4.3).

Because the T₂' decay from field inhomogeneities is constant and the T₂ decay of tissues is random, the spin-echo effectively compensates for T₂'. As a result, the spin-echo sequence eliminates the effects of local field inhomogeneities on dephasing, yielding information solely about T₂ and not T₂'.

In conventional spin-echo, only one phase-encoding step is executed during each TR, leading to one echo and filling of only one line in k-space. However, in clinical practice, fast or turbo spin-echo pulse sequences (15) are commonly used, where the number of phase-encoding steps per TR is substantially increased. This is achieved by using multiple 180° RF rephasing pulses to generate a series of echoes, known as an echo train. Each echo possesses a distinct amplitude of the phase encoding slope to fill a different line in k-space. Consequently, k-space is filled efficiently per TR, leading to reduced scan times.

1.2.4.2 Gradient-echo sequence

The gradient-echo sequence (Figure 1.6) consists of RF excitation pulses that are variable, typically with a flip angle of 90° or less. Following the first RF excitation pulse, T₂' dephasing due to B_0 field inhomogeneities occurs almost immediately similar as in the spin-echo sequence, and M_t will try to realign with B_0 according to the FID. The frequency-encoding gradient introduces dephasing in the transverse magnetization. To generate a gradient-echo, the same frequency-encoding gradient with opposite polarity is used to rephase the transverse

magnetization. Unlike spin-echo pulse sequences, gradient-echo sequences do not inherently compensate for T_2' magnetic field inhomogeneities.

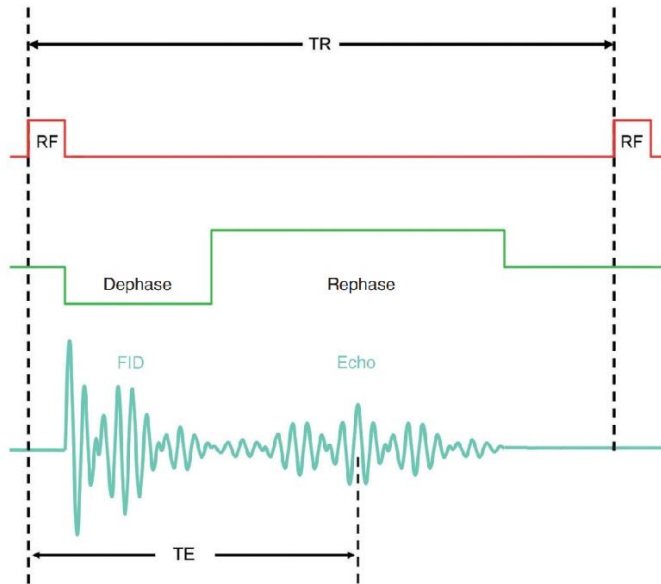


Figure 1.6: Basic gradient-echo pulse sequence, where the frequency-encoding gradient is employed for inducing dephasing and rephrasing to produce an echo. This figure is adapted and modified from (9) with permission of John Wiley and Sons.

1.2.4.3 T_1 - and T_2 -weighted images

As described in Section 1.2.2, the generation of MRI tissue contrast is substantially influenced by the intrinsic relaxation properties of tissues, particularly T_1 (the recovery of longitudinal magnetization M_z) and T_2 (the decay of transverse magnetization M_{xy}). Altering the TR and TE settings on the MR scanner enables the acquisition of distinct T_1 - and T_2 -weightings in an image, each emphasizing different aspects of tissue relaxation characteristics and providing critical information for clinical diagnosis. It is important to note that acquiring only single T_1 - and T_2 -weighted images is limited to qualitative analysis.

To obtain T_1 -weighted images, the TR and TE settings are manipulated to emphasize the contribution of T_1 relaxation. This is done by applying a short TE, which minimizes the influence of T_2 decay, and a short TR, which allows for distinct tissue contrast based on their unique T_1 recovery times. T_1 -weighted imaging is particularly valuable for visualizing anatomical structures. Conversely, for the acquisition of T_2 -weighted images, the MRI protocol involves a long TE and a long TR. The long TE ensures that T_2 decay has a prominent influence on the signal, highlighting differences in T_2 relaxation times among tissues. The long TR allows

for full T_1 recovery, effectively minimizing the contribution of T_1 to the image contrast. Consequently, T_2 -weighted images accentuate variations in tissue T_2 relaxation. T_2 -weighted imaging is particularly sensitive to water content. Consequently, it is a valuable tool for identifying diseases characterized by changes in tissue structure or water content, such as edema, stroke, or brain tumors.

In addition to T_1 - and T_2 -weighted images, a proton density-weighted image can be acquired, where both the influence of T_1 and T_2 relaxation needs to be minimized. This is achieved by a typically short TE and a long TR. The short TE ensures minimal T_2 decay, while the long TR allows sufficient time for complete T_1 recovery.

1.2.5 Quantitative MRI

In addition to the superior ability of MRI to define soft tissue structures and generate detailed anatomical images for qualitative analysis, MRI possesses the ability to provide quantitative or parametric information about biophysical tissue properties and microstructural processes, such as molecular diffusion, tissue perfusion, cell density, and microvessel structure. This field is known as quantitative MRI or multiparametric MRI, and represents a frontier of innovation that holds promise for healthcare (16,17). In quantitative MRI, appropriate mathematical models are employed to MRI signal data acquired through multiple measurements to extract these tissue properties, known as model fitting. This quantitative information can be utilized across a spectrum of healthcare applications, ranging from disease diagnosis, selection of treatments, and treatment monitoring.

Several scan protocols can be used for quantitative MRI. This thesis primarily focuses on diffusion-weighted imaging (DWI), which is comprehensively described in the subsequent section (Section 1.3). For additional context, this section provides a concise overview of another frequently employed quantitative MRI method for tumor characterization: dynamic contrast-enhanced MRI (DCE-MRI). It is worth noting that this thesis does not explore other quantitative MRI approaches, such as T_2 -mapping (18), which is sensitive to tissue water content, and MR Elastography (19,20), a technique for quantifying tissue stiffness by tracking the propagation of mechanical waves within the body. MR Elastography has demonstrated its utility in the diagnosis and assessment of conditions like liver fibrosis and cirrhosis (20).

1.2.5.1 Dynamic contrast-enhanced MRI (DCE-MRI)

Dynamic contrast-enhanced MRI (DCE-MRI) is the primary method for characterizing tissue microvasculature properties and providing perfusion-related parameters for cancer

characterization and monitoring of treatment response in the body (16,21–23). Acquiring DCE-MRI data involves the injection of a contrast agent into the bloodstream, typically utilizing the paramagnetic gadolinium in clinical practice. The presence of gadolinium in the tissues alters the local magnetic field, leading to an accelerated return of nuclear spins to their equilibrium states and a subsequent reduction in T_1 relaxation time. This reduction in T_1 relaxation time translates into heightened signal intensity on T_1 -weighted images, forming the foundation of DCE-MRI. Beyond the critical T_1 shortening effect, variations in the magnetic field induced by gadolinium also contribute to decreased T_2 and T_2^* relaxation times (16,21–23).

In DCE-MRI, a rapid T_1 -weighted imaging sequence is typically employed to capture a dynamic series of images during the passage of the contrast agent through the tissue of interest. This dynamic imaging approach allows for the temporal (in time) tracking of the distribution of the contrast agent. In the context of tumor assessment, the angiogenesis process often results in the development of new vessels characterized by increased permeability (4,24). As a consequence, these vessels exhibit a distinctive 'wash-in' and 'wash-out' pattern in response to the contrast agent. Thus, the temporal behavior in the tumor region on DCE-MRI manifests as a rapid signal increase (16,21–23).

To extract quantitative information from the temporal intensity curve obtained in DCE-MRI, the Extended Tofts-Kety (ETK) model (25) is commonly employed. The ETK model offers valuable insights into the tumor vasculature, blood vessel permeability, and the extravascular/extracellular volume fraction. Of particular interest is the parameter K_{trans} , which represents the volume transfer constant between blood plasma and the extravascular-extracellular space, often referred to as permeability. K_{trans} is particularly useful for characterizing the tumor microenvironment (26) as it differs substantially from normal soft tissue with respect to vascular structure and permeability (4,24).

Despite the strong contrast and great sensitivity to tissue perfusion, DCE-MRI is not without limitations and potential implications. One prominent concern is that the involvement of gadolinium-based contrast agents may be nephrotoxic in patients suffering from renal insufficiency (27–29). Additionally, the use of gadolinium-based contrast agents in DCE-MRI poses a considerable risk for patients who are pregnant or breastfeeding, with potential implications for a range of inflammatory, dermal, rheumatological conditions, as well as the risk of stillbirth or neonatal death (29,30). Recent evidence has also indicated a gadolinium deposition in the brain and other organs, even among individuals with normal renal function (29,31,32). These considerations underscore the importance of exploring non-contrast-

enhanced MRI techniques capable of providing information on tissue micro-environmental characteristics such as perfusion dynamics.

1.3 Diffusion-weighted imaging (DWI)

As an adjunct to DCE-MRI, diffusion-weighted MRI (DWI) has been shown to provide increased specificity in cancer diagnosis by providing quantitative clinical biomarkers (33–36). Unlike DCE-MRI, which relies on invasive contrast agents, DWI utilizes the random motion of water molecules, also known as diffusion, to generate image contrast.

This section describes the principles of Brownian motion and free diffusion, the basic diffusion pulse sequence, the apparent diffusion coefficient (ADC), and the mono-exponential ADC model for DWI data. This section is primarily based on information provided in the book ‘Diffusion MRI: from quantitative measurement to in vivo neuroanatomy’ by H. Johansen-Berg and T. E. Behrens (37), unless otherwise specified.

1.3.1 Brownian motion and free diffusion

On a microscopic level, diffusion is the intermingling of (water) molecules that randomly move according to Brownian motion due to thermal energy within a medium, initially described by Robert Brown in 1827 (38). Brownian motion is the random motion of particles originating from their collision with other fast-moving particles, which is illustrated in Figure 1.7. Using Fick’s law (39,40), Albert Einstein (41,42) established a mathematical formulation of Brownian motion, which can be written as:

$$x^2 = 2Dt, \quad (5)$$

where x^2 represents the mean squared displacement, D represents the diffusion coefficient and t is time. This equation was derived for the one-dimensional (1D) model, but can be generalized into n dimensions by:

$$x^2 = 2nDt. \quad (6)$$

On a macroscopic level in a free medium where there are no restrictions or barriers, there is so-called ‘free diffusion’, where diffusion occurs uniformly in all directions, referred to as ‘isotropic diffusion’. In such circumstances, the movement of diffusing particles follows a Gaussian probability distribution and adheres to the Einstein equations above.

1.3.2 The ‘basic’ diffusion sequence

Modern diffusion measurements in NMR and MRI find their origin in the pioneering work of Stejskal and Tanner in 1965 (43), who introduced the ‘pulsed gradient spin-echo’ (PGSE) sequence. The concept of Stejskal and Tanner is based on the separation of stationary spins from their moving or diffusing counterparts within the magnetic field.

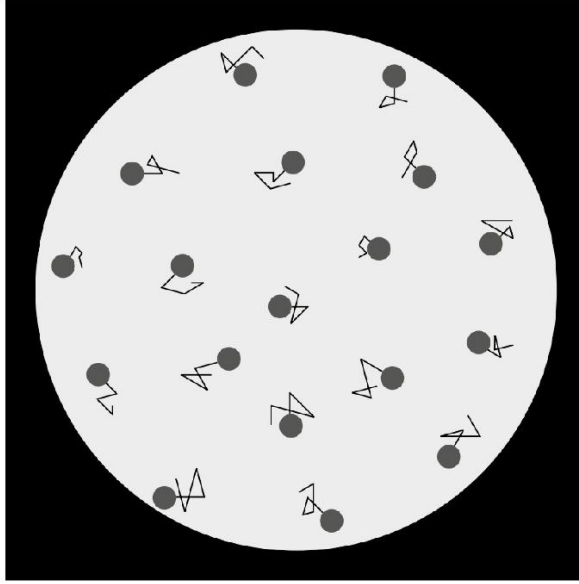


Figure 1.7: Illustration of Brownian motion. Brownian motion is a random, erratic movement of microscopic particles suspended in a fluid, driven by continuous collisions with surrounding fluid molecules. This figure is adapted and modified from (37) with permission of Elsevier.

The PGSE sequence combines two diffusion gradients within a spin-echo framework, consisting of a 90° RF excitation pulse followed by a 180° refocusing pulse, as illustrated in Figure 1.8. These two diffusion gradients enable the differentiation between stationary and moving or diffusing spins within the voxel. The first diffusion gradient is placed before the 180° pulse and induces a phase shift in the spins, setting the stage for the subsequent diffusion-related effects. Following this, the second diffusion gradient is applied after the 180° pulse and serves as a phase reverser that compensates for the dephasing incurred by the static spins. However, for spins or particles that diffuse randomly through the excited volume, the second gradient will not fully reverse the changes induced by the first gradient. This incomplete restoration originates from the phase dispersion to other randomly moving spins, leading to signal attenuation.

The attenuation in the diffusion signal depends on both the magnitude of diffusion and the amount of diffusion weighting, quantified by the b value [s/mm^2] (43). The b value is a tunable parameter that depends on the strength (G) and duration (δ) of the diffusion gradients, and time spacing Δ between the two diffusion gradients. It is expressed by:

$$b = \gamma^2 G^2 \delta^2 \left(\Delta - \frac{\delta}{3} \right), \quad (7)$$

where γ represents the gyromagnetic ratio and depends on the atomic nucleus similar to eq. (1). In addition to the amount of diffusion weighting, the diffusion gradient possesses a directionality. This directionality is particularly relevant in tissues where the diffusion is not uniform in all directions as explained in the subsequent sections.

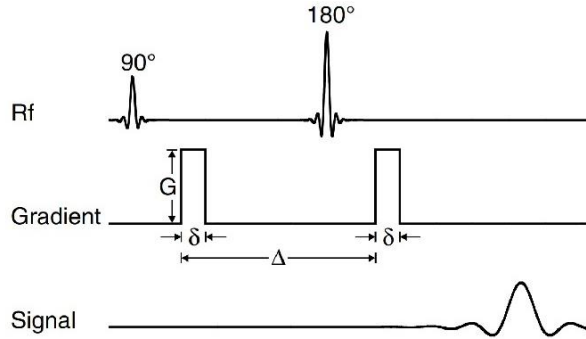


Figure 1.8: A schematic of the pulsed gradient spin-echo (PGSE) sequence introduced by Stejskal and Tanner. This sequence incorporates two diffusion gradients, positioned both before and after the 180° refocusing pulse. The degree of diffusion-weighting, represented by the b value, is contingent on factors such as gradient strength (G), the duration of the diffusion gradient (δ), and the time spacing Δ separating the two diffusion gradients. This figure is adapted and modified from (37) with permission of Elsevier.

1.3.3 The apparent diffusion coefficient (ADC)

In DWI, the rate of diffusion of water molecules within tissues is quantitatively measured by the apparent diffusion coefficient (ADC). The term ‘apparent’ in the ADC is used to indicate that the measured diffusion is not always a direct reflection of the true molecular diffusion of water in tissues. Instead, it is influenced by various factors, such as tissue microstructure, cell membranes, and perfusion effects. The ADC provides valuable insights into the apparent diffusion rate within a given tissue, reflecting the net effect of all diffusion-related interactions within the tissue microenvironment. Hence, DWI allows us to study the underlying microstructure without the use of exogenous contrast agents (33–36,44).

In tissues where the apparent diffusivity remains predominantly consistent across various tissue orientations within a voxel (i.e. demonstrating isotropic behavior), such as gray matter, it is generally sufficient to describe the diffusion properties with a single ADC. Conversely, in biological tissues like white matter, water movement is substantially affected by obstructions and constraints imposed by physiological barriers, such as cell membranes and fibers (Figure 1.9). These factors result in distinct diffusion orientations, giving rise to what is known as ‘anisotropic diffusion.’ In anisotropic diffusion, a singular ADC is insufficient for adequately characterizing the orientation-dependent movement of water.

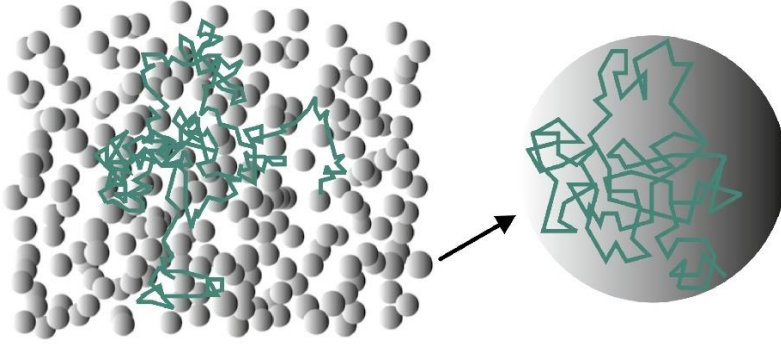


Figure 1.9: (left) An illustration of diffusion that is hindered by biological cells. (right) Inside the biological cells, diffusion may be restricted by the cellular membranes. This figure is adapted and modified from (37) with permission of Elsevier.

In the DWI images, the effect of ADC and anisotropy becomes more pronounced in the images as b values increase, resulting in tissues with elevated diffusion being seen as hypo-intense regions (i.e. low signal). Conversely, tissues with dense cellular packing, for instance tumors, appear as hyper-intense regions (i.e. high signal), exhibiting lower ADC values (34–36).

1.3.4 The mono-exponential ADC Signal Model

To obtain the ADC, a minimum of two measurements is required. Typically, these comprise: one baseline measurement with no diffusion weighting, denoted as $b = 0$ mm²/s and often referred to as S_0 ; and a second measurement $S(b)$ that is acquired at $b > 0$. Under the assumption that the displacement of water molecules is Gaussian, the signal attenuation is expressed by a mono-exponentially decay:

$$S(b) = S_0 e^{-bADC}. \quad (8)$$

The above equation provides the ADC within each voxel in one direction, forming the basis for the generation of quantitative ADC maps. To account for anisotropic diffusion, diffusion gradients are applied in multiple directions with a minimum of three measurements. These diffusion gradients are typically orthogonal, aligning with the three spatial orientations (x, y, z), each yielding directional-specific ADC values:

$$S_{xx}(b) = S_0 e^{-bADC_{xx}}, \quad (9)$$

$$S_{yy}(b) = S_0 e^{-bADC_{yy}}, \quad (10)$$

$$S_{zz}(b) = S_0 e^{-bADC_{zz}}. \quad (11)$$

Subsequently, the individual directional-specific ADC images are combined into a unified set known as the trace image (see Figure 1.10), which provides a composite representation of the

mean diffusivity of the tissues that considers all spatial orientations. Mathematically, the trace image can be described by taking the geometric mean of the individual images obtained from the three orthogonal directions:

$$S_{TRACE}(b) = \sqrt[3]{S_x S_y S_z} = S_0 e^{-b(ADC_{xx} + ADC_{yy} + ADC_{zz})/3} = S_0 e^{-bADC_{trace}}. \quad (12)$$

Importantly, the mono-exponential ADC model characterizes diffusion behavior within tissues for b values typically below 1500 s/mm^2 , it has limitations for higher b values ($b > 1500 \text{ s/mm}^2$).

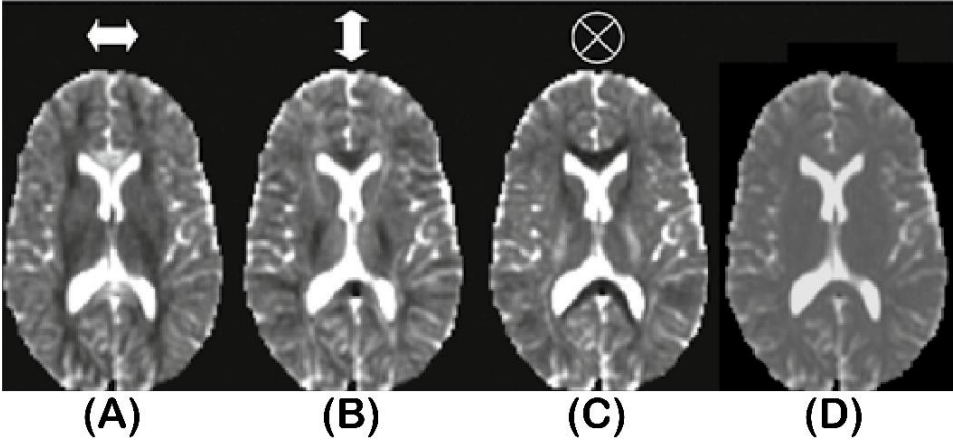


Figure 1.10: The effect of altering the diffusion encoding direction on the Apparent Diffusion Coefficient (ADC). **(A)** ADC map of diffusion encoding applied in the left-right direction. **(B)** ADC map of diffusion encoding applied in the anterior-posterior direction. **(C)** ADC map of diffusion encoding applied with direction in and out of the plane. **(D)** The trace image, showing the mean diffusivity of the three individual ADC images. Note the contrast reduction in the white matter in the trace image compared to the individual ADC maps. This figure is adapted and modified from (37) with permission of Elsevier.

1.3.5 Advanced diffusion models

In addition to the conventional mono-exponential ADC model, multi-compartment diffusion modeling offers the ability to characterize not only the ADC but also other properties such as cell size, axon diameter, orientation dispersion, and neurite density. At high b values ($b > 1500 \text{ s/mm}^2$), non-Gaussian diffusion behavior prevails due to substantial interactions among compartments and barriers within tissues, rendering the assumed Gaussian diffusion and the mono-exponential ADC model invalid. To account for these complexities, non-Gaussian Kurtosis diffusion imaging has been proposed (45), which can provide more detailed information about tissue microstructure and can be particularly useful in regions with complex fiber orientations. Non-Gaussian Kurtosis diffusion imaging is beyond the scope of this thesis, as well as other popular biophysical models (mainly applied in the brain) including neurite

orientation dispersion and density imaging (NODDI) (46), spherical mean technique (SMT) (47), and white matter tissue integrity (WMTI) (48).

Intriguingly, the DWI signal is not solely influenced by diffusion but also by blood flow, particularly noticeable at low b values ($b < 1500$ s/mm²). This phenomenon is explored in the subsequent section.

1.4 Intravoxel incoherent motion (IVIM)

In 1986, Le Bihan et al. (44) discovered that within DWI, the overall signal attenuation is not only affected by molecular diffusion but also by microcirculation of the blood in the capillary network. This phenomenon, which arises from the complex interplay of perfusion-related motions, introduces an additional layer of complexity to the measured DWI signal. Le Bihan et al. termed this concept ‘Intravoxel Incoherent Motion’ (IVIM) within biological tissues. They observed that these perfusion-related motions exhibit a remarkable level of incoherence, mirroring the apparent distribution of the capillary network. Unlike DCE-MRI, which is reliant on exogenous contrast agents, IVIM is based on DWI which does not require any intravenous contrast agent to provide perfusion-related biomarkers.

Despite IVIM’s early discovery in the 1980s, its journey to widespread clinical research was marked by controversy and skepticism (49). Advancements in imaging technology such as, echo-planar imaging (EPI) (50), which enables the rapid acquisition of a series of gradient echoes, and parallel imaging (51), which uses multiple receiver coils to simultaneously acquire data, have greatly accelerated image acquisition. These innovations have enabled the integration of IVIM imaging into clinical research, particularly for oncological applications (52). Here, IVIM imaging has been applied in various tissues in the body, including brain (53,54), pancreas (55,56), breast (57), prostate (58), head and neck (59), and liver (60).

This section of the doctoral thesis describes the IVIM model used to derive quantitative diffusion and perfusion-related parameters from the DWI data. Moreover, it describes several conventional fitting methods to extract these parameters. This section is primarily based on information provided in the book ‘Intravoxel Incoherent Motion (IVIM) MRI: Principles and Applications’ by Le Bihan et al. (61), unless otherwise specified.

1.4.1 The IVIM model

In 1988 (62), Le Bihan et al. expanded on the groundbreaking work in 1986 (44), where they explained in more detail the mathematical basis of the bi-exponential IVIM model to DWI data. The IVIM model is illustrated in Figure 1.11. This model can be seen as a two-compartment extension of the mono-exponential ADC model. The first compartment represents molecular diffusion, characterized by relatively slow diffusivity (D ; diffusion coefficient). The second compartment, often termed the microvascular or blood flow compartment, represents the dynamic interplay of blood within the microvascular network. Within this microvascular compartment, two essential components are relevant: a faster diffusion component (D^* ;

pseudo-diffusion coefficient), that is often associated with blood velocity information, and microvascular blood volume (f ; signal fraction or often termed ‘perfusion fraction’) representing the signal contribution from the pseudo-diffusion component and provides information about the contribution of blood flow to the overall DWI signal in a voxel. With these parameters, the bi-exponential IVIM model is expressed as follows:

$$S(b) = S_0 \left((1 - f)e^{-bD} + fe^{-b(D^*+D)} \right). \quad (13)$$

However, in modern implementations, it is typical to omit the D in the second exponential term, given that D^* predominantly influences this term, resulting in a simplified expression:

$$S(b) = S_0 \left((1 - f)e^{-bD} + fe^{-bD^*} \right). \quad (14)$$

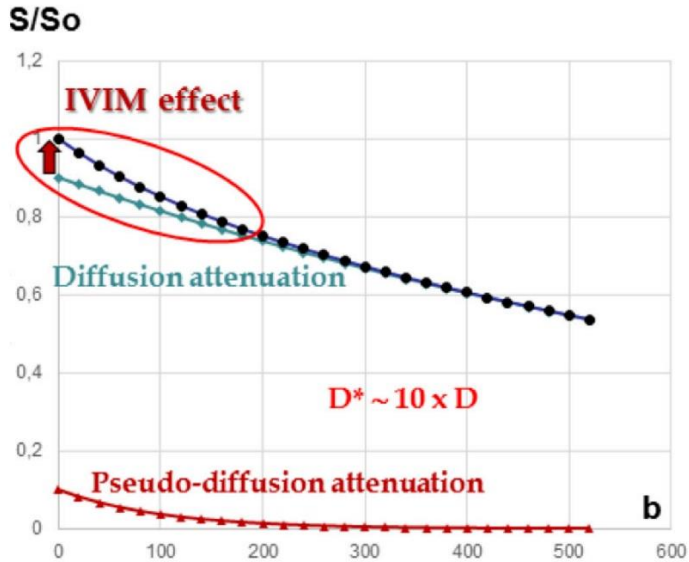


Figure 1.11: Plots illustrating the intravoxel incoherent motion (IVIM) effect across multiple b values (0 - 600 mm^2/s). The signal exhibits a bi-exponential shape due to the independent contributions of tissue diffusion and blood flow components. At particularly low b values ($b < 200$), the IVIM effect becomes particularly noticeable as a deviation of the tissue diffusion signal decay (blue). The bottom-red plot represents the pseudo-diffusion attenuation. This figure is adapted and modified from (52), CC-BY-4.0.

Similar to the mono-exponential ADC model, the IVIM model relies on the assumption that the displacement of water molecules is Gaussian within tissue. This is particularly valid for b values below 1500 s/mm^2 , with the IVIM model typically finding its application in the domain of b values below 1000 s/mm^2 . In practice, multiple DWI images are acquired, often numbering 10 to 20 b values, and with often more b values in the range of 0–200 s/mm^2 . This lower b value range offers valuable insights into perfusion dynamics.

1.4.2 Fitting methods for IVIM

Extracting IVIM model parameters is typically done by fitting the IVIM model to the DWI data. However, this is a challenging ill-posed inverse problem (signal to parameters) where multiple sets of plausible parameters can explain the noisy DWI signals. This challenge primarily originates from two interrelated factors. First, the perfusion fraction f is generally lower than 30%, and often between 5 and 10% (63), particularly for applications in the brain (53,54). Due to these small perfusion fractions, the DWI signal predominantly comprises the diffusion component (D), with only a minimal contribution from the pseudo-diffusion component (D^* and f). Coupled with the intrinsic challenge of a low signal-to-noise ratio (SNR) inherent to DWI, even minor fluctuations in signal measurements can escalate into substantial inaccuracies. Notably, D^* is particularly sensitive to noise. These factors and challenges often lead to solutions converging to local minima rather than the global optimum. Yet, due to the factors described above, even finding the global optimum can yield highly uncertain parameter estimates.

Several techniques are proposed to improve the precision and accuracy of primarily the perfusion-related parameters, f and D^* . This section describes the most common fitting strategies, including some advanced Bayesian approaches.

1.4.2.1 One-step nonlinear least-squares

The most common approach to fit the IVIM model to the DWI data is employing ‘one-step’ nonlinear least squares (LSQ) based on the Levenberg-Marquardt algorithm (64,65) or trust-region algorithm (66,67). These algorithms necessitate initial guesses (commonly setting D^* at 10 times higher than D) and can incorporate bound constraints. This method is the most straightforward as it simultaneously fits the parameters D , f , and D^* . However, despite its apparent simplicity, using this voxelwise approach in practical applications frequently leads to imprecise parameter estimates. Consequently, it is widely acknowledged that the one-step LSQ fitting method often falls short for clinical IVIM imaging.

1.4.2.2 Two-step nonlinear least-squares or segmented fitting

In response to the instability and inaccuracy issues of the one-step LSQ approach, an extension known as the ‘two-step’ nonlinear least squares or segmented approach has been proposed (68). The segmented approach assumes that the value of D^* is approximately one or two orders of magnitude greater than D , often expressed as $D \ll D^*$. Therefore, the influence of D^* can be considered negligible at sufficiently high b values, with the threshold frequently set at $b = 200$ s/mm². At these high b values, typically between 200 and 1000 s/mm², the pseudo-diffusion

contribution diminishes to negligible levels, simplifying the IVIM model to a mono-exponential form:

$$S(b) = S_0(1 - f)e^{-bD} = S_{int}e^{-bD}, \quad (15)$$

where S_{int} is the y-intercept at $b = 0$. In the first step of the segmented approach, the IVIM parameters are typically extracted by taking the natural logarithm of the data, where the slope of the data provides an estimate of D . The parameter f can be derived by extrapolating the mono-exponential fit to S_{int} , with f calculated as:

$$f = \frac{S_0 - S_{int}}{S_0}, \quad (16)$$

In the second step of the segmented approach, the focus shifts to estimating D^* . This is achieved by employing the LSQ algorithm for all b values, and is done by either fixing both D and f or fixing only D , based on the outcomes obtained in the first step of the segmented approach. In the latter case where only D is fixed, the values for both D^* and f are estimated.

Collecting a comprehensive IVIM dataset with 10 to 20 b value measurements is a time-intensive endeavor. To reduce scan time, one could use a simplified version of the segmented approach that excludes the estimation of D^* , and focuses solely on fitting D and f , using only the initial step of the segmented approach. A special case is the acquisition of only 3 b values, which has been shown to hold diagnostic value (69,70).

Several studies that have compared the segmented approach to the one-step LSQ fitting (71–74) clearly indicate that, although LSQ yields a superior fit to the data (with lower residuals in the sum of squares), the segmented approach has demonstrated greater robustness in parameter estimation, especially concerning the perfusion-related parameters. In this thesis, additional comparisons between LSQ and the segmented approach are provided.

1.4.2.3 Bayesian interference

An alternative and more advanced approach to IVIM parameter estimation involves employing Bayesian inference. Bayesian parameter estimation lies in the concept of a joint posterior parameter distribution. According to Bayes' theorem (75), the posterior distribution can be expressed as follows:

$$P(\theta|data) = \frac{P(data|\theta) \times P(\theta)}{P(data)}, \quad (17)$$

where θ is the set of parameters, $P(data|\theta)$ is the likelihood of observing the data given θ , $P(\theta)$ is the prior distribution for θ , and $P(data)$ is the marginal likelihood (also known as the

evidence) and it acts as a normalizing constant. Within the Bayesian framework, an iterative process updates the posterior distribution, where the acquired posterior distribution serves as the new prior for the subsequent iteration. The final posterior distribution is used to obtain parameter estimates, often summarized by the mean or median value.

The Bayesian approach to IVIM parameter estimation has several notable advantages that set it apart from the other fitting methods mentioned above. First, unlike conventional approaches that necessitate initial parameter guesses, Bayesian inference eliminates this requirement. This freedom from initial guesses mitigates the risk of converging into local minima during optimization. Second, Bayesian inference provides a measure of uncertainty (quantifies uncertainty) for each parameter estimate in the form of a probability distribution. Third, the Bayesian framework allows for the incorporation of prior information or assumptions, shaping the parameter estimation process based on additional knowledge. Studies comparing Bayesian methods to conventional approaches frequently noted parameter estimates that exhibit improved accuracy and precision (76,77).

Alternative approaches within the Bayesian framework have also been proposed: a shrinkage prior (78), which aids in reducing the influence of outliers and noise, enhancing the robustness of parameter estimation; and a spatial homogeneity prior (79), assuming that neighboring voxels share similar parameter values and using this local information to generate smoother parameter maps, thereby reducing spatial variability.

However, despite being recognized as a promising approach to enhance IVIM modeling, Bayesian inference has some limitations. Parameter estimates have still demonstrated limited repeatability and substantial variability, particularly for f and D^* (80–82). Additionally, the computational demands associated with Bayesian methods, coupled with the relative unfamiliarity of such techniques, have hindered their widespread adoption in practical applications. Crucially, a comprehensive comparison study by While (83), encompassing both simulations and *in vivo* data, revealed that the use of Bayesian approaches in IVIM modeling may lead to biased parameter estimates or the disappearance of relevant features due to over-smoothing.

In summary, given the challenges concerning the mentioned estimators for IVIM, including sensitivity to noise and outliers, reliance on model assumptions, computational demands, inherent biases, and concerns regarding repeatability and variability, there is a need for alternative fitting approaches to facilitate the clinical implementation of IVIM.

1.5 Artificial Intelligence

In the last decade, significant advancements in big data capabilities and computational power have ushered in a new era of possibilities for Artificial Intelligence (AI). AI involves the development of computer systems capable of executing tasks requiring human-like intelligence, such as learning, problem-solving, and decision-making. These advancements have paved the way for a range of techniques and algorithms enabling computers to uncover intricate patterns within extensive datasets. The continuous evolution of AI, specifically in deep learning and artificial deep neural networks (DNNs), has been a driving force behind these breakthroughs. The conceptualization of DNNs traces its origins back to 1958 when Frank Rosenbalt introduced the notion of ‘Perception’ (84). However, it was only within the last decade that DNNs received scientific recognition when they outperformed other high-profile image analysis benchmarks. Perhaps the first is the renowned ImageNet Large-Scale Visual Recognition Challenge of 2012 (85), where a DNN outperformed the second-best error rate on the image classification task.

Nowadays, the domain of AI stands as one of the most expansive fields of study, exerting a profound influence across every facet of our daily lives. Its impact spans wide-ranging applications, exemplified by innovations in natural language processing (NLP) (86), and speech recognition and synthesis (87,88). Beyond language, this wave of innovation has led to advances in computer vision - a field of artificial intelligence that enables machines to interpret and understand visual information. This extends to medical imaging, with every year yielding a growing number of AI-driven publications in the field of radiology (89).

This doctoral thesis specifically focuses on the application of AI in diffusion parameter estimation, with a primary emphasis on fitting the IVIM model to DWI data. Within this context, this section of the thesis delves into the basic principles of DNNs and gives an overview of various network architectures, learning strategies, and hyperparameters. This section is primarily based on information provided in the book ‘Deep Learning’ by Goodfellow et al. (84), unless otherwise specified.

1.5.1 Basic principles of deep neural networks

Deep neural networks (DNNs) are composed of artificial neurons, also referred to as nodes or hidden units, which are motivated by the structure of biological neurons found in the human brain but are simplified mathematical entities. Figure 1.12 shows an illustration of an artificial neuron. In the context of DNNs, these neurons are typically organized into successive layers,

as exemplified in Figure 1.13 for a multi-layer perceptron (MLP) (Section 1.5.2), following a specific architecture. Each neuron within this architecture receives input from one or more other neurons and processes this input by computing a weighted sum, achieved through vector multiplication of the input with a weight matrix. The weight values or ‘weights’ of the weight matrix represent the strength of the connection between neurons, indicating the input’s influence on the computational process. To introduce flexibility and enhance the ability of the network to generalize, a bias term is integrated into the weighted summation. Eventually, the result undergoes transformation through a nonlinear function known as the activation function, resulting in the output. These transformed outputs can be forwarded to another neuron in the subsequent layer, continuing this computational sequence. Mathematically, the output of an artificial neuron can be expressed by:

$$y_j = f\left(\sum_{i=1}^n x_i w_i + b\right), \quad (18)$$

where each input x_i has an associated weight w_i . The sum of all weighted inputs plus a bias term b is then passed through a nonlinear activation function f , resulting in an output y_j .

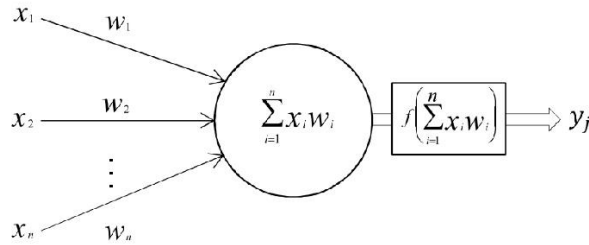


Figure 1.12: Representation of an artificial neuron or hidden unit that serves as a building block of deep neural networks (DNNs). Every input, represented as x_i , is assigned a corresponding weight, w_i . The sum of these weighted inputs, $x_i w_i$, undergoes a transformation through a nonlinear activation function, denoted as f . This process converts the preactivation state of the neuron into an output, denoted as y_j . To maintain simplicity, the inclusion of bias terms has been omitted. The resulting output, y_j , subsequently acts as the input for an artificial neuron in the subsequent layer. This figure is adapted and modified from (90), CC-BY-4.0.

Training a DNN is the process whereby the network learns to make accurate predictions or classifications based on input data. This training involves an iterative process of adjusting the internal parameters of the network such as the weights and biases. Before the start of the training, these parameters are initialized with random values. During the training process, data is introduced into the network through the ‘input layer’, which can be composed of various data types such as images, text, or numerical values. Each neuron in this layer represents a specific feature or input variable. Following this, the output from the input layer is forwarded to the

'hidden layers', consisting of multiple hidden units. These hidden layers transform the input data through a series of weighted connections and activation functions. The depth and width of the hidden layers contribute to the capacity of the network to learn intricate patterns. The final layer, often referred to as the 'output layer', produces the network's predictions. The number of units or neurons in this output layer depends on the nature of the task, with regression tasks having a single output neuron and classification tasks having one neuron per class. In deep learning diffusion and IVIM parameter estimation, the number of hidden units in the output layer often corresponds to the number of model-parameters being estimated.

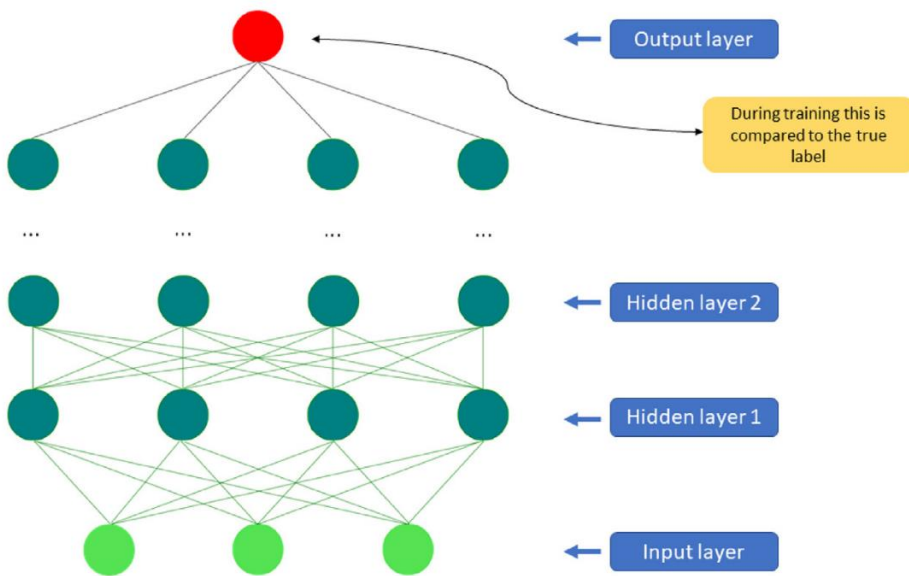


Figure 1.13: Example of a multi-layer perceptron (MLP). All artificial neurons within a layer are intricately linked to every neuron in the subsequent layer, forming a fully-connected network. In the training phase, input data is fed into the input layer. It then traverses through the hidden layers before reaching the output layer. In the output layer, predictions are produced that represent a hierarchical representation of the input data. The difference between the predicted output and the actual label based on the loss function undergoes processing through an activation function. This dissimilarity is subsequently backpropagated to adjust the network's weights and biases using a gradient descent optimization algorithm, aiming to minimize the prediction error. This figure is adapted and modified from (91), CC-BY-4.0.

Throughout the training process, predictions are generated based on the specific nature of the task. Subsequently, a loss function quantifies the difference between predicted and ground truth (actual values). Mean squared error (MSE) also termed L2-loss is a common choice as loss function in parameter estimation. Other loss functions include mean absolute error (L1), cross-entropy loss, and Dice loss (91). Once the loss function is defined, a gradient descent

optimization algorithm, commonly referred to as an optimizer, is employed. Through a series of iterations and weight updates in a process known as backpropagation, the optimization algorithm strives to find the global minimum of the network's loss function. Eventually, if the training is performed correctly, the model converges, which leads to a DNN that can make accurate predictions when presented with new, unseen data.

The successful deployment of DNNs often relies upon effectively managing two critical phenomena: underfitting and overfitting. Underfitting occurs when a DNN fails to capture the underlying patterns within the data effectively. In such instances, the model is often too simplistic to discern intricate relationships, resulting in suboptimal performance. This can be addressed by increasing the learning capacity of DNNs, i.e. increasing the number of hidden layers or hidden units (see also Section 1.5.4 below). Conversely, overfitting occurs when a network becomes too closely adapted to the training data so that it loses its applicability to unseen data. Mitigating overfitting is a challenge that involves finding a balance between model complexity and generalization. Several regularization methods have been proposed that could help to prevent overfitting, with a selection of them described in Section 1.5.4.

1.5.2 Types of network architectures

DNNs operate in network architectures that define the structure and connectivity of layers and artificial neurons. Network architectures serve as a blueprint for how information propagates within the network, how features are extracted and modified, and how predictions or decisions are made.

Typical DNN architectures consist of an encoder and a decoder component. The encoder transforms input data into a lower-dimensional or latent representation, retaining essential information and valuable features. The decoder reconstructs this encoded representation into a format resembling the original input data. In deep learning diffusion parameter estimation, it is common to only utilize the encoder component of the DNN for the transformation of input data into a latent space representing the parameters of the biophysical model. Therefore, in the case of the IVIM model, the output often includes three units or neurons, representing D , f , and D^* , and may also incorporate an additional fourth unit, representing S_0 . This S_0 parameter enables the network to perform noise correction.

This section explores three prominent network architectures: the multi-layer perceptron (MLP), convolutional neural networks (CNNs), and transformers.

1.5.2.1 Multi-layer perceptron (MLP)

MLPs, also known as a feedforward neural networks or deep feedforward networks, are one of the simplest but most fundamental DNN architectures. It consists of multiple layers of artificial neurons, where each neuron is fully-connected to every neuron in the adjacent layers, as illustrated in Figure 1.13. MLPs are versatile and can be applied to a wide range of problems, including regression and classification tasks. They are particularly effective for structured data, where the relationships between features are not spatially or temporally dependent. MLPs for deep learning diffusion parameter estimation are extensively explored within this thesis.

1.5.2.2 Convolutional neural networks (CNNs)

CNNs (92,93) are a category of MLPs that employ convolution layers with predefined kernels to operate on image data within its hidden layers. CNNs are designed to learn hierarchical representations, capturing local patterns and spatial dependencies.

The inherent inductive bias of CNNs, which is rooted in the assumptions of locality and weight sharing, along with their translational equivariance properties, enables them to identify patterns and features in images regardless of their position or translation. CNNs have revolutionized the field of computer vision, image-processing, and medical image analysis (91,94,95). Paper III explores CNNs for deep learning diffusion parameter estimation.

A typical CNN architecture is depicted in Figure 1.15. The first hidden layer often consists of a convolutional layer to extract features from the input image. Convolution layers employ kernels applied to the input matrix (2D data), resulting in a smaller output matrix. This kernel represents a learnable weight matrix. To preserve the matrix size, padding can be employed around the input matrix, i.e. enlarging the image. Zero padding is the most common type of padding where the input image is surrounded by zeros. Subsequent hidden layers often encompass additional convolution layers, pooling layers, or fully-connected layers. By stacking multiple convolutional layers, the network can capture a hierarchy of features. This enables the learning of features and patterns at different levels. Early layers tend to capture simple features like edges, corners, and textures, while deeper layers progressively learn more complex and abstract features.

Following the application of a convolution layer, CNNs often incorporate pooling layers to reduce the spatial dimensions of the feature maps. Common pooling techniques include max pooling, which selects the maximum value within a kernel, and average pooling, which computes the average of the values within the kernel. The final layers typically consist of fully-

connected layers responsible for making predictions or classifications based on the extracted features.

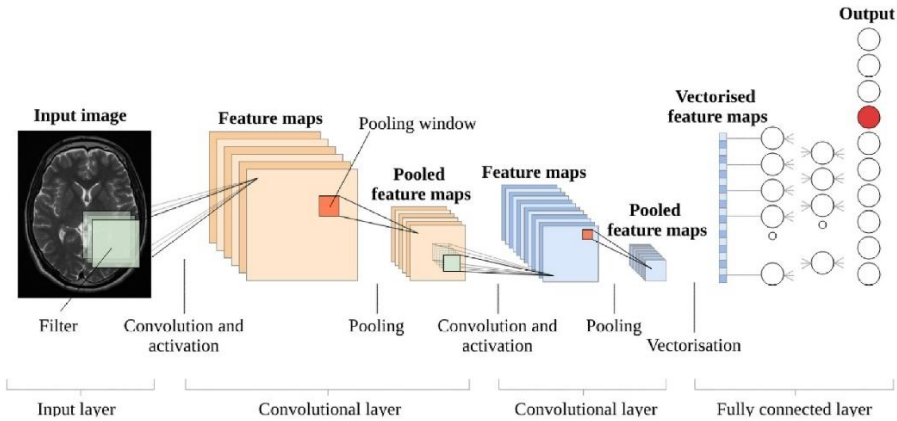


Figure 1.14: Building blocks of a typical CNN. A CNN consists of an input layer (in this case a brain scan), followed by several convolution layers and pooling layers. At the end, CNNs typically have fully-connected layers that compute the final output. This figure is adapted and modified from (91), CC-BY-4.0.

1.5.2.3 Transformers

Transformer neural networks (96), also known as transformers or attention models, were first introduced in the renowned paper of Vaswani et al. (96) ‘Attention Is All You Need’. The typical transformer architecture comprises an encoder stack and a decoder stack, as depicted in Figure 1.16. The decoder stack consists of a part that is solely a decoder component, which is followed by an encoder-decoder component that combines the outcomes of the encoder stack and the decoder component. Since the introduction of transformers, they have emerged as a groundbreaking innovation in the fields of natural language processing (NLP) and sequential data (86). Perhaps the most well-known application of transformers for NLP is the powerful tool ‘Chat-GPT’, a language model trained on massive text datasets in multiple languages, capable of generating human-like responses to text. Recently, transformers have also garnered substantial interest in computer vision applications (97).

Transformers are based on attention mechanisms that can capture long-range dependencies within sequences. This doctoral thesis, in particular Paper III, explores the concepts of self-attention (96) and the recently proposed neighborhood-attention (98) within transformers for deep learning parameter estimation.

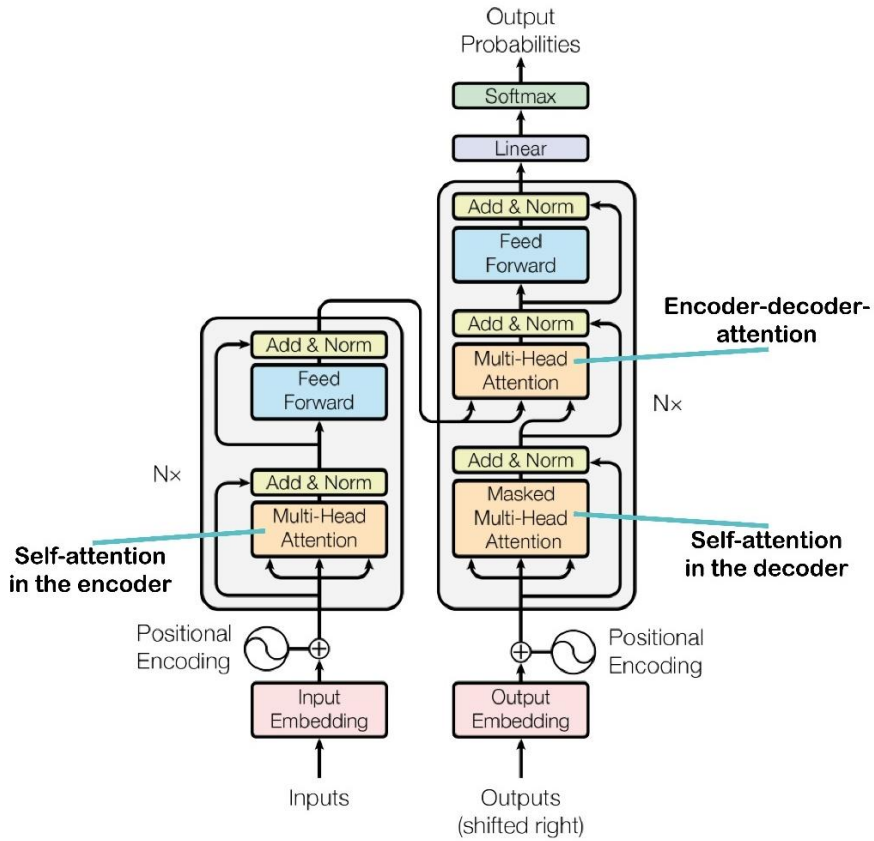


Figure 1.1: The transformer architecture, consisting of an encoder component, a decoder component and an encoder-decoder component. This figure is adapted and modified from (91), with permission of Google.

Self-attention

Central to transformers is the concept of self-attention (96), allowing each element in a sequence to weigh its relationships with all other elements in the same sequence. In the transformer, prior to the application of the self-attention module, the input sequence undergoes an initial step involving embedding and positional encoding. This process generates embedding vectors for each element within the sequence. Within the self-attention module, the embedded sequence proceeds through a linear layer, which generates three separate matrices, known as Query (Q), Key (K), and Value (V) matrices. In essence, self-attention involves a series of learnable transformations, embeddings, and matrix multiplications utilizing these Q, K, and V matrices. Self-attention is mathematically expressed as follows:

$$\text{self-attention}(Q,K,V)= \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (1)$$

where d is the embedding dimension and T a transpose operation.

When determining self-attention, a ‘factor’ matrix is derived through the dot product of the Query and Key matrices. Conceptually, Q can be interpreted as a series of elements for which attention is computed, while the elements in K represent those to which Q directs its attention. Therefore, the dot product between the Query and Key matrices quantifies the relevance between each pair of elements. The outcome undergoes a SoftMax operation, which normalizes the factor matrix into a probability distribution ranging from 0 to 1, summing all scores up to 1. Multiplying the output of the Softmax with the Value matrix (i.e. the sequence itself) yields a weighted sum (also known as the attention score) of all elements within the Value matrix. This attention score serves as the output of the self-attention module. Following the attention module, the outcome is processed by fully-connected layers to transform the attention score into an appropriate output for further processing.

Transformers typically employ multi-headed attention, where the attention mechanism is applied multiple times in parallel, each with a different set of learned weight matrices for Q, K, and V. These different sets of weights allow the model to attend to different parts of the input sequence and capture different patterns or relationships within the data.

In a transformer architecture (Figure 1.16), self-attention is applied at two key locations: in the encoder stack, where the source sequence pays attention to itself; and at the initial part of the decoder stack (decoder component), where the target sequence pays attention to itself. Additionally, attention is also performed later in the decoder stack, specifically in the encoder-decoder component, called encoder-decoder attention. Here, the Query matrix is derived from the target sequence (output decoder self-attention) and the Key and Value matrices from the source sequence (output encoder self-attention). Hence, it computes the relevance of each element in a target sequence to each element in the source sequence. A transformer model can encompass multiple attention blocks in each of these transformer components to improve its overall capacity and performance.

Despite the widespread utility of self-attention models in NLP tasks (86), applying self-attention to vision-related tasks poses challenges (97). In the self-attention concept, each token within the sequence can attend to any other token within that same sequence, irrespective of their relative positions or the sequence order, as it is permutation-invariant. This characteristic results in a quadratic increase in complexity with respect to the number of tokens. This quadratic nature results in high memory demands, especially when applying self-attention modules to images, as memory requirements scale with the number of pixels. This prevents self-attention modules from being easily applicable to computer vision tasks (97). Recent

innovations have introduced variants of attention mechanisms that address some of these challenges (99,100), among them the concept of neighborhood-attention (98).

Neighborhood-attention

Neighborhood-attention, as introduced by Hassani et al. (98), represents a modified variant of self-attention that localizes attention to the nearest neighboring pixels. This approach involves a sliding window operation using the Q, K, and V pixels within a kernel. The definition of neighborhood-attention, as outlined in the paper of Hassani et al., can be described as:

$$\mathbf{neighborhood-attention}_k(\mathbf{i}) = \mathbf{softmax} \frac{A_i^k}{\sqrt{d}} \mathbf{V}_i^k, \quad (20)$$

with i being the i -th pixel to which neighborhood-attention is being applied, A_i^k is the dot product of the i -th input Query projection with the Key projections of its k nearest neighbors within the applied kernel, d represents the embedding dimension, and V_i^k represents the i -th input's k nearest neighboring Value projections. The computation of A_i^k can be expressed as:

$$A_i^k = \begin{bmatrix} \mathbf{Q}_i \mathbf{K}_{\rho_1(i)}^T + \mathbf{B}_{(i, \rho_1(i))} \\ \mathbf{Q}_i \mathbf{K}_{\rho_2(i)}^T + \mathbf{B}_{(i, \rho_2(i))} \\ \vdots \\ \mathbf{Q}_i \mathbf{K}_{\rho_k(i)}^T + \mathbf{B}_{(i, \rho_k(i))} \end{bmatrix}, \quad (21)$$

where $\rho_j(i)$ denotes i 's j -th nearest neighbor. This calculation is a modification of the dot product of Q and K within self-attention, where in the context of neighborhood-attention, it is restricted to the k nearest neighbors of pixel i . V_i^k can be expressed as:

$$\mathbf{V}_i^k = [\mathbf{V}_{\rho_1(i)}^T \ \mathbf{V}_{\rho_2(i)}^T \ \dots \ \mathbf{V}_{\rho_k(i)}^T]^T. \quad (22)$$

As neighborhood-attention exclusively operates on the nearest neighbors, it introduces local inductive biases and maintains translational equivariance, a characteristic shared with CNNs. Additionally, it conserves memory usage, by only considering the attention matrix for the k nearest neighbors, which stands in contrast to self-attention where all pixels in the image are involved in the computation.

Neighborhood-attention finds its integration within NATTEN, a Python package that incorporates specialized C++ and Cuda kernels specifically designed for neighborhood-attention. Through NATTEN, neighborhood-attention transformers, which are hierarchical transformers constructed with multiple neighborhood-attention blocks, can run fast with efficient memory utilization. In the original paper of Hassani et al., these neighborhood-attention transformers outperformed or were comparable to other sliding-window attention

(Swin) transformers (99) and convolution-based models (ConvNeXt) (101) in downstream vision tasks such as image classification and segmentation.

1.5.3 Types of learning strategies

There are several types of learning strategies in terms of the level of supervision for training DNNs. This section describes supervised learning, unsupervised learning, and semi-supervised learning.

1.5.3.1 Supervised learning

Supervised learning is one of the most common learning strategies for training DNNs. In this approach, the network is provided with a labeled dataset. The primary objective of supervised learning is to learn a mapping from inputs to outputs that minimizes the discrepancy between predicted and actual targets or ground truth. When provided with sufficient labeled data, supervised learning can yield high-quality models with strong predictive capabilities. However, there are challenges associated with this approach. Accessing labeled data can be time-consuming and expensive. Furthermore, in many applications, obtaining ground truth annotations can be impractical or infeasible.

In the context of supervised learning for diffusion parameter estimation, the network is trained using the biophysical model-parameters as ground truth. Typically, the input of the network comprises the measured or simulated DWI signals, while the output of the network are parameter estimates for the biophysical model. A supervised network is optimized using a loss function, often the MSE, between the network’s output parameters $\hat{\theta}_{net}$ and the ground truth parameters $\hat{\theta}$, denoted in this thesis as “parameters-MSE”:

$$\text{parameters-MSE} = L(\hat{\theta}, \hat{\theta}_{net}) = \sum_{\text{parameters}} \|\hat{\theta} - \hat{\theta}_{net}\|^2. \quad (23)$$

In IVIM parameter estimation, these $\hat{\theta}$ and $\hat{\theta}_{net}$ typically include D, f, and D*, and optionally S₀.

One of the major challenges in supervised learning in the context of diffusion and IVIM parameter estimation is the scarcity of reliable ground truth data, primarily due to the ill-posed nature of the inverse problem. Conventional fitting methods, such as least squares and other estimators, are known to be highly sensitive to noise and rely on specific assumptions (Section 1.4.2), rendering them unsuitable to serve as ground truth for training supervised networks. An alternative approach for generating ground truth for training supervised networks is to consider synthetic data. Within this thesis, supervised learning is thoroughly explored for deep learning

diffusion parameter estimation. Paper II explores supervised learning in the context of IVIM parameter estimation using MLPs and conducts a thorough exploration of the limitations associated with this learning strategy. Paper III explores whether incorporating spatial information can enhance supervised deep learning diffusion parameter estimation, with application to IVIM.

1.5.3.2 Unsupervised learning

An alternative to relying on labeled data is unsupervised learning. In unsupervised learning, the objective of the network is to discover underlying patterns or structures within the data without the use of explicit target labels. Clustering is a common example of unsupervised learning, where samples are automatically grouped into clusters based on their most distinct features. Unsupervised learning has the advantage of not requiring labeled data, making it suitable for scenarios where ground truth is challenging or unavailable. However, because there is no guidance of labeled data, it is challenging to assess the quality of learned representations.

In the context of unsupervised learning for diffusion parameter estimation, the network is trained using the measured or simulated DWI signals as ground truth. Similar to supervised learning, the input of the network encompasses the measured or simulated DWI signals, and its output provides parameter estimates for the biophysical model. However, in unsupervised learning, these parameter estimates are used to predict DWI signals, denoted as $S_{net}(b)$, by utilizing the biophysical model, such as the IVIM model. Subsequently, the unsupervised networks are optimized using a loss function, typically the MSE, between the input signal $S(b)$ and the predicted signal $S_{net}(b)$, denoted in this thesis as “signals-MSE”:

$$\mathbf{signals-MSE} = L(S(b), S_{net}(b)) = \sum_{b \in B} \|S(b) - S_{net}(b)\|^2. \quad (24)$$

As these unsupervised networks employ the biophysical model to predict signals for network optimization, they are often referred to as 'physics-informed' DNNs.

This doctoral thesis explores various aspects of unsupervised learning for deep learning diffusion and IVIM parameter estimation. Paper I identifies specific issues associated with a prior unsupervised approach for IVIM parameter estimation (102), and subsequently addresses them by introducing various network hyperparameters. Paper II conducts a thorough exploration of the limitations associated with unsupervised learning for IVIM parameter estimation. Lastly, Paper III investigates whether incorporating spatial information can enhance unsupervised deep learning diffusion parameter estimation, with application to IVIM.

1.5.3.3 Semi-supervised learning

Semi-supervised learning combines elements of both supervised and unsupervised learning to address some of the limitations of each approach. In this paradigm, a model is trained on a dataset containing a small portion of labeled data and a more extensive set of unlabeled data.

Semi-supervised learning typically involves two stages. First, the network is trained in a general supervised manner using the available labeled data. Subsequently, this model is used to predict labels for the unlabeled data, a process known as pseudo-labeling. After labeling the unlabeled data, the model is retrained on the entire dataset, comprising both the original labeled data and the newly pseudo-labeled data. By leveraging the abundance of unlabeled data, this approach can make efficient use of limited labeled data, potentially leading to improved model performance.

Semi-supervised learning is particularly valuable when labeled data is scarce or expensive to obtain. However, the quality of the ‘pseudo-labels’ assigned to the unlabeled data is not assured, which can substantially affect the effectiveness of semi-supervised learning. Semi-supervised learning is not explored in this thesis.

1.5.4 Hyperparameters

In addition to the different network architectures and learning strategies, the effectiveness of DNNs is also determined by a set of tunable network parameters, known as hyperparameters. Hyperparameters can be thought of as the configuration settings that control various aspects of the network’s training process, and they must be carefully tuned to achieve optimal performance. This section outlines several hyperparameters that have a substantial impact on the training process of the network. Papers I and II explore these hyperparameters thoroughly with the application to deep learning IVIM parameter estimation.

1.5.4.1 Training hyperparameters: learning rate, optimizer, batch size

Perhaps the most crucial DNN training hyperparameter for achieving training stability and optimal convergence is the learning rate. The learning rate controls the step size at which the optimizer updates the intrinsic network parameters.

As mentioned in Section 1.5.1, the optimizer determines the update rule for adjusting the network's parameters during training. The optimizer strives to find the optimal set of network parameters that results in the best possible network performance. Several optimizers have been proposed (103). One of the most commonly used optimizers is Adam (104), which is an adaptive optimization algorithm that adjusts learning rates for each parameter individually,

incorporates momentum to accelerate convergence, and includes bias correction to stabilize training.

Another important hyperparameter is the batch size, which determines how many data samples are processed together in each forward and backward pass during training. A smaller batch size provides a more frequent update to the model but may lead to noisy gradients, while a larger batch size offers more stable updates but requires higher memory capacity. The choice of batch size often involves trade-offs between training speed and convergence stability, and tuning the learning rate (105). For example, a large batch size may require a smaller learning rate to avoid overshooting, while a small batch size may require a larger learning rate to escape local minima.

1.5.4.2 Network hyperparameters: hidden units, hidden layers, activation function

The number of hidden units and number of hidden layers are network hyperparameters that control the learning capacity of the DNN. In general, a larger network has a stronger capacity for learning features in the data but also takes a longer time to train. The number of hidden units in each layer defines the capacity of the network to capture and represent features in the data. A higher number of units can lead to overfitting if not adequately regularized, while a lower number may result in underfitting. In addition, the depth of a network, defined by the number of hidden layers, substantially impacts its representational power. Deeper networks can model intricate features but may suffer from vanishing gradients during training.

Another crucial hyperparameter within the network architecture is the activation function that introduces non-linearity to the DNN, enabling it to model complex relationships within the data. Common activation functions include the sigmoid function and exponential linear unit (106).

1.5.4.3 Regularization: batch normalization, dropout, early stopping

There are several regularization techniques proposed to mitigate overfitting, a phenomenon where the model performs well on training data but fails to generalize effectively to unseen data. These regularization techniques are tools in the DNN toolkit and require tuning to achieve the desired regularization effect. Therefore, we also refer to them in this thesis as hyperparameters.

Two widely used regularization techniques are batch normalization and dropout. Batch normalization (107) normalizes the activations of a layer across a mini-batch of data during training. It ensures that activations have a consistent mean and variance, and therefore stabilizes training and often accelerates convergence. Dropout (108) is a regularization technique that addresses overfitting by randomly deactivating a subset of hidden units during each forward

and backward pass. This probabilistic strategy generates an ensemble of sub-networks during training, thereby enhancing the robustness of the network and reducing its dependence on specific neurons.

Another approach to regularize DNNs and prevent overfitting is through ‘early stopping’. During training, an early-stopping criterion continuously monitors the performance of the model on a test dataset. If this performance begins to decrease, the training process is stopped after a specified number of epochs (typically set at 10 epochs). This technique aims at preventing the model from over-optimizing on the training data, enhancing its capacity to generalize to unseen data.

Chapter 2

Objectives

The overarching objective of this doctoral thesis was to explore the use of deep learning to estimate clear, detailed maps of perfusion-based biomarkers from diffusion-weighted magnetic resonance imaging (DWI). The primary goal was to enhance parameter estimation for the intravoxel incoherent motion (IVIM) model within DWI, where successful outcome will allow for better cancer diagnosis and treatment monitoring, without the need for contrast injection.

The specific aims of this thesis were to:

- Enhance a prior promising unsupervised deep learning IVIM fitting approach by optimizing several network hyperparameters through extensive simulations, and compare the optimized approach to other conventional fitting approaches using in vivo data from pancreatic cancer patients (Paper I).
- Perform a comprehensive analysis of deep learning IVIM parameter estimation by exploring the impact of training features (including training distribution) for both unsupervised and supervised learning using simulations and in vivo data from brain cancer patients (Paper II).
- Develop novel deep learning approaches to parameter estimation that can leverage spatial information to improve estimator performance, with application to diffusion modeling and IVIM (Paper III).

Chapter 3

Summary of papers

This chapter summarizes the importance of each research paper in this thesis, helping to provide a foundational understanding of the core themes, methods, and outcomes explored in each work. The full papers are presented in their published or submitted format at the end of this thesis.

3.1. Paper I

Improved unsupervised physics-informed deep learning for intravoxel incoherent motion modeling and evaluation in pancreatic cancer patients

Misha Pieter Thijs Kaandorp, Sebastiano Barbieri, Remy Klaassen, Hanneke W. M. van Laarhoven, Hans Crezee, Peter T. While, Aart J. Nederveen, Oliver J. Gurney-Champion

Magnetic Resonance in Medicine, 86(4), 2250-2265.

The intravoxel incoherent motion (IVIM) model for diffusion-weighted imaging (DWI) demonstrates great promise in the estimation of predictive and prognostic cancer imaging biomarkers. However, conventional IVIM fitting approaches such as conventional LSQ and Bayesian approaches exhibit long fitting times and demonstrate poor image quality or biased parameter estimates, thereby preventing clinical implementation of IVIM. Recent work by Barbieri et al. (102) introduced an unsupervised deep learning approach for IVIM fitting, denoted as IVIM-NET_{orig}. This unsupervised approach demonstrated superior performance in terms of both accuracy and fitting time compared to the LSQ and Bayesian approaches. However, that study focused only on volunteer data and limited its exploration of network hyperparameters. Moreover, initial observations of Paper I identified unexpected parameter correlations between the perfusion-related parameter estimates. Therefore, the aim of Paper I was to optimize IVIM-NET_{orig} by exploring modifications of the network's architecture and several hyperparameters in simulations. We evaluated our optimized approach using in vivo data from patients with pancreatic ductal adenocarcinoma (PDAC), i.e. pancreatic cancer, receiving neoadjuvant chemoradiotherapy (CRT).

We initially implemented IVIM-NET_{orig} (102), which was an MLP (Section 1.5.2.1), with three hidden layers, where the number of hidden units was equal to the number of b values. We considered many variants of this original IVIM-NET by introducing and altering various hyperparameters. These hyperparameters included:

- An extra fit parameter S_0 to allow the network to correct for noise.
- Sigmoid activation functions at the end of the network instead of absolute activation functions.
- Varying the number of hidden layers.
- Adding dropout regularization (108) (Section 1.5.4.3).
- Adding batch normalization (107) (Section 1.5.4.3).

- A novel parallel network design, where each parameter is predicted in parallel by independent subnetworks.
- Tuning the learning rate for optimal convergence.

All networks were trained in the same unsupervised manner (Section 1.5.3.2). The training was stopped when there was no improvement in the test loss over a continuous span of 10 epochs, e.g. early stopping (Section 1.5.4.3).

In simulations, we sampled IVIM parameters from a uniform distribution: $0.5 \times 10^{-3} \leq D \leq 3 \times 10^{-3}$ mm²/s, $5 \leq f \leq 55\%$, $10 \times 10^{-3} \leq D^* \leq 100 \times 10^{-3}$ mm²/s, and $S_0=1$. These were used to generate 100,000 DWI signals, considering 12 b values (0, 5, 10, 20, 30, 40, 60, 150, 300, 500, 700 mm²/s) using eq. (14). Rician noise was added to the signals such that they possessed a relative SNR of 20. As this is an unsupervised approach, training was done on the same data as the evaluation data (i.e. test data). We trained and tested our various networks with different hyperparameter configurations on this simulated data. We assessed the performance of these networks in terms of accuracy, independence, and consistency by calculating the normalized root-mean-square error (NRMSE), Spearman’s correlation ρ , and the coefficient of variation (CV_{NET}), respectively. The best-performing network, referred to as IVIM-NET_{optim}, was compared to a LSQ approach and a Bayesian approach at different SNRs.

Subsequently, IVIM-NET_{optim}’s performance was evaluated using an independent in vivo dataset comprising 23 patients with PDAC. Among these patients, fourteen had not undergone any treatment between two scan sessions, and nine had received CRT between these sessions. Our evaluation included a qualitative analysis of the parameter maps, focusing on feature clarity and visually assessing the consistency of the fit to the IVIM signal in pairs of neighboring voxels. Additionally, we conducted a quantitative assessment by evaluating test-retest repeatability. This involved calculating the intersession within-subject standard deviation (wSD) for each IVIM parameter using the data from the patients with repeated baseline scans. Bland-Altman plots were generated using data from both patient cohorts to identify significant parameter changes following CRT. Comparisons were also made to LSQ and a Bayesian approach.

Figure 3.1 shows a subset of the results of the simulation study in Paper I. The originally proposed unsupervised approach, IVIM-NET_{orig}, exhibited strong correlations between the perfusion-related parameters ($\rho(D^*,f) = 0.74$). IVIM-NET_{optim} resolved these correlations ($\rho(D^*,f) = 0.22$) (Figure 3.1A) by having a network consisting of a parallel network design architecture with two hidden layers, batch normalization, dropout of 10%, sigmoid constraints,

the inclusion of the fit parameter S_0 , and a learning rate of 3×10^{-5} . IVIM-NET_{optim} demonstrated improved accuracy to IVIM-NET_{orig} (NRMSE(D) = 0.177 vs 0.196; NRMSE(f) = 0.220 vs 0.267; NRMSE(D*) = 0.386 vs 0.393). Furthermore, it demonstrated substantially better accuracy in comparison to conventional LSQ and Bayesian approaches, particularly at low SNR (Figure 3.1B).

Figure 3.2 shows a subset of the results of the in vivo study in Paper I. Deploying IVIM-NET_{optim} to in vivo data from PDAC patients produced more detailed and significantly less noisy parameter maps than the other conventional fitting methods (Figure 3.2A). Furthermore, IVIM-NET_{optim} detected the most significant parameter changes in D and f across the entire cohort undergoing CRT (Figure 3.2B).

In conclusion, this study improved the accuracy, independence, and consistency of both diffusion and perfusion-related parameters from IVIM-NET by changing the network architecture and tuning hyperparameters. When deploying IVIM-NET_{optim} to in vivo DWI data from pancreatic cancer patients, IVIM-NET_{optim} demonstrated to be considerably faster and to produce less noisy and more detailed parameter maps compared to alternative state-of-the-art fitting methods. Furthermore, in this cohort, IVIM-NET_{optim} had a substantially better test-retest repeatability and was able to detect the most individual patients with significant changes in the IVIM parameters throughout radiotherapy. Therefore, in Paper I, we recommended IVIM-NET_{optim} for accurate, informative, and consistent IVIM fitting to DWI data.

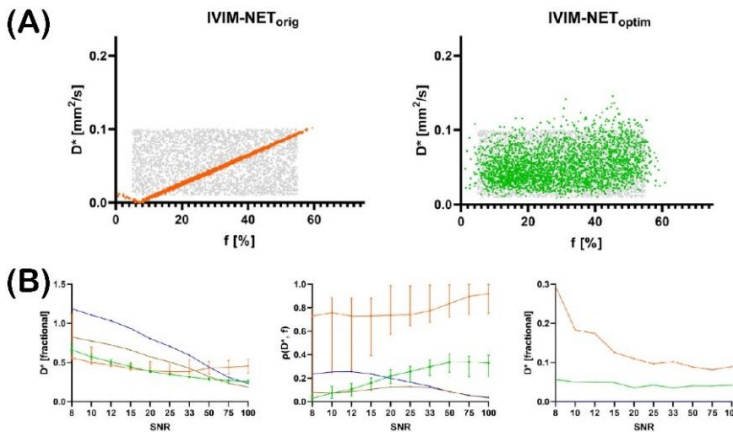


Figure 3.1: A subset of the results of the simulation study performed in Paper I. **(A)** D^* plotted against f for IVIM-NET_{orig} (left) and IVIM-NET_{optim} (right). The ground truth parameter values are presented in gray. **(B)** NRMSE (left), $\rho(D, f)$ (center), and CV_{NET} (right) plots of the estimated D^* IVIM parameter computed at several SNRs for the least squares (LS; blue), Bayesian (brown), IVIM-NET_{orig} (orange), and IVIM-NET_{optim} (green) approaches to IVIM fitting. The 5 to 95 percentiles of IVIM-NET for 50 repeated pieces of training are plotted as error bars and the median value is used for the line plot.

Summary of papers

The plots show that $\text{IVIM-NET}_{\text{orig}}$ is highly inconsistent in producing IVIM parameters for multiple repeated trainings at all SNRs, and that $\text{IVIM-NET}_{\text{optim}}$ outperforms $\text{IVIM-NET}_{\text{orig}}$. As the least squares and Bayesian approaches are deterministic, their CV_{NET} was zero and not plotted. The least squares and Bayesian approaches were superior at high SNR. Note the clear parameter correlations between the perfusion-related parameter estimates ($\rho(D^*, f)$) in $\text{IVIM-NET}_{\text{orig}}$, and their absence for $\text{IVIM-NET}_{\text{optim}}$.

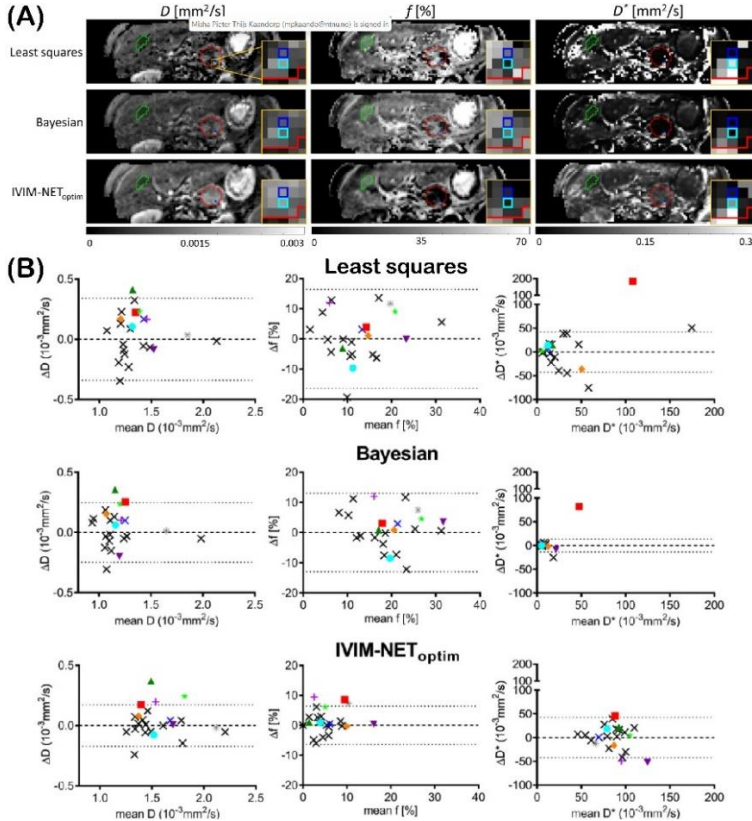


Figure 3.2: A subset of the results of the in vivo study on pancreatic ductal adenocarcinoma (PDAC) patient data performed in Paper I. **(A)** IVIM parameter maps of the least squares, Bayesian, and $\text{IVIM-NET}_{\text{optim}}$ approaches for a PDAC patient of the treated cohort before chemoradiotherapy. The red ROI represents the PDAC (tumor), and the green ROI represents 2D homogenous liver tissue. The two highlighted blue regions correlate to the voxels from the log-plots (not illustrated; see the full version of Paper I). In the parametric maps computed by $\text{IVIM-NET}_{\text{optim}}$, the tissues appear more homogeneous, particularly in the liver, the kidneys, and around the tumor ROI. **(B)** Bland-Altman plots of the least squares, Bayesian, and $\text{IVIM-NET}_{\text{optim}}$ approaches showing the mean and difference (Δ) between the intersession repeatability patients (black crosses) and the mean and Δ between pre- and post-treatment patients (colored symbols), which represents the treatment effects. The dotted lines indicate the 95% confidence intervals of the test-retest data. Colored measurements that exceed the 95% confidence intervals were considered significant to treatment response. $\text{IVIM-NET}_{\text{optim}}$ detected the most patients with significant parameter changes (10 significant parameter changes) after chemoradiotherapy. In comparison, the least squares and Bayesian approaches detected only two and three significant parameter changes, respectively.

3.2. Paper II

Deep learning intravoxel incoherent motion modeling: Exploring the impact of training features and learning strategies

Misha Pieter Thijs Kaandorp, Frank Zijlstra, Christian Federau, Peter T. While

Magnetic Resonance in Medicine, 90(1), 312-328.

While the improved unsupervised deep learning approach of paper I, IVIM-NET_{optim}, successfully addressed unexpected parameter correlations in the originally proposed unsupervised approach (102) and demonstrated improved estimator performance using in vivo data from pancreatic cancer patients, its application to the brain revealed poor anatomy generalization and elevated D^* values (109). Other research indicated that supervised deep learning approaches (Section 1.5.3.1) may exhibit training data bias (110). Therefore, the aim of paper II was to explore the impact of key training features, including the effect of the training data, for both unsupervised and supervised learning for IVIM model fitting. The main motivation of this work was not to develop a network that could improve IVIM parameter estimation, but rather to investigate the possible limitations of voxelwise deep learning IVIM parameter estimation.

One in vivo data set and two synthetic data sets were used in training and evaluating unsupervised and supervised networks:

1. In vivo data from glioma patients: This data set consisted of 28 glioma patients acquired at 16 b values: 0, 10, 20, 40, 80, 110, 140, 170, 200, 300, 400, 500, 600, 700, 800, 900 mm^2/s .
2. Uniform distribution (*synthetic*): IVIM parameters were sampled from a uniform distribution: $0 \leq S_0 \leq 1$, $0 \times 10^{-3} \leq D \leq 3 \times 10^{-3} \text{ mm}^2/\text{s}$, $0 \leq f \leq 50\%$, and $3 \times 10^{-3} \leq D^* \leq 100 \times 10^{-3} \text{ mm}^2/\text{s}$, wherefrom 100,000 DWI signals were generated using eq. (14), considering similar 16 b values as described above. Rician noise was added to the signals such that when $S_0=1$ the SNR was 200.
3. Patient distribution (*synthetic*): To provide a realistic patient distribution, we obtained IVIM parameter combinations from the segmented fit applied to the in vivo data. We then used these parameter combinations to simulate DWI signals.

In our first experiment, the primary focus was to evaluate the influence of the learning rate, network size, and the suitability of the early-stopping criterion. For this purpose, we

implemented the original network of IVIM-NET_{orig} (102) and altered these hyperparameters. These networks were trained and tested on the synthetic uniform distribution (dataset (2) above). We trained these networks to full convergence, surpassing the early-stopping point, which is different from both the original unsupervised approach (102) and the approaches described in Paper I, including IVIM-NET_{optim}. Assessment encompassed an evaluation of network stability, loss convergence, and generalizability for both unsupervised and supervised learning. We used the hyperparameters corresponding to the most stable network in further experiments.

Four further networks were trained by considering the two different learning strategies (unsupervised, supervised) and the two synthetic data sets above (uniform distribution, patient distribution). These networks were evaluated at different validation points during training, including the early-stopping point. We compared the parameter estimates with the ground truth for individual data points of the uniform distribution test set. We further qualitatively assessed each network by evaluating parameter maps and root-mean-square error (RMSE) for a representative slice of a synthetic glioma patient (dataset (3) above). In addition, we evaluated the performance of our simulated networks using the in vivo data from glioma patients (dataset (1) above) to demonstrate their suitability in clinical settings. In these evaluations, we also compared the networks to an unsupervised network directly trained on the in vivo signals, LSQ, the segmented approach, and IVIM-NET_{optim}.

In simulations, we showed that both unsupervised and supervised learning benefited from utilizing a network with more hidden units, particularly a number that exceeded the number of b values. Important to note is that this is a different approach than was used in the previous studies of Barbieri et al. (102) and Paper I, which set the number of hidden units to be equal to the number of b values. This increase in the number of hidden units ensured that the network had sufficient learning capacity and resulted in enhanced convergence speed. Additionally, using a lower learning rate ensured stable convergence.

Figure 3.3 shows a subset of further results of the simulation study in Paper II. This figure shows that extending the training process beyond early stopping resulted in reduced parameter correlations and parameter error, providing an alternative to exhaustive hyperparameter optimization. However, extensive training resulted in an increased sensitivity to noise, especially for parameter estimates with low SNR or low f (inherent uncertainty). Here, unsupervised parameter estimates exhibited variability similar to conventional LSQ fitting. Conversely, supervised learning demonstrated enhanced precision but was prone to a bias

towards the mean of the training distribution, which resulted in relatively smooth, yet possibly deceptive parameter maps. The in vivo results (Figure 3.4) were in broad agreement with the simulations, where fitting residuals were almost identical between approaches, particularly for the unsupervised networks. In addition, IVIM-NET_{optim} demonstrated inferior performance compared to the alternative approaches.

While the apparent sensitivity to noise in unsupervised learning may be undesirable, it could be argued that the corresponding variability observed in the parameter maps is indicative of the underlying uncertainty, which is indeed useful information. This uncertainty is exemplified by the contrasting D^* maps between approaches, despite the similar residual maps (Figure 3.4), and illustrates the difficulty in estimating D^* in the brain.

In conclusion, this study explored the impact of key training features in unsupervised and supervised learning for IVIM model fitting. It underscored that the effectiveness of these learning strategies is heavily dependent on design choices and emphasized the importance of conducting a thorough evaluation when using either approach. This comprehensive evaluation is essential for identifying and addressing potential biases in the process.

Summary of papers

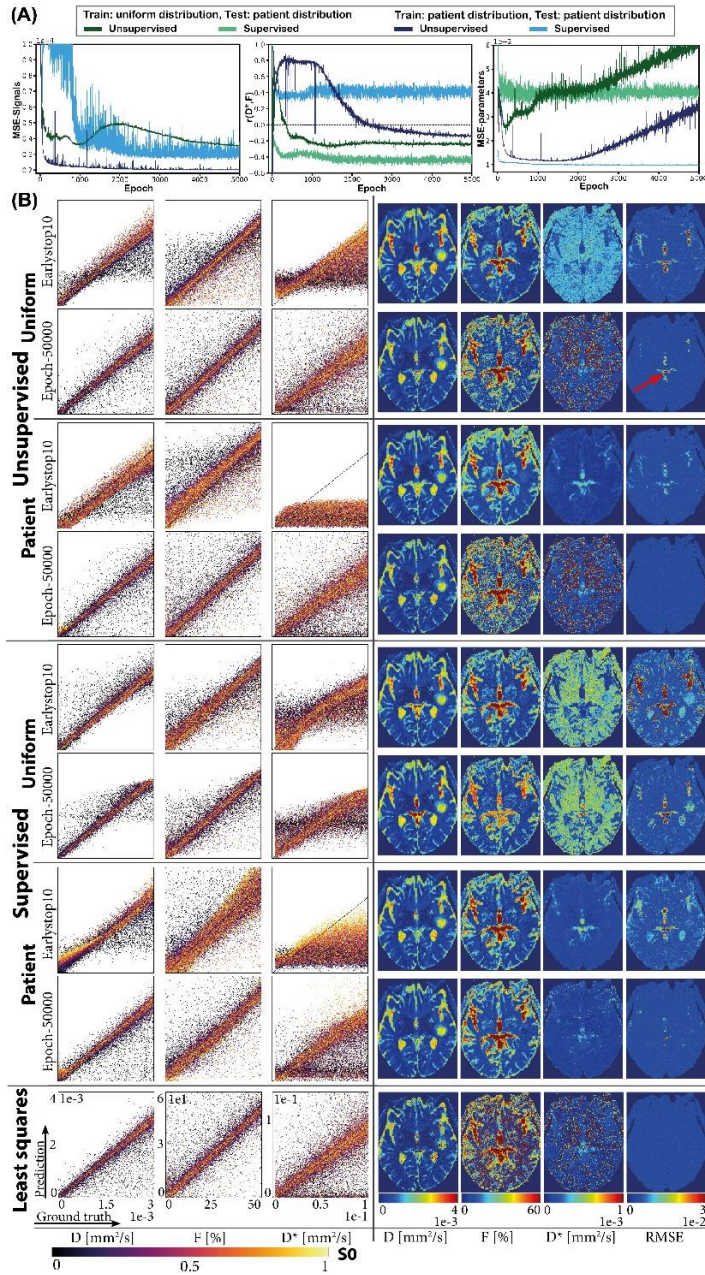


Figure 3.3: This figure shows a subset of the results of the simulation study performed in Paper II. It shows scatter plots of estimated parameter values (D , f , D^*) against ground truth (*left*), and IVIM parameter maps and RMSE maps (calculated between the predicted DWI signal and the measured DWI signal) for a representative slice of a synthetic glioma patient (*right*). Both scatter plots and parameter maps are displayed at the early-stopping point (Earlystop10) and at epoch 50000, representing full convergence. These maps are displayed for the four networks trained either unsupervised or supervised on synthetic data from either the uniform distribution (denoted ‘Uniform’ vertically) or patient distribution (denoted ‘Patient’ vertically). Corresponding least squares solutions (*bottom*)

and ground truth parameter maps (*right top*) are also shown. The red arrow in the RMSE maps indicate out-of-distribution data for the uniform distribution. The purple arrow in the ground truth $b=0$ image indicates the location of a tumor. The figure shows that at the early-stopping point both unsupervised and supervised learning are biased towards the mean of the training distribution. Extending the training beyond early stopping demonstrated that unsupervised approaches mimic least squares fitting, while for supervised learning the bias towards the training distribution remains.

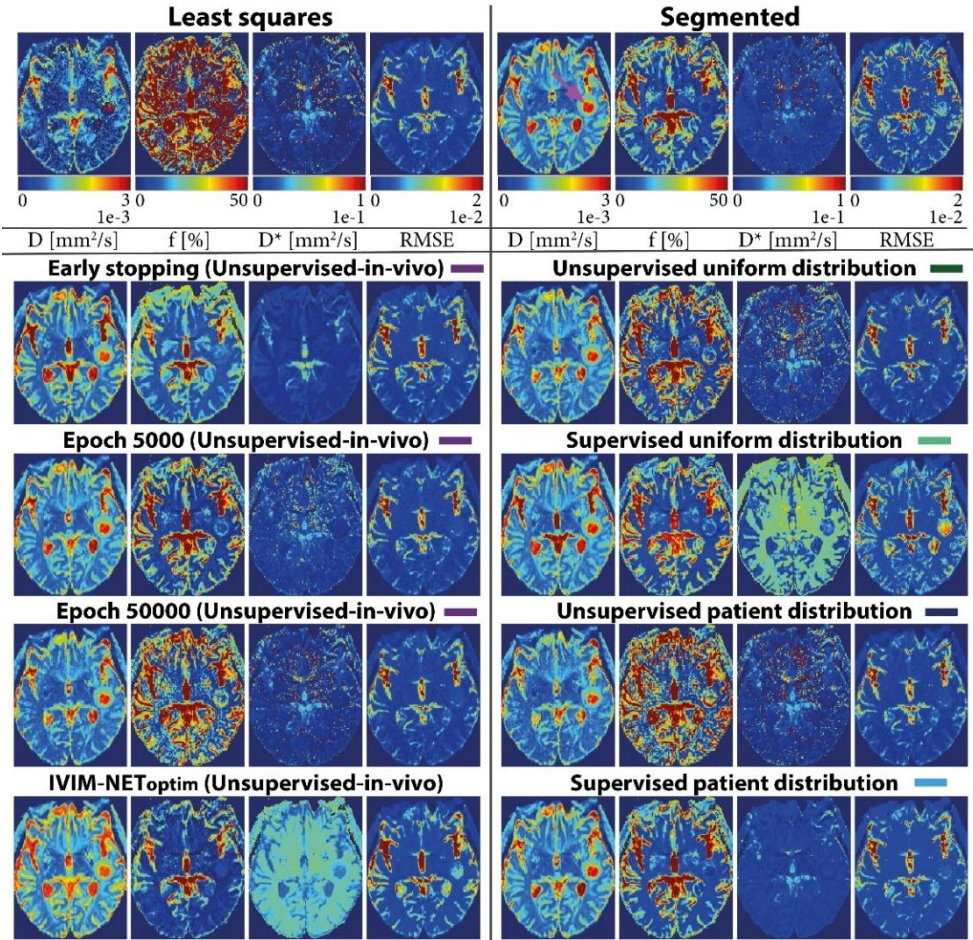


Figure 3.4: IVIM parameter maps and RMSE maps (calculated between the predicted DWI signal and the measured DWI signal) for a representative slice from the in vivo glioma patient data. These maps are displayed for the four networks trained either unsupervised or supervised on synthetic data from either the uniform distribution or patient distribution, plus an extra network that is trained directly on the in vivo signals (unsupervised). For the unsupervised network trained directly on the in vivo data (*left*), maps corresponding to three stopping points are displayed (Early stopping, Epoch-5000, and Epoch-50000), whereas the maps shown for the four synthetically-trained networks (*right*) are after training for 50,000 epochs. Corresponding maps are also shown for least squares (*top left*), the segmented approach (*top right*), and IVIM-NET_{optim} (*bottom left*). The purple arrow in the D map for the segmented approach indicates the location of a tumor.

3.3. Paper III

Incorporating spatial information in deep learning parameter estimation with application to the intravoxel incoherent motion model in diffusion-weighted MRI

Misha Pieter Thijs Kaandorp, Frank Zijlstra, Davood Karimi, Ali Gholipour, Peter T. While

Submitted to Medical Image Analysis (Dec. 2023)

The tissue microenvironment is generally characterized by local homogeneity, with properties like diffusion and perfusion maintaining consistency between adjacent voxels (17). Leveraging potential correlations between relevant signals in neighboring voxels should therefore enhance model-parameter fitting. However, estimators like LSQ and voxelwise deep neural networks, such as MLPs, do not incorporate such spatial information (see also Paper II). Contrastingly, CNNs and attention models (transformers) (96) are capable of incorporating such spatial information. In the context of parameter estimation, transformers have demonstrated superior performance over CNNs in capturing spatial correlations when applied to diffusion tensor imaging (DTI) (111). In this approach, networks were trained using a reconstructed high-quality DTI as ground truth for upscaling low-quality data. Yet, obtaining such high-quality can be challenging in many signal analysis problems, and noise may still influence the acquired data. In contrast, simulating training data allows the generation of any conceivable MRI signal from representative ground truths, and realistic correlations between neighbors can be introduced in a realistic fashion.

Building upon these observations, in paper III we explored means to incorporate spatial information into deep learning parameter estimation in biophysical modeling by training on synthetic data in patches. We demonstrated this approach for the IVIM model in DWI. The details of these investigations are outlined in four sub-studies:

1. Exploration of network architectures and learning strategies

In the first sub-study, various network architectures were explored for their potential to incorporate spatial information through either unsupervised or supervised training. The architectures included a voxelwise network (conventional MLP), a convolution-neighborhood network (MLP with a 2D-convolution in its first layer with kernel size 3), and a transformer-neighborhood network (employing self-attention). See Section 1.5.2 for more information about these architectures. These networks underwent training, both unsupervised or supervised, on DWI signals simulated by uniformly sampling 3×3 patches of IVIM parameters: $0 \leq S_0 \leq 1$, $0 \times 10^{-3} \leq D \leq 3 \times 10^{-3} \text{ mm}^2/\text{s}$, $0 \leq f \leq 50\%$, and $3 \times 10^{-3} \leq D^* \leq 100 \times 10^{-3} \text{ mm}^2/\text{s}$, considering 16 b values

(0, 10, 20, 40, 80, 110, 140, 170, 200, 300, 400, 500, 600, 700, 800, 900 mm²/s) using eq. (14). Rician noise was added to the signals such that when $S_0=1$ the SNR was 200. Each patch was generated with a random number of neighbors correlated to its center (neighbors-random, Figure 3.5A), meaning that those neighbors shared identical parameters. Additionally, nine subsets, denoted as ‘neighbors-N’ (N=0 to 8), were considered, where N represented a specific number of neighbors correlated to the center pixel (Figure 3.5C). The special case where all eight neighbors were correlated is denoted as ‘neighbors-all’ (Figure 3.6A).

We evaluated the performance of the networks trained on neighbors-random in terms of the convergence rate and the final loss, specifically signals-MSE for unsupervised learning (Section 1.5.3.2) and parameter-MSE for supervised learning (Section 1.5.3.1). This assessment utilized a neighbors-random test set comprising 100,000 patch-wise sets of DWI signals. The loss function was computed using the center pixel data. Additionally, each network was tested on the neighbors-all test set (Figure 3.6A), and the estimated parameters were compared with the ground truth. These results were also compared against parameters obtained through LSQ fitting, both on the center pixel data (i.e. voxelwise), as well as on the signal average of each 3×3 patch. A final comparison was made of the final loss among the three supervised-trained networks when applied to the test subsets containing specific numbers of correlated neighbors (neighbors-N).

In these simulation experiments, we found that only supervised learning derived benefits from incorporating spatial information, with no advantage observed for unsupervised learning (Figure 3.5B). Additionally, transformers outperformed CNNs in leveraging spatial information, demonstrating superior convergence speed and reduced loss (Figure 3.5B), and matching or surpassing the CNNs trained for each specific number of correlated neighbors (Figure 3.5C). Furthermore, Figure 3.6B shows that the supervised networks demonstrated an apparent reduced sensitivity to noise, akin to signal averaging. This is indicated by the narrowing of the spread of values along the diagonal for these supervised networks, where a similar narrowing of the spread of values was observed when applying LSQ to the signal average of the 3×3 patches. Therefore, in the subsequent sub-studies, we only considered supervised training of transformer-neighborhood networks for incorporating spatial information.

Summary of papers

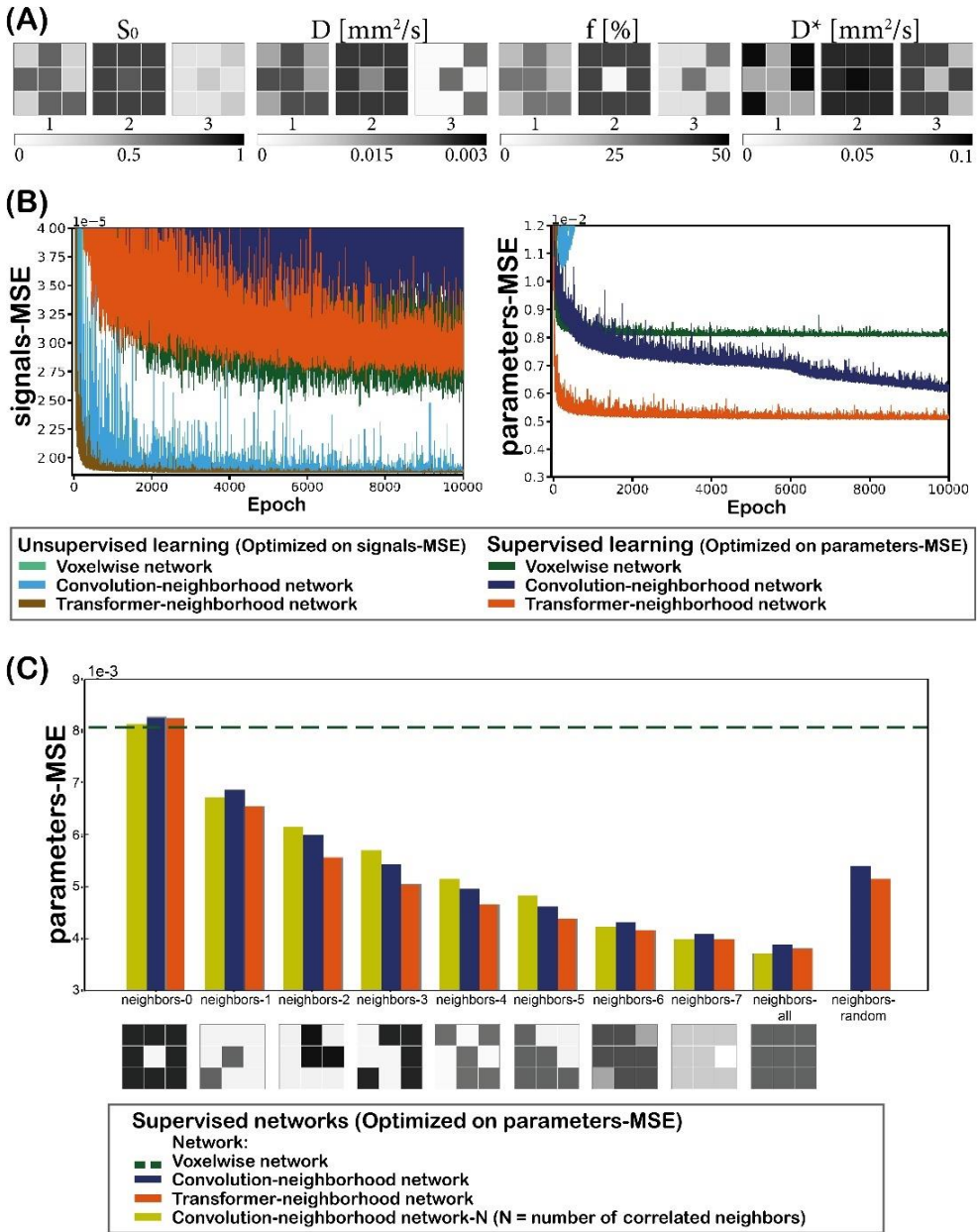


Figure 3.5: (A) Example 3×3 patches from the neighbors-random test set. (B) Test curves showing the metrics signals-MSE and parameters-MSE for the three networks (voxelwise network, convolution-neighborhood network, and transformer-neighborhood network) over 10,000 epochs. The networks were trained and tested unsupervised (optimized on signals-MSE) or supervised (optimized on parameters-MSE) on neighbors-random data. (C) Performance of the supervised networks trained on neighbors-random, evaluated for each specific number of correlated neighbors (neighbors-N). The convolution-neighborhood networks trained specifically for each test subset (yellow) represent the proxies for the optimal expected performance.

Summary of papers

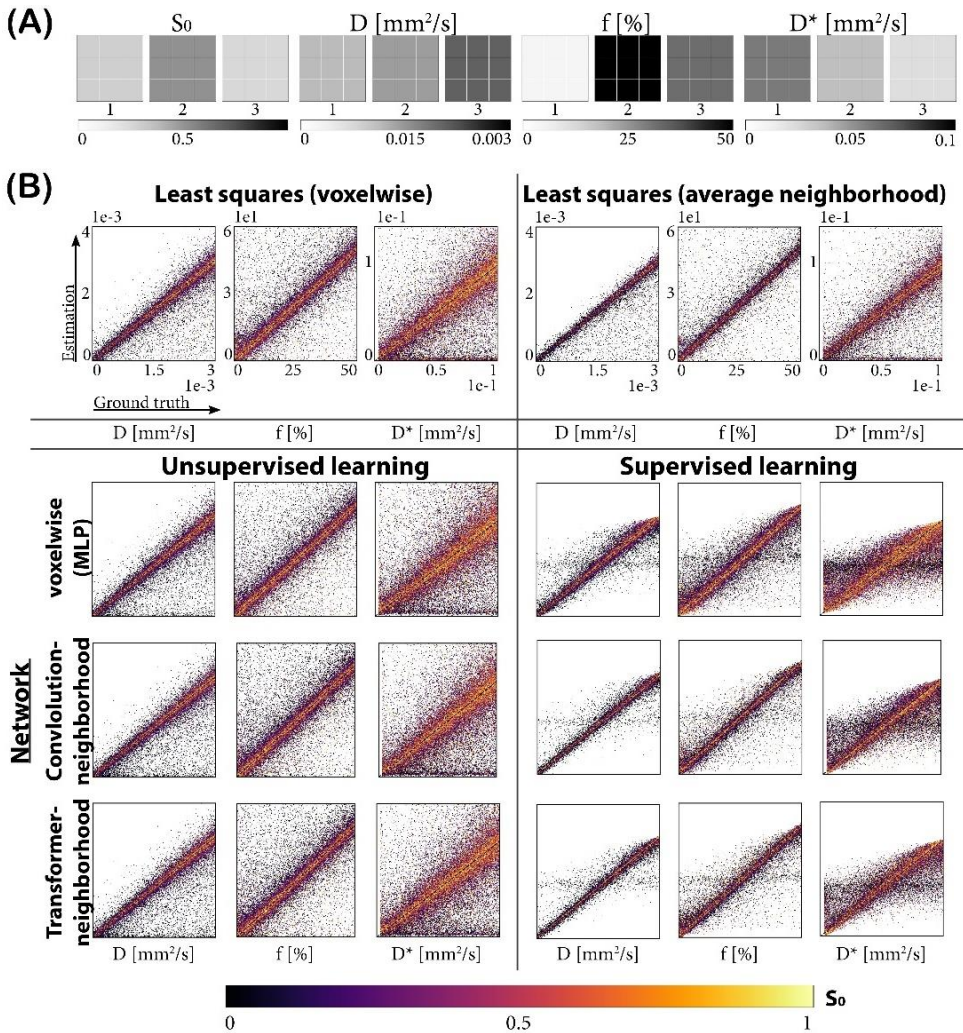


Figure 3.6: (A) Example 3×3 patches from the neighbors-all test set, where all neighbors are correlated in the patch (i.e. share identical underlying IVIM parameters). (B) Scatter plots comparing estimated parameter values against ground truth values for the networks described in Figure 3.5 tested on the neighbors-all test set, consisting of 100,000 3×3 patches of DWI signals. Corresponding plots are also shown for least squares applied to the center pixel of each patch (*top left*), and applied to the signal average of each 3×3 neighborhood (*top right*). All data points are colored by their S_0 -value, where $S_0=0$ (*black*) corresponds to SNR=0 and $S_0=1$ (*bright yellow*) corresponds to SNR=200.

2. Exploration of training with larger receptive fields and different attention concepts

In the second sub-study, we explored whether training supervised on synthetic data in patches with larger receptive fields can further improve the performance of transformers. We considered two types of transformers (Section 1.5.2.3): the first utilized self-attention (96); and the second neighborhood-attention (98). We trained both networks with increasing receptive field up to 17×17 , where self-X (transformer with self-attention) and NATTEN-X (transformer with neighborhood-attention) denote a receptive field of $X \times X$ voxels. Beyond a 7×7 receptive field, self-attention networks exhibited excessive memory usage. Therefore, for receptive fields of 7×7 and larger, neighborhood-attention was employed. We simulated synthetic training data in patches that matched the receptive field of the corresponding transformer-neighborhood network, following a similar approach to the data generation of the neighbors-random dataset in sub-study 1 above (i.e. following a uniform distribution, and patches with correlations generated in a random fashion), denoted as patch-uni.

Due to the lack of in vivo ground truths for a quantitative evaluation of our methods, we performed a quantitative assessment by testing the networks on 40 novel, synthetic, fractal-noise-based maps (128×128 pixels; 16 b values), which provided spatially-correlated ground truths (see full Paper III for details on how to generate these fractal-noise maps). Evaluation involved assessing parameter maps qualitatively. We performed a quantitative assessment by comparing median absolute percentage error (MDAPE), median percentage bias (MDPE), and mean absolute percentage error (MAPE). Comparisons were also made to LSQ.

Figure 3.7 shows a subset of the results of sub-study 2 in Paper III. The quantitative assessment using the fractal-noise maps demonstrated that training with larger receptive fields resulted in improved accuracy and reduced inherent supervised bias within low SNR regions, while preserving edge-like structures. The use of neighborhood-attention permitted training with larger receptive fields compared to self-attention, contributing to overall enhanced performance. Self-attention was limited to smaller receptive fields due to excessive memory demands.

Summary of papers

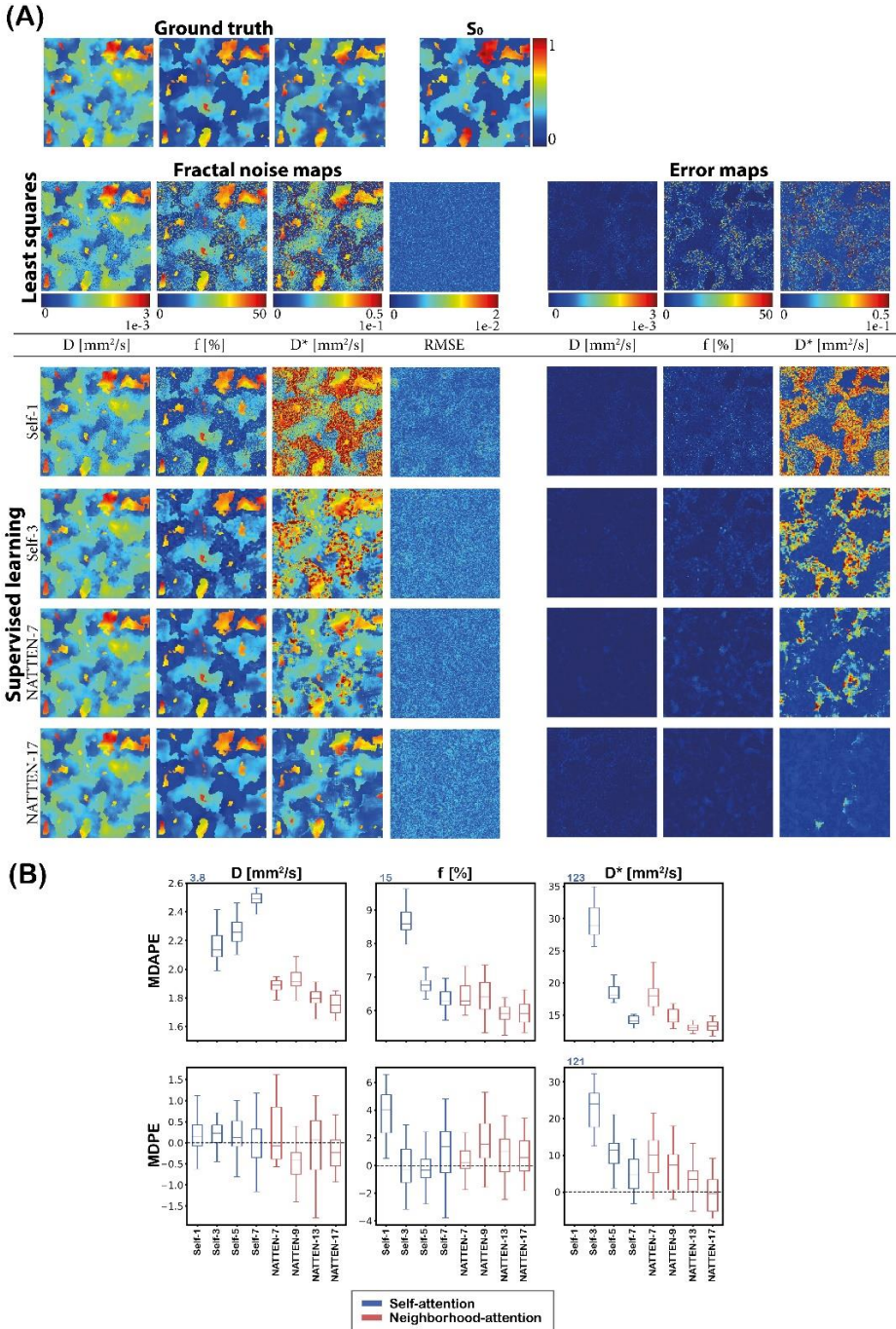


Figure 3.7: (A) Example IVIM fractal-noise parameter maps, RMSE maps, and error maps estimated for the transformer-neighborhood networks trained supervised with different receptive fields on patch-uni training data. Corresponding maps of the least squares fit and ground truth (*left top*) are also shown. (B) Boxplots displaying MDAPE and MDPE calculated for the transformer-neighborhood networks tested on the fractal-noise test set, evaluated over the last 20 epochs to illustrate variability.

3. Exploration of the representativeness of different training data

In the third sub-study, we explored additional considerations aimed at advancing our approach by generating training data that more closely resembles the characteristics of the test data. In order to quantitatively investigate our strategies for improving our methods, we utilized the same fractal-noise parameter maps of the previous exploration (see sub-study 2 above), providing us with spatially-correlated ground truths and control over the data generation process. We considered five different training sets in this sub-study (see full Paper III for more details on how to generate these training sets):

1. **Patch-uni:** This dataset is entirely independent of a test set (see sub-study 2).
2. **Mask-uni:** In this training set, we matched the underlying spatial variation of the test set. Here, we sampled patches from tissue masks of fractal-noise maps simulated with comparable random generation statistics as the test set. Subsequently, we sampled for each masked region in the patch parameter values from the uniform distribution.
3. **Patch-dist:** In this training set, we matched the underlying spatial variation of the test set. In this approach, we sampled parameter values from Gaussian distributions, which were defined to approximate the parameter values for each tissue type in the fractal-noise test set.
4. **Mask-dist:** In this training set, we matched both the spatial variation and the underlying distribution of parameter values of the test set. We generated mask-based patches in the same manner as mask-uni, and assigned parameter values for each tissue-specific mask by sampling from the Gaussian distributions defined for patch-dist.
5. **Test:** We generated patches of training data simulated in the same way as the test set.

We trained three transformer-neighborhood networks (Self-3, NATTEN-7, and NATTEN-17) on these training sets. Performance was evaluated on the fractal-noise test set, where we assessed MDAPE, MDPE, and MAPE.

Figure 3.8 shows the results of sub-study 3 in Paper III. This figure shows that transformers trained on synthetic possessing spatial variations more representative of the characteristics of the test set resulted in only marginal improvements in performance. Conversely, introducing more representative parameter distributions of the test set yielded a more substantial performance enhancement. It is possible that the inherent supervised bias contributed to this improved performance. Notably, these enhancements were most pronounced at smaller receptive fields (i.e., Self-3), whereas for larger receptive fields (NATTEN-17) improvements

were less apparent. This finding suggests that the use of representative masks and representative parameter distributions might be dispensable in the training process provided that a substantially large receptive field is used to mitigate supervised bias. In this context, training on all possible combinations of neighboring correlations and utilizing a uniform distribution could present a practical and more generalizable approach, facilitating easier implementation.

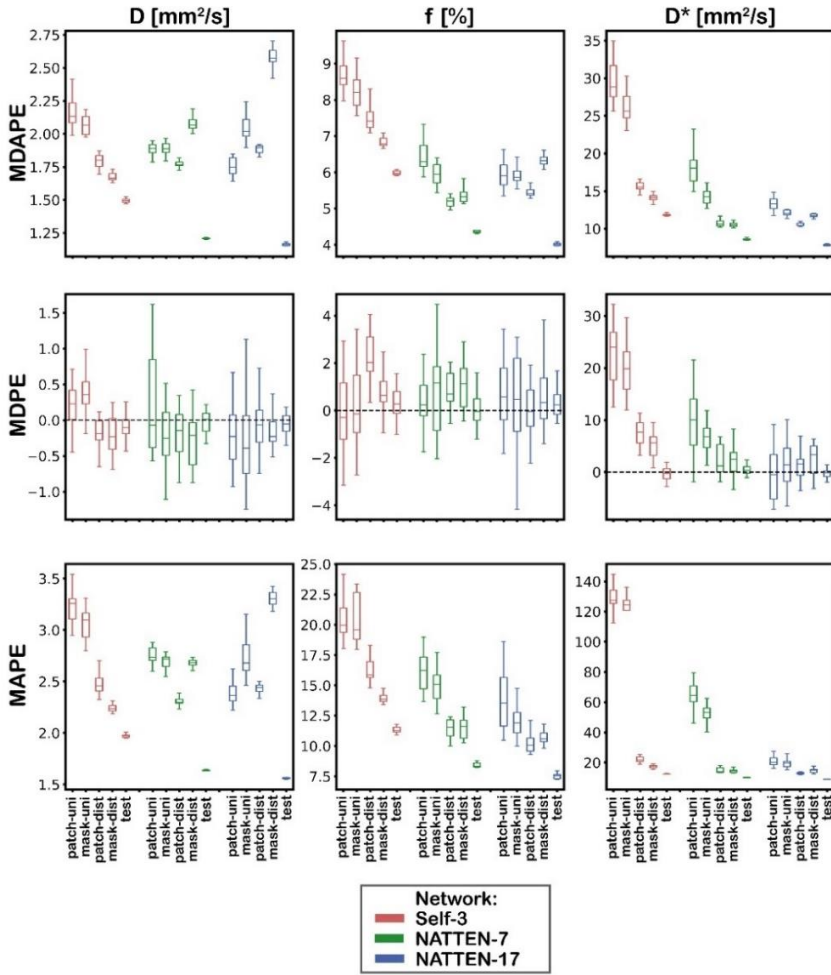


Figure 3.8: Boxplots displaying MDAPE, MDPE, and MAPE calculated for the transformer-neighborhood networks applied to the fractal-noise test set, evaluated over the last 20 epochs to illustrate variability. The metrics were calculated for transformer-neighborhood networks that were trained supervised using different synthetic data, ranging from patch-uni (trained independently of the test data) to data that is more representative of the test data. The latter involves leveraging information on the underlying spatial variation ('mask-') and/or the underlying distribution of the parameter values ('-dist') from the entire IVIM fractal-noise test set. The networks were trained with either self-attention (Self-3) or neighborhood-attention (NATTEN-7, NATTEN-17). As a baseline comparison, we also included results for networks trained on synthetic data that had been generated in the same way as the test set ('test').

4. In vivo analysis

In the final sub-study, we investigated the performance of our methods trained with different receptive fields and different parameter distributions (i.e. patch-uni and patch-dist as described in sub-study 3) using an in vivo brain dataset comprising multiple repetitions of acquired data. Our assessment involved a qualitative analysis of parameter maps and RMSE maps. Additionally, we calculated precision metrics across all repetitions for the transformer-neighborhood networks trained supervised with different receptive fields.

Figure 3.9 shows a sub-set of our in vivo study in this final sub-study, which is in broad alignment with the findings from our simulations (sub-study 2 and 3). Similarly to the findings in sub-study 2, we observed a lower D^* in white matter regions (e.g., low SNR regions) for networks trained with larger receptive fields. This observation suggests a potential reduction in inherent supervised bias and an improvement in estimator performance. Additionally, in line with the findings of sub-study 3, the parameter maps estimated by the networks trained on the different training distributions (e.g. patch-uni and patch-dist) exhibited increased similarity at larger receptive fields. Considering that in simulations, this heightened similarity correlated with reduced bias and enhanced accuracy (refer to Figure 3.8), we speculate that similar improvements in estimator performance may be extrapolated in the in vivo application for the networks with an increased receptive field.

In conclusion, this study presented a novel deep learning approach for parameter estimation in biophysical modeling, where neighborhood-attention models are trained supervised on synthetic data in patches possessing spatial correlations between neighboring pixels. This enabled attention models to effectively leverage spatial information, resulting in enhanced estimator performance compared to CNNs and conventional fitting methods. This approach brings challenging ill-posed signal analysis problems, such as IVIM, closer to clinical implementation. Additionally, the novel synthetic fractal-noise maps enabled quantitative assessment of our approaches, and may open up new possibilities for research and development within quantitative medical image analysis.

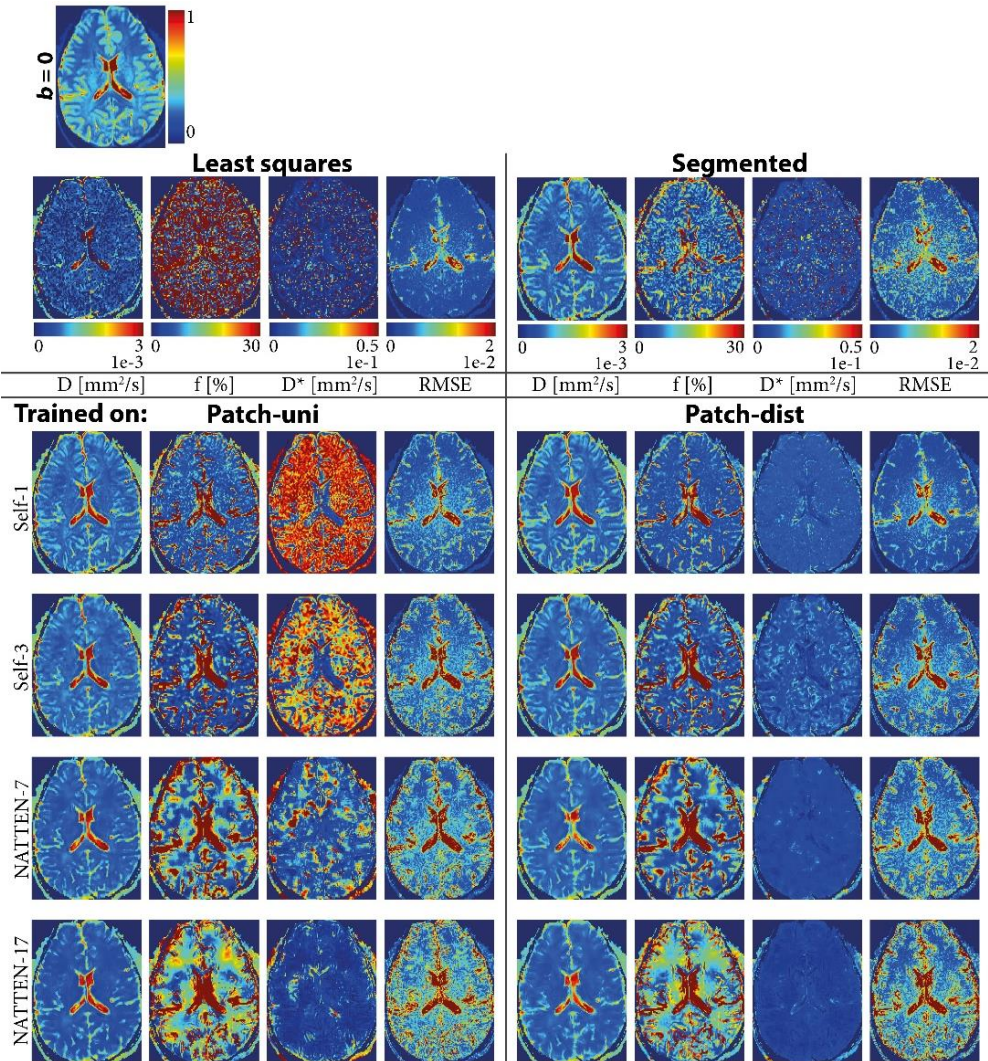


Figure 3.9: IVIM parameter maps and RMSE maps for a representative slice from the in vivo brain volunteer data, estimated by the transformer-neighborhood networks trained supervised with different receptive fields. These networks were trained using either a uniform distribution for each parameter (patch-uni; *left*) or a distribution that is more representative of the in vivo data (patch-dist; *right*). The networks were trained with either self-attention (Self-1, Self-3) or neighborhood-attention (NATTEN-7, NATTEN-17). Also shown is the $b = 0$ image (*top left*), along with corresponding maps for the least squares and segmented approaches.

Chapter 4

Discussion

Throughout this thesis, we have recognized limitations of conventional fitting methods in achieving clinically reliable diffusion and perfusion parameter estimates for clear, detailed maps of cancer biomarkers from DWI data. This underscored the compelling need for innovative approaches within the diffusion modeling framework. Consequently, the primary objective of this thesis has been to investigate the application of deep learning for this purpose. The preceding chapters and scientific papers of this doctoral thesis have conducted an in-depth exploration of deep learning diffusion parameter estimation, with a particular focus on the IVIM model for DWI.

This discussion chapter provides a critical analysis of the findings of this thesis, explores their implications, and considers potential future directions for research. These are structured around the central themes and subjects explored in the thesis. However, note that these themes are interconnected and interdependent.

4.1 Learning strategies for deep learning parameter estimation

4.1.1. Unsupervised learning

While the study of Barbieri et al. (102) and the research conducted in Paper I suggested that unsupervised learning held great promise for IVIM parameter estimation, subsequent research in Paper II and Paper III have raised substantial questions on the performance of this training method. Paper II revealed that early stopping in unsupervised learning can lead to correlated and biased parameter estimates. It further showed that when training was extended, unsupervised estimates displayed a variability similar to least squares. Furthermore, Paper III demonstrated that unsupervised learning does not benefit from incorporating spatial information. Notably, despite the improved performance of the optimized unsupervised network from Paper I (IVIM-NET_{optim}) in the pancreas, it exhibited poor performance when applied to liver data (112) and when applied to brain data in both Paper II and in a study conducted by Spinner et al. (109). These findings question the reliability of IVIM-NET_{optim}.

Various other unsupervised approaches have been explored in the context of IVIM parameter estimation (113–120). For instance, one study proposed an unsupervised CNN for IVIM fitting (117). Other studies modified the proposed IVIM-NET of Paper I for three-compartment models (118,119), or utilized unsupervised deep learning approaches for dynamic-exponential IVIM modeling (120). Additionally, unsupervised networks have found application in DCE-MRI for the prediction of the Extended Tofts-Kety model-parameters (121). Considering the limitations identified for unsupervised learning in this thesis, the effectiveness of these approaches may need careful evaluation, similar to the concerns raised for IVIM-NET_{optim}.

4.1.2. Supervised learning

Supervised learning presents its own set of challenges when it comes to parameter estimation. Paper II demonstrated that supervised IVIM parameter estimates exhibit inherent biases towards the mean of the training distribution, especially for regions with low SNR. This bias is particularly noticeable when the model is trained on a uniform distribution, aligning with research performed by Gyori et al. (122), who reported similar findings for the spherical SMT (47) model. These intrinsic training biases might potentially manifest in other microstructural model-fitting approaches employing supervised learning (123–126), including the first application of supervised learning for IVIM fitting (127).

While supervised learning is prone to a training bias, Paper III demonstrated that it is possible to mitigate this bias by training on spatially-correlated synthetic data. By incorporating spatial

information from the relevant neighbors, the sensitivity to noise of the supervised networks was apparently reduced in a manner akin to signal averaging. This resulted in enhanced accuracy, marking a unique strength of supervised learning that distinguishes it from other approaches.

To conclude, both unsupervised and supervised deep learning approaches bring unique advantages and challenges to the domain of diffusion parameter estimation. The apparent sensitivity to noise of unsupervised deep learning methods, while possibly undesirable, does offer insights into the underlying uncertainty, which is valuable information for analysis. For future research directions, it is of crucial importance to underscore the advantages of incorporating spatial information through supervised deep learning approaches for diffusion parameter estimation, leading ultimately to improved estimator performance.

Considering that the decision-making process of many deep learning-based methods is typically a black box (128,129), ensuring a comprehensive assessment is imperative. Therefore, quantifying uncertainty (130,131) should be strongly prioritized, as solely relying on visual assessment may not suffice. From a clinical perspective, uncertainty quantification can provide clinicians with a measure of confidence associated with predictions or diagnostic information. This allows them to make more informed and cautious decisions. Future work should consider integrating uncertainty quantification into deep learning diffusion parameter estimation to further enhance our understanding of the potential of these approaches in a clinical setting.

4.2 Training and test data for deep learning parameter estimation

4.2.1. Training data

The most straightforward approach for training data in the context of deep learning diffusion parameter estimation is to use the in vivo data, i.e. unsupervised learning. However, findings from Paper II and Paper III unveil a crucial limitation of this approach when applied to IVIM. The in vivo DWI signal is corrupted by noise and therefore the network has no access to uncorrupted ground truth. Hence, unsupervised parameter estimates are always contaminated by this noise. Consequently, training on such in vivo signals is unlikely to yield accurate underlying parameter estimates.

An alternative to using in vivo signals as ground truth involves employing simulated ground truth data, which is utilized in supervised learning techniques. In this thesis, two distinct methods for generating distributions of synthetic training data are explored: a uniform synthetic distribution; and a patient-derived synthetic distribution. Training on a uniform distribution offers the advantage of encompassing a structured and wide range of parameter values, facilitating a comprehensive training of the entire parameter space. It is important to note that the inherent supervised bias is particularly observable for the uniform distribution as this distribution misaligns with the characteristics of a real in vivo test dataset, as particularly demonstrated in Paper II.

The alternative approach where a DNN is trained supervised on a distribution of parameters derived from in vivo patient data has been utilized in Paper II and by others (132,133). This method involves using predefined parameter distributions, derived from conventional fitting approaches applied to in vivo data, to train the network. This distribution is more likely to resemble the in vivo test set distribution than the uniform distribution. As supervised approaches are inherently biased towards the training distribution, such training strategies might give the appearance of improved performance. However, as elucidated in paper II, this approach leads to a shift in the mean of the training distribution that comes closer towards that of the test set. Hence, the inherent supervised bias remains and is only apparently reduced. This can still create the illusion of smoother and reliable parameter maps, while remaining entirely misleading. Furthermore, these approaches also rely on assumptions that conventional fitting methods can offer reliable ground truth values, despite their well-known sensitivity to noise. As discussed in Paper III, an argument could be made in favor of training on a uniform distribution, highlighting its potential for generalizability across various test datasets and

anatomical variations, as this approach is not dependent on prior assumptions of a specific test set.

Another important concept explored in this thesis is integrating prior knowledge regarding spatial variation from the test set into the synthetic training data. As highlighted in Paper III, the tissue microenvironment typically exhibits local homogeneity. Therefore, in Paper III, we developed deep learning approaches that can leverage such spatial information to improve estimator performance. As there is a lack of ground truth data, we generated synthetic training data in patches with spatially-correlated ground truth. By training on such synthetic data, the networks learned to identify relevant neighboring voxels and leverage the correlated information to improve estimator performance. Therefore, this approach of generating synthetic data is highly recommended for improved deep learning diffusion parameter estimation.

Another key aspect highlighted in this thesis is the importance of training data being representative of the test data. The inherent uncertainty introduced by out-of-distribution data (ODD), i.e. data lying outside the training distribution but within the test distribution, can pose challenges in network predictions and calibration (134). Guo et al. (135) discussed how a lack of exposure to diverse input data and patterns can lead to overconfident predictions and misclassifications. One approach to characterize the DNN's response to ODD data is employing uncertainty modeling techniques (136), such as dropout Bayesian approaches (137).

ODD data arises primarily due to limited data or knowledge about the system being modeled. In this thesis, particularly in Paper III, synthetic data proves instrumental in addressing this 'epistemic uncertainty' by not relying on a physical dataset but by generating representative signals that are potentially present in the test set. However, generating reliable synthetic data requires precise knowledge of the test data, posing a challenge in diffusion signal modeling given the numerous processes affecting the in vivo DWI signal, as highlighted in Paper III. These include various physiological noise effects (not only IVIM), artifacts, and partial volume effects. Importantly, conventional estimators disregard these effects. However, a notable advantage of synthetic data approaches lies in their adaptability to incorporate such external influences in a flexible and controllable manner, potentially enhancing estimator performance.

4.2.2. Test data

Just as the selection of training data demands careful deliberation, the choice of test data for performance evaluation necessitates thorough consideration, with most factors overlapping those relevant to training data. While it might be tempting to employ in vivo data as test data for validation due to its direct relevance to the application, this approach presents a fundamental

limitation. In vivo data lacks ground truths for quantitative assessment, rendering it capable of providing only qualitative insights. Given the absence of reliable ground truth, in vivo results could be potentially misinterpreted and misleading.

To enable a quantitative analysis of parameter estimation approaches, it is imperative to incorporate ground truth data in the evaluation phase. Utilizing synthetic data generated with a uniform distribution of parameter values for evaluation can serve as an important benchmark for evaluating DNN behavior. As demonstrated in Paper II, this approach can effectively highlight limitations or biases that may emerge during the learning process. Furthermore, these assessments can be particularly well visualized through ground truth versus estimated parameter scatter plots. However, it is important to note that while a uniform distribution may suffice for voxelwise approaches, it may not provide a comprehensive evaluation of the performance of networks that leverage spatial information.

Paper III presented a novel approach for quantitative assessment that incorporates spatial information through the utilization of synthetically generated fractal-noise parameter maps. These maps offer spatially-correlated ground truth data, a critical element in the quantitative evaluation of the methods and key findings outlined in Paper III. Without this evaluation, the deep learning methods developed in Paper III would only be subjected to a qualitative assessment using in vivo data, known to hide biases in deep learning and therefore less appropriate.

As highlighted above, it is important that synthetic data is representative of the system being modeled, and thus to real in vivo data. Paper III underscored that in vivo data typically exhibits slight variations in parameter values within each tissue type. Consequently, fractal-noise can be harnessed to introduce variability in simulated parameter estimates, offering a potentially more realistic representation of tissue heterogeneity and diversity compared to using fixed values. These variations could be incorporated when generating patch-based training data. Note, these could also be introduced for patch-uni or patch-dist (see Paper III for details), as attention models are permutation-invariant, rendering the spatial structure within a patch irrelevant for learning correlations among neighbors. This versatility makes fractal-noise a desirable approach in quantitative medical image analysis.

The variations introduced by fractal-noise could also be applied to synthetic data with a more spatially realistic appearance. As an example, a digital brain phantom, as proposed by Guerquin-Kern et al. (138), could be used where tissue variations are generated by fractal-noise rather than a fixed value. Similarly, in approaches like Synthmap (139), which generates

realistically appearing synthetic brain data with tissue segmentations and specific naturally occurring parameters for each tissue type, fractal-noise can replace a fixed value for each tissue type. Indeed, such spatially realistically appearing datasets can serve both for training and testing. However, it is important to note that using such datasets for training might have limited generalizability to other anatomies compared to, for example, training on patch-uni or patch-dist. As underscored above, various physiological noise effects affecting the in vivo DWI signal (not only IVIM) can be incorporated into synthetic data, and thus in synthetically fractal-noise generated maps, to enhance its authenticity.

To conclude, the selection and characterization of training and test datasets are critical aspects in deep learning parameter estimation, that profoundly impact the performance and generalizability of DNNs. Relying solely on in vivo data for training and test data is not sufficient due to the lack of underlying ground truth. Simulated data offers ground truth for training and testing. However, important considerations should be made to ensure the representativeness of simulated training data to real test data, including the distribution of parameter values and the spatial variance. Thus, understanding and addressing out-of-distribution (ODD) data becomes pivotal for constructing robust deep learning networks.

Moving forward, future research should explore possibilities of integrating processes that influence the in vivo DWI signal, extending beyond IVIM, into the synthetic data approaches proposed in this thesis. This exploration should aim to produce a higher quality synthetic dataset that better represents in vivo data. Noteworthy, these distinctive capabilities differentiate our approaches from other estimators that neglect these intricate processes, offering a more comprehensive and adaptable approach to model-parameter estimation.

Additionally, future work should explore further the potential of fractal-noise within medical imaging. The ability to introduce realistic variability with minimal user-defined parameters, particularly advantageous in deep learning that requires large datasets for training, underscores the promise of fractal-noise. Synthetic data of this nature are not only promising for avenues in parameter estimation, but could also open up new possibilities of research in the field of quantitative medical image analysis. For instance, fractal-noise could be utilized in the evaluation and data generation of super-resolution (140) or model-based reconstruction (141,142) algorithms.

4.3 Network architectures for deep learning parameter estimation

In this doctoral thesis, we have explored the use of several network architectures for diffusion parameter estimation, including multi-layer perceptrons (MLPs), convolutional neural networks (CNNs), and transformers.

Within this thesis, MLPs have been thoroughly explored for diffusion parameter estimation. MLPs proved to be highly adaptable and computationally efficient, providing a robust framework to investigate various learning strategies and training hyperparameters for deep learning IVIM parameter estimation. However, it is important to note that MLPs do not take into consideration the spatial dependencies inherent in diffusion data. Consequently, MLPs may be better suited as a faster alternative to conventional least squares methods, which also do not incorporate spatial information.

In contrast, as evidenced in Paper III, CNNs and transformers can incorporate spatial information by exploiting relationships between diffusion signals in neighboring voxels when trained in a supervised manner. Paper III demonstrated that transformers outperformed CNNs in terms of accuracy and convergence speed for deep learning diffusion parameter estimation (applied to IVIM). These findings align with the study conducted by Davood et al. (111), which demonstrated that a transformer network outperformed a CNN when trained on high-quality diffusion tensor imaging data for reconstructing the diffusion tensor from low-quality data. This prompts for a reconsideration of CNNs as the default operators for computer vision tasks in medical imaging.

The superiority of transformers can be attributed to their attention mechanism, enabling them to effectively capture complex long-range spatial dependencies while being permutation-invariant. In contrast, CNNs exhibit an inductive bias on locality which impairs their ability to capture long-range spatial dependencies in an image. While this bias has been instrumental in the success of CNNs, it might not be ideal for parameter estimation, potentially introducing complications in identifying correlated neighbors as emphasized in Paper III.

This thesis explored the application of self-attention (96) and the recently proposed neighborhood-attention (98) within transformers for deep learning parameter estimation. Neighborhood-attention enabled training with larger receptive fields than self-attention, potentially incorporating more spatially-correlated neighbors into the network's optimization process, leading to improved fitting. As noted in Paper III, neighborhood-attention introduces

overlapping neighborhood-attention blocks with a specific kernel size in its layers, potentially complicating the identification of correlated neighbors within a patch, in a manner similar to CNNs, due to the local inductive biases. However, self-attention is constrained to smaller receptive fields due to memory constraints. Ultimately, neighborhood-attention achieved superior accuracy by training on larger receptive fields. Thus, in scenarios where transformer networks can be trained on larger receptive fields beyond the capacity of self-attention, opting for neighborhood-attention networks may be the more advantageous choice.

Alternative transformer concepts that have not been explored in this thesis are vision transformers (ViTs) (100). ViTs, introduced by Dosovitskiy et al. in the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition", interpret images as a sequence of patches and process them through a standard self-attention transformer. ViT networks can capture long-range dependencies in images based on self-attention. ViTs have found use in many applications within medical imaging, such as classification (143), segmentation (144–147), and object detection (148), where they outperformed other CNN approaches. Intriguingly, recent research suggests that ViTs exhibit errors more consistent with those of humans than CNNs (149–151). As tissues are typically locally homogenous and may not exhibit long-range dependencies across the entire image, the focus of this thesis was not on ViTs. In Paper III, we focused on networks that emphasize local correlations between neighboring voxels. Nevertheless, future work should explore the potential of ViTs in deep learning diffusion parameter estimation.

One possible limitation of ViTs is that they require pre-training on large datasets, and most ViT-based networks, including TransUNet (152), are pre-trained on ImageNet (153). When initialized with random weights instead of pre-training, CNNs tend to outperform ViTs (154). Pre-training on ImageNet may be suboptimal due to the substantial domain gap between natural and medical imaging modalities. Research by Xie et al. (155) demonstrated that their ViT network, pre-trained on large datasets of medical images, exhibited a substantial improvement in medical image classification compared to training on ImageNet. This underscores the importance of obtaining large datasets in medical imaging to fully harness the potential of deep learning techniques and transformers. However, medical imaging datasets are substantially smaller than natural image datasets, primarily due to factors such as cost, privacy concerns, and the scarcity of certain diseases. This makes the efficient training of DNNs in the medical domain challenging. These findings emphasize the critical importance of designing a training dataset that closely aligns with the test data, as discussed in Section 4.2.

A noteworthy extension of transformers involves hybrid transformer-CNN approaches, aiming to leverage the combined strengths of transformers, such as capturing long-range dependencies, and CNNs, adept at capturing local patterns. One well-known example is TransUNet (152), along with other approaches (156–159), specifically designed for image segmentation. TransUNet has been applied in various image segmentation tasks, including setting a new record (at the time of publication) on the automated cardiac diagnosis challenge (160). Future research should explore whether hybrid transformer-CNN approaches can further improve the performance of deep learning diffusion parameter estimation. However, as indicated by our findings in Paper III, the attention mechanism might just prove superior when exploiting relationships between diffusion signals in neighboring voxels compared to convolution approaches.

To conclude, these findings emphasize the importance of selecting an appropriate network architecture tailored to the specific demands and characteristics of the diffusion parameter estimation task. While research regarding the use of transformers in diffusion parameter estimation has been somewhat limited, the growing prominence of transformers in computer vision (97), including the work performed in this thesis, indicates substantial promise for further advancing their integration into deep learning diffusion parameter estimation. Further research is imperative to unlock the full potential of transformers and assess their suitability in diverse clinical settings.

4.4 Hyperparameters for deep learning parameter estimation

Hyperparameters play a pivotal role in the performance of DNNs for diffusion parameter estimation. These parameters are not learned from the data but are set prior to training, affecting various aspects of the training process and the network's generalization capability. In this thesis, particularly in Papers I and II, the impact of several key hyperparameters were explored on the accuracy and convergence of DNNs.

4.4.1 Learning rate

Our experiments highlight the pivotal role of selecting an appropriate learning rate in training DNNs. Learning rate is related to the stochastic nature of the DNN training process due to various sources of randomness during training (161). This stochasticity arises from factors such as the random initialization of weights and the shuffling of training data batches. These stochastic elements contribute to the variability observed in training the DNNs, even when trained on the same dataset multiple times, as demonstrated throughout this thesis. Achieving consistency in performance is essential, as it is undesirable for a network to predict a valid solution at one time and a suboptimal solution at another. Recognizing this, Paper I identified the need to quantify stochasticity as a performance measure, referred to as consistency. Here, consistency was evaluated by training the network 50 times on identical data with random initializations. Koopmans et al. (162) applied a similar strategy by training an IVIM-NET network 100 times. Engaging in multiple training sessions is indeed a highly time-consuming and exhausting pursuit, prompting the exploration of more efficient approaches.

One of the straightforward approaches to mitigate this variability in training is to alter the learning rate. Paper II demonstrated that a higher learning rate might lead to faster convergence but may also induce undesirable oscillating behavior with big spikes. This leads to poor consistency, where the optimization process might 'overshoot' the minimum of the loss function. Therefore, caution is advised when using high learning rates. Conversely, a low learning rate enhances consistency, but excessively low rates should be avoided to prevent sluggish network convergence. Hence, there is a need for careful tuning in the selection of learning rates.

While our focus in this thesis has been on fixed learning rates to maintain experiment consistency, it is generally acknowledged that dynamic learning rates can substantially improve convergence speed. A common dynamic learning rate is a monotonically decreasing learning rate scheduler, where the learning rate decreases at a fixed number of epochs by a factor or

percentage. Another widely used dynamic learning rate is the ‘reduce learning rate on plateau’ (ReduceLRonPlateau) technique, which adjusts the learning rate when a plateau in network performance is detected. It is crucial to note that these methods should not be applied to different training sets in the same manner. Each application requires careful tuning, as these approaches are susceptible to prolonged plateaus in test loss.

As an example, in Paper III, a prolonged plateau appeared in the test loss for the convolution-neighborhood network before a subsequent dip occurred (at epoch 6000 in Figure 1). Therefore, a poor configuration of these dynamic learning rate methods could lead to the underestimation of network performance, resulting in suboptimal outcomes. Moreover, these methods heavily depend on the number of batches per epoch, which may not be desirable. It could be argued that a well-tuned learning rate, combined with a ReduceLRonPlateau method towards the end of the network may offer the optimal balance between achieving a fast and consistent network.

Another noteworthy dynamic learning rate technique is ‘cyclical learning rates’ (163), where the learning rate oscillates during training between reasonable boundary values, allowing the network to escape from local minima more effectively and potentially converge faster.

4.4.2 Network capacity

Concerning the capacity of DNNs, the structural aspects of a deep learning network, particularly its depth and width, profoundly influence its ability to capture intricate data relationships. In Paper I, an approach was taken where the number of hidden units equated to the number of measured b values. This approach has also been employed by several other studies (102,118,119). However, subsequent research within this thesis has led to the conclusion that this approach is not recommended. Paper II underscored that such a choice can severely limit the capacity of DNNs. During the initial exploration phase, emphasis should be placed on the selection of a configuration suitable for the given task. While a suitable configuration of these network hyperparameters encompasses a relatively wide range, it is advisable to explore several configurations in terms of eventual loss and computational demands. Considering network configurations that lean towards greater capacity could be advantageous to ensure that the network’s complexity is sufficient and prevent underestimating its learning potential, even if this comes at the expense of increased computational demands.

Concerning activation functions, Paper II showed that utilizing different activation functions between each layer in the network did not produce a considerable performance improvement, provided that the network was trained until true convergence.

4.4.3 Regularization

Regularization techniques such as early stopping, batch normalization (107), and dropout (108), were initially suggested as network hyperparameters that mitigate overfitting and enhance the performance of IVIM-NET_{optim} in Paper I. However, Paper II unveiled some crucial insights regarding these methods. Firstly, it was discovered that early stopping, if applied improperly by terminating the training process before reaching true convergence, could introduce biases and correlations in parameter estimation. These correlations were also present in the original unsupervised deep learning IVIM approach of Barbieri et al (102), as highlighted in Paper I. To address this issue, Paper II found that extending training beyond the point of early stopping could resolve these undesirable correlations and reduce errors. In addition, Paper II also indicated potential drawbacks of using batch normalization or dropout (also the stochastic gradient descent or SGD optimizer). Paper II demonstrated that these methods could even lead to suboptimal performance in certain cases. Therefore, caution and comprehensive evaluation are advised when using these regularization techniques.

4.4.4 Concerning hyperparameter optimization and IVIM-NET

This section provides a critical overview of key aspects of IVIM-NET_{optim}, considering recent discoveries.

In the case of IVIM-NET_{optim}, sigmoid activation functions were employed to the output of the network. The rationale behind introducing these sigmoid activation functions was to confine the parameter values within biophysically plausible ranges, which was intended to facilitate the convergence of the network towards more accurate solutions. However, it was observed that in regions characterized by low SNR and thus heightened uncertainty, parameter estimates predicted by IVIM-NET_{optim} are apparently biased towards the center of the sigmoid functions. It is worth noting that this phenomenon has not been formally documented in the scientific literature or the scientific papers within this thesis. Yet, this apparent bias of the sigmoid functions can explain the unexpectedly high values of D^* observed in anatomies where IVIM-NET_{optim} has been applied. For instance, IVIM-NET_{optim} demonstrated elevated D^* in white matter (e.g. low SNR regions) when applied to brain data (109) (also observed in Paper II), and similar high D^* values when applied to liver data (112). Consequently, these findings discourage the use of sigmoid activation functions to the output of the network. Instead, our findings indicate that employing absolute value activations functions to the output of the network, as a means to avoid the possibility of exploding gradients due to negative outputs, represents a viable alternative.

Another substantial factor in the unforeseen shortcomings of IVIM-NET_{optim} was the fact that the network was exclusively optimized within simulated settings, where RMSE was solely used as a measure of accuracy. However, as emphasized by Epstein et al. (132), RMSE is considered an insufficient performance measure alone. The sensitivity to outliers of RMSE tends to favor methods that consistently produce parameter estimates close to the mean parameter values, thereby endorsing biased approaches that estimate parameters towards those means. In the specific case of IVIM-NET_{optim}, the center of the sigmoid functions closely mirrored the mean of the test distribution in the performed simulations of Paper I. Combined with the network's shallow architecture and the early-stopping criterion, this resulted in an estimator that predicted estimates closely aligned with the mean of the test distribution, yielding a low RMSE, but an apparently biased approach. Therefore, and in line with the findings of Epstein et al. (132), it is advisable not to rely solely on (low) RMSE as the exclusive metric for evaluating parameter estimation performance. Instead, it should always be accompanied by an evaluation of bias and variance. Paper II demonstrated that such evaluations can be effectively visualized through scatter plots that compare ground truth parameters to the estimated values.

To conclude, hyperparameter tuning is a crucial aspect in the development of DNNs for diffusion parameter estimation. Techniques like grid search can be employed to systematically search the hyperparameter space and identify the optimal configuration. However, it is worth noting that this process can be time-consuming and exhausting. This thesis found that the impact of individual hyperparameters appeared to be reduced after training extensively. Future research should explore whether regularization techniques can considerably enhance the performance of deep learning diffusion parameter estimation, rather than solely expediting convergence. Ultimately, the primary focus should be on designing a robust training dataset, as hyperparameters cannot compensate for a low-quality dataset.

4.5 Communication, exploitation, and reproducibility

The primary means of communicating the findings presented in this thesis was via the publication of research articles in international scientific journals. Both Paper I (164) and Paper II (165) have been published in the prestigious Journal of Magnetic Resonance in Medicine (MRM), one of the top-tier journals specifically focusing on clinical MRI research with an emphasis on methodology. It is noteworthy that Paper I received a certificate as being one of the top research articles in MRM, making an immediate impact on the community. In addition, Paper I was highlighted and selected by the International Society for Magnetic Resonance in Medicine (ISMRM) to appear in a virtual issue of MRM entitled “Diffusion MRI: From Research to Clinic” (166), highlighting the evolution of the field over the previous three years.

One of the important characteristics of Paper I is the provision of openly accessible code upon publication (167), which is both comprehensive and well-documented. This code encompasses a diverse range of examples, from simulations to in vivo experiments. This availability facilitated the reproducibility, validation, and advanced the collective knowledge in the field. The methods from Paper I have been implemented and adapted in several studies (109,112,118,119), demonstrating the impact of Paper I on subsequent research endeavors. The codes for both Paper II (168) and Paper III (169) have also been openly accessible. Given that deep learning is a relatively new concept within medical imaging and parameter estimation, sharing code plays a pivotal role in familiarizing researchers with this innovative technology.

Considering recent assessments and discoveries related to the approach advocated in Paper I, it is important to highlight that declaring a method as a groundbreaking innovation might be premature without thorough evaluations across various anatomies. Therefore, while the approaches presented in Paper III show great potential, we have adopted a cautious stance, refraining from advocating a singular best method. Instead, we have taken a more suggestive approach, acknowledging the potential while emphasizing the necessity for further research.

4.6 Research ethics

Ethical considerations hold paramount significance in every scientific endeavor, particularly when dealing with sensitive medical data. This section on research ethics is primarily based on the following review papers in medical ethics (170,171) and their relation with AI (172–174).

Fundamental to research involving human subjects is the ethical principle of obtaining informed consent. Participants have the ethical right to comprehensive information regarding the research, including its objectives, potential benefits, and associated risks. Participants should be assured of their right to withdraw from the study at any time without facing adverse consequences. Institutional and local ethics committees play a pivotal role in safeguarding the rights and well-being of research participants under these principles. These committees should meticulously assess the benefits, harms, and risks associated with both the systems to be developed and the acquired data. Researchers and developers of AI methodologies have the ethical responsibility of upholding the dignity and rights of the participants. This commitment extends to securing patient data, refraining from its commercialization, and ensuring its utilization aligns with the ethical consent provided by the participants.

Within the context of this thesis, all participants contributing to the utilized datasets - a pancreatic tumor dataset from Amsterdam UMC hospital (Netherlands), a brain tumor dataset from the University of Lausanne (Switzerland), and a brain dataset from a healthy volunteer at St. Olav's University Hospital (Norway) – have granted written informed consent. Moreover, the studies have received ethical approval from local medical ethics committees, affirming a steadfast adherence to ethical standards throughout the entire research process.

Anonymization is another essential step to safeguard the privacy of individual participants. Following data collection, it is crucial to implement security measures to guarantee the confidentiality and integrity of the medical data. Restricted access to datasets for authorized personnel, along with data storage in secure environments adhering to local and international data protection regulations, is paramount. In this thesis, these guidelines were carefully followed. Prior to data utilization, personally identifiable information, including names and medical record numbers, underwent removal or encryption to proactively mitigate any potential identification risks. Access to the datasets was restricted to authorized personnel and data was securely stored on local servers.

It is important to highlight that this thesis extensively utilized synthetically generated data. In addition to the quantitative benefit, the use of synthetic data offered the advantage of creating

diverse representative datasets for training and testing without compromising the privacy of individuals. Ensuring a diverse range of data is crucial to prevent the development of AI models with a narrow perspective and to facilitate generalization. Acquiring a substantial variety often demands large in vivo datasets, and synthetic data circumvents this. It must be noted that ethical research values inclusivity and diversity as a means to promote equitable outcomes.

Sharing data between institutions is often challenging for several reasons related to privacy, security risks, interoperability challenges, and regulatory concerns. Consequently, the processes of sharing and obtaining relevant data are often time-consuming. However, synthetic datasets allow researchers to openly collaborate and contribute to collective knowledge without compromising ethical standards regarding data ownership and confidentiality.

In line with this, there is a strong recommendation to make datasets and code publicly available. This promotes transparency, reproducibility, collaboration, education and training, and the advancement of scientific knowledge. Indeed, it is crucial to adhere to all ethical considerations as described above when engaging in this practice, including informed consent and anonymization. In this thesis, in vivo examples, code for generating synthetic data, and other relevant code have been made publicly available on GitHub (167–169). It is strongly encouraged to pursue this standard of open science in research.

Chapter 5

Conclusion

The primary objective of this thesis was to improve the robustness, accuracy, and precision of diffusion and perfusion-related parameters from DWI data. The overarching goal was to improve the mapping of cancer biomarkers through the application of deep learning. The specific focus was on improving the parameter estimation of the intravoxel incoherent motion (IVIM) model within diffusion-weighted imaging (DWI), offering an alternative to potentially harmful dynamic contrast-enhanced MRI (DCE-MRI) methods.

Fitting the IVIM model to DWI data is a challenging ill-posed signal analysis problem, primarily due to the low SNR inherent to DWI and the low perfusion fractions typically encountered. Conventional fitting methods such as least squares, the segmented approach, and Bayesian inference are highly sensitive to noise, rely on model assumptions, computational demands, and inherent biases, making them unsuitable for clinical use. This has prompted the exploration of alternative fitting methods within the context of IVIM. In this pursuit, deep learning emerges as a promising solution, with the potential to overcome challenges in diffusion and IVIM parameter estimation. Moreover, deep learning proves itself to be a versatile tool, offering a myriad of possibilities and adaptability to user needs. Given the flexibility of deep learning, customization and validation become crucial for optimal system performance. This thesis extensively explored various facets of deep learning, providing a comprehensive evaluation of learning strategies such as unsupervised and supervised learning. It also thoroughly explored various other aspects of deep learning, including different training and test data, hyperparameters, and the utilization of various deep learning architectures such as MLPs, CNNs, and transformers.

While unsupervised learning for DNNs may appear straightforward and promising, this thesis demonstrated that optimizing DNNs on measured and noisy signals results in parameter

Conclusion

estimates that display a variability similar to least squares. A crucial factor is to ensure that the network is trained until true convergence to alleviate training bias and correlated parameters. The variability similar to least squares fitting was consistently observed regardless of the training data or network architecture employed, where unsupervised learning showed no noticeable benefit from leveraging spatial information when employing CNNs or transformer neural networks. Additionally, it is essential to note that unsupervised deep learning approaches introduced in this thesis, such as IVIM-NET_{optim} in Paper I, may lead to suboptimal performance when applied to different anatomies. Based on these findings, unsupervised deep learning may not be the preferred approach for improved parameter estimation.

Conversely, supervised approaches have been considered susceptible to training data bias and are generally discouraged within the community. However, this thesis introduced an innovative supervised approach that can effectively leverage neighboring information to improve supervised deep learning diffusion parameter estimation. This approach, where attention models are trained supervised on spatially-correlated synthetic data in patches, can efficiently leverage correlations within neighboring voxels when present. This strategy apparently reduces the sensitivity to noise in a similar manner to signal averaging, resulting in a substantial improvement in parameter estimation and a reduction in inherent supervised bias. Furthermore, a particular advantage of the proposed method is its sensitivity to genuine underlying heterogeneity, a feature notably absent in other conventional fitting methods. These features make the proposed method highly desirable and show the potential of supervised learning for diffusion parameter estimation

A pivotal aspect of parameter estimation involves appropriately validating the designed estimators. While using in vivo signals as test data for validation has direct relevance to the application, it presents a fundamental limitation. In vivo data lacks ground truths for quantitative assessment, offering only qualitative insights that could potentially be misconstrued and misleading. Such qualitative assessments could obscure biases, particularly important in deep learning, where approaches might smooth away pathologies. In quantitative MRI, the pursuit of obtaining reliable quantitative parameter maps necessitates ground truth data for quantitative evaluation.

In response to this challenge, this thesis presents an innovative approach: fractal-noise maps possessing spatially-correlated ground truth. This novel approach has yielded crucial findings within this thesis, demonstrating that neighborhood-attention networks trained on synthetic data with larger receptive fields further improve the performance of supervised deep learning

Conclusion

diffusion parameter estimation, as exemplified for the IVIM model. Additionally, fractal-noise may open up new avenues the field of quantitative medical image analysis.

In conclusion, this thesis has substantially advanced the field of deep learning diffusion parameter estimation, marking a substantial step towards the clinical implementation of complex diffusion signal analysis problems, especially in the context of IVIM for DWI. These advancements go beyond the scope of cancer, holding potential for various other diseases characterized by substantial changes in diffusion and perfusion tissue characteristics. This indicates a promising avenue for enhancing capabilities in screening, early detection, diagnosis, and treatment monitoring across a spectrum of medical conditions.

This thesis serves as a cornerstone for future research endeavors aimed at refining the reliability, accuracy, and precision of our methods. Initial exploration should concentrate on enhancing the representativeness of synthetic data, both in training and testing, to align more closely with *in vivo* data. This can be achieved by incorporating prior assumptions about the *in vivo* data into the network's training process, potentially enhancing overall performance. In this pursuit, fractal-noise could offer realistic representations of tissue heterogeneity and diversity. Furthermore, subsequent research should prioritize uncertainty quantification, offering interpretable and explainable approaches to guide fellow researchers, physicians, and clinicians in deciding whether to adopt and trust the proposed methods.

However, perhaps the most important future step should involve a comprehensive comparative study to ascertain whether DNNs confer clinical advantages over other estimators. This study should particularly emphasize clinical performance and alignment with the expectations of radiologists. Crucially, it must encompass diverse anatomies with varying noise levels and tissue characteristics, complemented by histological characterization and tumor classification. Ideally, these datasets should be made open source to address the scarcity of clinical data and encourage open research. This comprehensive clinical assessment is pivotal for establishing a trusting relationship between radiologists and AI models, laying the foundation for the practical implementation of deep learning diffusion and IVIM parameter estimation in clinical settings.

Chapter 6

References

1. Egeblad M, Nakasone ES, Werb Z. Tumors as organs: Complex tissues that interface with the entire organism. *Dev. Cell* 2010;18:884–901 doi: 10.1016/j.devcel.2010.05.012.
2. Hanahan D, Weinberg R. The hallmarks of cancer. *Cell* 2000;57:70 doi: 10.1107/S2059798322003928.
3. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell* 2011;144:646–674 doi: 10.1016/j.cell.2011.02.013.
4. Hanahan D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* 2022;12:31–46 doi: 10.1158/2159-8290.CD-21-1059.
5. Anderson NM, Simon MC. The tumor microenvironment. *Curr. Biol.* 2020;30:R921–R925 doi: 10.1016/j.cub.2020.06.081.
6. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* 2021;71:209–249 doi: 10.3322/caac.21660.
7. Carioli G, Bertuccio P, Boffetta P, et al. European cancer mortality predictions for the year 2020 with a focus on prostate cancer. *Ann. Oncol.* 2020;31:650–658 doi: 10.1016/j.annonc.2020.02.009.
8. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA. Cancer J. Clin.* 2023;73:17–48 doi: 10.3322/caac.21763.
9. Westbrook C, Talbot J. *MRI in Practice*. John Wiley & Sons; 2018.
10. Purcell EM, Torrey HC, Pound R V. Resonance Absorption by Nuclear Magnetic Moments in a Solid. *Phys. Rev.* 1946;69:681.
11. Bloch F. Nuclear induction. *Phys. Rev.* 1946;70:460–474 doi: 10.1103/PhysRev.70.460.
12. Mansfield P, Grannell PK. NMR Diffraction in Solids. *J Phys C Solid State* 1973;6, L422–L4 doi: 10.1088/0022-3719/6/22/007.
13. Lauterbur PC. Image formation by induced local interactions: examples employing nuclear magnetic resonance. *Nature* 1973;242:190–191 doi: 10.1038/242190a0.
14. Ridgway JP. Cardiovascular magnetic resonance physics for clinicians: part I. *J. Cardiovasc. Magn. Reson.* 2010;12(1), 1–2 doi: 04.2009/JCPSP.207210.

15. Schmitt F, Stehling MK, Turner R, Kiefer B. Turbo Spin-Echo Imaging. In: Echo-Planar Imaging: Theory, Technique and Application. ; 1998. pp. 583–604.
16. Olsson LE, Johansson M, Zackrisson B, Blomqvist LK. Basic concepts and applications of functional magnetic resonance imaging for radiotherapy of prostate cancer. *Phys. Imaging Radiat. Oncol.* 2019;9:50–57 doi: 10.1016/j.phro.2019.02.001.
17. Novikov DS, Fieremans E, Jespersen SN, Kiselev VG. Quantifying brain microstructure with diffusion MRI: Theory and parameter estimation. *NMR Biomed.* 2019;32:1–53 doi: 10.1002/nbm.3998.
18. Chavhan GB, Babyn PS, Thomas B, Shroff MM, Mark Haacke E. Principles, techniques, and applications of T2*-based MR imaging and its special applications. *Radiographics* 2009;29:1433–1449 doi: 10.1148/rg.295095034.
19. Mariappan YK, Glaser KJ, Ehman RL. Magnetic resonance elastography: A review. *Clin. Anat.* 2010;23:497–511 doi: 10.1002/ca.21006.
20. Manduca A, Bayly P V., Ehman RL, et al. MR elastography: Principles, guidelines, and terminology. *Magn. Reson. Med.* 2021;85:2377–2390 doi: 10.1002/mrm.28627.
21. Khalifa F, Soliman A, El-Baz A, et al. Models and methods for analyzing DCE-MRI: A review. *Med. Phys.* 2014;41 doi: 10.1118/1.4898202.
22. Gordon Y, Partovi S, Müller-Eschner M, et al. Dynamic contrast-enhanced magnetic resonance imaging: fundamentals and application to the evaluation of the peripheral perfusion. *Cardiovasc. Diagn. Ther.* 2014;4:147–64 doi: 10.3978/j.issn.2223-3652.2014.03.01.
23. Hylton N. Dynamic contrast-enhanced magnetic resonance imaging as an imaging biomarker. *J. Clin. Oncol.* 2006;24:3293–3298 doi: 10.1200/JCO.2006.06.8080.
24. Carmeliet P, Jain RK. Angiogenesis in cancer and other diseases. 2000:249–257.
25. Tofts PS, Brix G, Buckley DL, et al. Estimating kinetic parameters from dynamic contrast-enhanced T1- weighted MRI of a diffusable tracer: Standardized quantities and symbols. *J. Magn. Reson. Imaging* 1999;10:223–232 doi: 10.1002/(SICI)1522-2586(199909)10:3<223::AID-JMRI2>3.0.CO;2-S.
26. O’Connor JPB, Jackson A, Parker GJM, Roberts C, Jayson GC. Dynamic contrast-enhanced MRI in clinical trials of antivascular therapies. *Nat. Rev. Clin. Oncol.* 2012;9:167–177 doi: 10.1038/nrclinonc.2012.2.
27. Grobner T. Gadolinium - A specific trigger for the development of nephrogenic fibrosing dermopathy and nephrogenic systemic fibrosis? *Nephrol. Dial. Transplant.* 2006;21:1104–1108 doi: 10.1093/ndt/gfk062.
28. Elmholdt TR, Pedersen M, Jørgensen B, et al. Nephrogenic systemic fibrosis is found only among gadolinium-exposed patients with renal insufficiency: A case-control study from Denmark. *Br. J. Dermatol.* 2011;165:828–836 doi: 10.1111/j.1365-2133.2011.10465.x.
29. Cheong BYC, Wilson JM, Preventza OA, Muthupillai R. Gadolinium-Based Contrast Agents: Updates and Answers to Typical Questions Regarding Gadolinium Use. *Texas Hear. Inst. J.* 2022;49:7–9 doi: 10.14503/THIJ-21-7680.
30. Ray JG, Vermeulen MJ, Bharatha A, Montanera WJ, Park AL. Association between MRI exposure during pregnancy and fetal and childhood outcomes. *JAMA - J. Am. Med. Assoc.* 2016;316:952–961 doi: 10.1001/jama.2016.12126.
31. Weinreb JC, Rodby RA, Yee J, et al. Use of intravenous gadolinium-based contrast media

- in patients with kidney disease: Consensus statements from the American college of radiology and the national kidney foundation. *Radiology* 2021;298:28–35 doi: 10.1148/RADIOL.2020202903.
32. Kanda T, Fukusato T, Matsuda M, et al. Gadolinium-based contrast agent accumulates in the brain even in subjects without severe renal dysfunction: Evaluation of autopsy brain specimens with inductively coupled plasma mass spectroscopy. *Radiology* 2015;276:228–232 doi: 10.1148/radiol.2015142690.
33. Partridge SC, DeMartini WB, Kurland BF, Eby PR, White SW, Lehman CD. Quantitative diffusion-weighted imaging as an adjunct to conventional breast MRI for improved positive predictive value. *Am. J. Roentgenol.* 2009;193:1716–1722 doi: 10.2214/AJR.08.2139.
34. Messina C, Bignone R, Bruno A, et al. Diffusion-weighted imaging in oncology: An update. *Cancers (Basel)*. 2020;12:1–28 doi: 10.3390/cancers12061493.
35. Chilla GS, Tan CH, Xu C, Poh CL. Diffusion weighted magnetic resonance imaging and its recent trend-a survey. *Quant. Imaging Med. Surg.* 2015;5:407–22 doi: 10.3978/j.issn.2223-4292.2015.03.01.
36. White NS, McDonald C, Farid N, et al. Diffusion-weighted imaging in cancer: Physical foundations and applications of Restriction Spectrum Imaging. *Cancer Res.* 2015;74:4638–4652 doi: 10.1158/0008-5472.CAN-13-3534.Diffusion-weighted.
37. Johansen-Berg H, Behrens TE. *Diffusion MRI: from quantitative measurement to in vivo neuroanatomy.* Elsevier; 2009.
38. Brown R. Xxvii. a brief account of microscopical observations made in the months of june, july and august 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *Philos. Mag.* 1828.
39. Fick A. Über diffusion. *Ann Phys* 1955.
40. Fick A. Concerns diffusion and concentration gradient. *Ann Phys Lpz* 1955.
41. Einstein A. Über die von der molekularkinetischen Theorie der wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann Phys.* 1905;4:549–560.
42. Einstein A. *Investigations on the Theory of the Brownian Movement.* New York Dover Publ. Inc 1926.
43. Stejskal EO, Tanner JE. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. *J. Chem. Phys.* 1965;42:288–292 doi: 10.1063/1.1695690.
44. Le Bihan D, Breton E, Lallemand D, Grenier P, Cabanis E, Laval-Jeantet M. MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology* 1986.
45. Steven AJ, Zhuo J, Melhem ER. Diffusion kurtosis imaging: An emerging technique for evaluating the microstructural environment of the brain. *Am. J. Roentgenol.* 2014;202:26–33 doi: 10.2214/AJR.13.11365.
46. Zhang H, Schneider T, Wheeler-Kingshott CA, Alexander DC. NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage* 2012;61:1000–1016 doi: 10.1016/j.neuroimage.2012.03.072.
47. Kaden E, Kelm ND, Carson RP, Does MD, Alexander DC. Multi-compartment microscopic diffusion imaging. *Neuroimage* 2016;139:346–359 doi: 10.1016/j.neuroimage.2016.06.002.

-
48. Fieremans E, Jensen JH, Helpert JA. White matter characterization with diffusional kurtosis imaging. *Neuroimage* 2011;58:177–188 doi: 10.1016/j.neuroimage.2011.06.006.
 49. Le Bihan D. Intravoxel Incoherent Motion Perfusion MR Imaging: A Wake-Up Call. *Pharm. J.* 2008.
 50. Ordidge R. The development of echo-planar imaging (EPI): 1977-1982. *Magn. Reson. Mater. Physics, Biol. Med.* 1999;9:117–121 doi: 10.1016/S1352-8661(99)00056-3.
 51. Hammernik K, Klatzer T, Kobler E, et al. Learning a variational network for reconstruction of accelerated MRI data. *Magn. Reson. Med.* 2018;79:3055–3071 doi: 10.1002/mrm.26977.
 52. Le Bihan D. What can we see with IVIM MRI? *Neuroimage* 2019;187:56–67 doi: 10.1016/j.neuroimage.2017.12.062.
 53. Federau C, Cerny M, Roux M, et al. IVIM perfusion fraction is prognostic for survival in brain glioma. *Clin. Neuroradiol.* 2017;27:485–492 doi: 10.1007/s00062-016-0510-7.
 54. Federau C, Meuli R, O'Brien K, Maeder P, Hagmann P. Perfusion measurement in brain gliomas with intravoxel incoherent motion MRI. *Am. J. Neuroradiol.* 2014;35:256–262 doi: 10.3174/ajnr.A3686.
 55. Klaassen R, Steins A, Gurney-Champion OJ, et al. Pathological validation and prognostic potential of quantitative MRI in the characterization of pancreas cancer: preliminary experience. *Mol. Oncol.* 2020;14:2176–2189 doi: 10.1002/1878-0261.12688.
 56. Kim B, Lee SS, Sung YS, et al. Intravoxel incoherent motion diffusion-weighted imaging of the pancreas: Characterization of benign and malignant pancreatic pathologies. *J. Magn. Reson. Imaging* 2017;45:260–269 doi: 10.1002/jmri.25334.
 57. Li Y, Wang Z, Chen F, et al. Intravoxel incoherent motion diffusion-weighted MRI in patients with breast cancer: Correlation with tumor stroma characteristics. *Eur. J. Radiol.* 2019;120 doi: 10.1016/j.ejrad.2019.108686.
 58. Cui Y, Li C, Liu Y, et al. Differentiation of prostate cancer and benign prostatic hyperplasia: comparisons of the histogram analysis of intravoxel incoherent motion and monoexponential model with in-bore MR-guided biopsy as pathological reference. *Abdom. Radiol.* 2020;45:3265–3277 doi: 10.1007/s00261-019-02227-5.
 59. Noij DP, Martens RM, Marcus JT, et al. Intravoxel incoherent motion magnetic resonance imaging in head and neck cancer: A systematic review of the diagnostic and prognostic value. *Oral Oncol.* 2017;68:81–91 doi: 10.1016/j.oraloncology.2017.03.016.
 60. Li YT, Cercueil JP, Yuan J, Chen W, Loffroy R, Wang YXJ. Liver intravoxel incoherent motion (IVIM) magnetic resonance imaging: A comprehensive review of published data on normal values and applications for fibrosis and tumor evaluation. *Quant. Imaging Med. Surg.* 2017;7:59–78 doi: 10.21037/qims.2017.02.03.
 61. Le Bihan D, Mami I, Federau C, Sigmund EE eds. *Intravoxel incoherent motion (IVIM) MRI: principles and application.* CRC Press; 2018.
 62. Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, Laval-Jeantet M. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. *Radiology* 1988;168:497–505.
 63. Lemke A, Stieltjes B, Schad LR, Laun FB. Toward an optimal distribution of b values for intravoxel incoherent motion imaging. *Magn. Reson. Imaging* 2011;29:766–776 doi: 10.1016/j.mri.2011.03.004.

-
64. K. Levenberg. A method for the solution of certain non-linear problems in Least squares. *Q. Appl. Math.* 1944;2:164–168 doi: 10.1177/002205741408001401.
65. Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. *An algorithm least-squares Estim. nonlinear parameters* 1963;11(2), 431.
66. Moré JJ, Sorensen DC. Computing a trust region step. *SIAM J. Sci. Stat. Comput.* 1983;4:553–572.
67. Coleman TF, Li Y. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.* 1996;6:418–445 doi: 10.1137/0806023.
68. Pekar J, Moonen CTW, van Zijl PCM. On the precision of diffusion/perfusion imaging by gradient sensitization. *Magn. Reson. Med.* 1992;23:122–129 doi: 10.1002/mrm.1910230113.
69. Penner AH, Sprinkart AM, Kukuk GM, et al. Intravoxel incoherent motion model-based liver lesion characterisation from three b-value diffusion-weighted MRI. *Eur. Radiol.* 2013;23:2773–2783 doi: 10.1007/s00330-013-2869-z.
70. Concia M, Sprinkart AM, Penner AH, et al. Diffusion-weighted magnetic resonance imaging of the pancreas: Diagnostic benefit from an intravoxel incoherent motion model-based 3 b-value analysis. *Invest. Radiol.* 2014;49:93–100 doi: 10.1097/RLI.0b013e3182a71cc3.
71. Merisaari H, Movahedi P, Perez IM, et al. Fitting methods for intravoxel incoherent motion imaging of prostate cancer on region of interest level: Repeatability and gleason score prediction. *Magn. Reson. Med.* 2017;77:1249–1264 doi: <https://dx.doi.org/10.1002/mrm.26169>.
72. Cho GY, Moy L, Zhang JL, et al. Comparison of fitting methods and b-value sampling strategies for intravoxel incoherent motion in breast cancer. *Magn. Reson. Med.* 2015;74:1077–1085 doi: 10.1002/mrm.25484.
73. Suo S, Lin N, Wang H, et al. Intravoxel incoherent motion diffusion-weighted MR imaging of breast cancer at 3.0 tesla: Comparison of different curve-fitting methods. *J. Magn. Reson. Imaging* 2015;42:362–370 doi: 10.1002/jmri.24799.
74. Meeus EM, Novak J, Withey SB, Zarinabad N, Dehghani H, Peet AC. Evaluation of intravoxel incoherent motion fitting methods in low-perfused tissue. *J. Magn. Reson. Imaging* 2017;45:1325–1334 doi: 10.1002/jmri.25411.
75. Rice JA. *Mathematical statistics and data analysis*. Cengage Learning; 2006.
76. Barbieri S, Donati OF, Froehlich JM, Thoeny HC. Impact of the calculation algorithm on biexponential fitting of diffusion-weighted MRI in upper abdominal organs. *Magn. Reson. Med.* 2016;75:2175–2184 doi: 10.1002/mrm.25765.
77. Gurney-Champion OJ, Klaassen R, Froeling M, et al. Comparison of six fit algorithms for the intravoxel incoherent motion model of diffusion-weighted magnetic resonance imaging data of pancreatic cancer patients. *PLoS One* 2018;13:1–18 doi: 10.1371/journal.pone.0194590.
78. Orton MR, Collins DJ, Koh DM, Leach MO. Improved intravoxel incoherent motion analysis of diffusion weighted imaging by data driven Bayesian modeling. *Magn. Reson. Med.* 2014;71:411–420 doi: 10.1002/mrm.24649.
79. Freiman M, Perez-Rossello JM, Callahan MJ, et al. Reliable estimation of incoherent motion parametric maps from diffusion-weighted MRI using fusion bootstrap moves. *Med. Image Anal.* 2013;17:325–336 doi: 10.1016/j.media.2012.12.001.
80. Barbieri S, Donati OF, Froehlich JM, Thoeny HC. Comparison of intravoxel incoherent

motion parameters across MR imagers and field strengths: Evaluation in upper abdominal organs. *Radiology* 2016;279:784–794 doi: 10.1148/radiol.2015151244.

81. Kakite S, Dyvorne H, Besa C, et al. Hepatocellular carcinoma: Short-term reproducibility of apparent diffusion coefficient and intravoxel incoherent motion parameters at 3.0T. *J. Magn. Reson. Imaging* 2015;41:149–156 doi: 10.1002/jmri.24538.

82. Lecler A, Savatovsky J, Balvay D, et al. Repeatability of apparent diffusion coefficient and intravoxel incoherent motion parameters at 3.0 Tesla in orbital lesions. *Eur. Radiol.* 2017;27:5094–5103 doi: 10.1007/s00330-017-4933-6.

83. While PT. A comparative simulation study of bayesian fitting approaches to intravoxel incoherent motion modeling in diffusion-weighted MRI. *Magn. Reson. Med.* 2017;78:2373–2387 doi: 10.1002/mrm.26598.

84. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016.

85. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 2012 doi: 10.1201/9781420010749.

86. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. *Multimed. Tools Appl.* 2023;82:3713–3744 doi: 10.1007/s11042-022-13428-4.

87. Oord A van den, Dieleman S, Zen H, et al. WaveNet: A generative model for raw audio. 2016:1–15.

88. Xiong W, Wu L, Allea F, Droppo J, Huang X, Stolcke A. The Microsoft 2017 Conversational Speech Recognition System. *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.* 2018;2018-April:5934–5938 doi: 10.1109/ICASSP.2018.8461870.

89. Decuyper M, Maebe J, Van Holen R, Vandenberghe S. Artificial intelligence with deep learning in nuclear medicine and radiology. *EJNMMI Phys.* 2021;8 doi: 10.1186/s40658-021-00426-y.

90. Vieira S, Pinaya WHL, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neurosci. Biobehav. Rev.* 2017;74:58–75 doi: 10.1016/j.neubiorev.2017.01.002.

91. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Z. Med. Phys.* 2019;29:102–127 doi: 10.1016/j.zemedi.2018.11.002.

92. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc. IEEE* 1998;86:2278–2323 doi: 10.1109/5.726791.

93. Alzubaidi L, Zhang J, Humaidi AJ, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Springer International Publishing; 2021. doi: 10.1186/s40537-021-00444-8.

94. Sarvamangala DR, Kulkarni R V. Convolutional neural networks in medical image understanding: a survey. *Evol. Intell.* 2022;15:1–22 doi: 10.1007/s12065-020-00540-3.

95. Anwar SM, Adnan M, Muhammad Q, Majdi A, Khan MK. Medical image analysis using convolutional neural networks a review. *J. Med. Syst.* 2018;42:1–13.

96. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 2017;2017-Decem:5999–6009.

97. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in Vision: A

-
- Survey. *ACM Comput. Surv.* 2022;54:1–30 doi: 10.1145/3505244.
98. Hassani A, Walton S, Li J, Li S, Shi H. Neighborhood attention transformer. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 2023:6185–6194.
99. Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proc. IEEE Int. Conf. Comput. Vis.* 2021:9992–10002 doi: 10.1109/ICCV48922.2021.00986.
100. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv Prepr. arXiv2010.11929* 2020.
101. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2022;2022-June:11966–11976 doi: 10.1109/CVPR52688.2022.01167.
102. Barbieri S, Gurney-Champion OJ, Klaassen R, Thoeny HC. Deep learning how to fit an intravoxel incoherent motion model to diffusion-weighted MRI. *Magn. Reson. Med.* 2020;83:312–321 doi: 10.1002/mrm.27910.
103. Choi D, Shallue CJ, Nado Z, Lee J, Maddison CJ, Dahl GE. On Empirical Comparisons of Optimizers for Deep Learning. *arXiv Prepr. arXiv1910.05446* 2019.
104. Kingma DP, Ba JL. Adam: A method for stochastic optimization. *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* 2015:1–15.
105. Smith SL, Kindermans PJ, Ying C, Le Q V. Don't decay the learning rate, increase the batch size. *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.* 2018:1–11.
106. Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.* 2016:1–14.
107. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd Int. Conf. Mach. Learn. ICML 2015* 2015;1:448–456.
108. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 2014;15:1929–1958.
109. Spinner GR, Federau C, Kozerke S. Bayesian inference using hierarchical and spatial priors for intravoxel incoherent motion MR imaging in the brain: Analysis of cancer and acute stroke. *Med. Image Anal.* 2021;73:102144 doi: <https://dx.doi.org/10.1016/j.media.2021.102144>.
110. Gyori NG, Palombo M, Clark CA, Zhang H, Alexander DC. Training data distribution significantly impacts the estimation of tissue microstructure with machine learning. *Magn. Reson. Med.* 2022;87:932–947 doi: 10.1002/mrm.29014.
111. Karimi D, Gholipour A. Diffusion tensor estimation with transformer neural networks. *Artif. Intell. Med.* 2022;130:102330 doi: 10.1016/j.artmed.2022.102330.
112. Scalco E, Rizzo G, Mastropietro A. The quantification of IntraVoxel incoherent motion – MRI maps cannot preserve texture information: An evaluation based on simulated and in-vivo images. *Comput. Biol. Med.* 2023;154 doi: 10.1016/j.combiomed.2022.106495.
113. Lim JP, Blumberg SB, Narayan N, et al. Fitting a directional microstructure model to diffusion-relaxation mri data with self-supervised machine learning. In: *International Workshop on Computational Diffusion MRI*. Springer Nature Switzerland; 2022. pp. 77–88.

-
114. Huang H, Liu B, Xu Y, Zhou W. Synthetic-to-real domain adaptation with deep learning for fitting the intravoxel incoherent motion model of diffusion-weighted imaging. *Med. Phys.* 2023;50:1614–1622 doi: 10.1002/mp.16031.
115. Huang HM. An unsupervised convolutional neural network method for estimation of intravoxel incoherent motion parameters. *Phys. Med. Biol.* 2022;67 doi: 10.1088/1361-6560/ac9a1f.
116. Wu J, Kang T, Lan X, et al. IMPULSED model based cytological feature estimation with U-Net: Application to human brain tumor at 3T. *Magn. Reson. Med.* 2023;89:411–422 doi: 10.1002/mrm.29429.
117. Vasylechko SD, Warfield SK, Afacan O, Kurugol S. Self-supervised IVIM DWI parameter estimation with a physics based forward model. *Magn. Reson. Med.* 2021 doi: 10.1002/mrm.28989.
118. Voortter PHM, Backes WH, Gurney-Champion OJ, et al. Improving microstructural integrity, interstitial fluid, and blood microcirculation images from multi-b-value diffusion MRI using physics-informed neural networks in cerebrovascular disease. *Magn. Reson. Med.* 2023;1657–1671 doi: 10.1002/mrm.29753.
119. Troelstra MA, Van Dijk AM, Witjes JJ, et al. Self-supervised neural network improves tri-exponential intravoxel incoherent motion model fitting compared to least-squares fitting in non-alcoholic fatty liver disease. *Front. Physiol.* 2022;13:1–15 doi: 10.3389/fphys.2022.942495.
120. Zhou XX, Wang XY, Liu EH, et al. An Unsupervised Deep Learning Approach for Dynamic-Exponential Intravoxel Incoherent Motion MRI Modeling and Parameter Estimation in the Liver. *J. Magn. Reson. Imaging* 2022;56:848–859 doi: 10.1002/jmri.28074.
121. Ottens T, Barbieri S, Orton MR, et al. Deep learning DCE-MRI parameter estimation: Application in pancreatic cancer. *Med. Image Anal.* 2022;80:102512 doi: 10.1016/j.media.2022.102512.
122. Gyori NG, Palombo M, Clark CA, Zhang H, Alexander DC. Training data distribution significantly impacts the estimation of tissue microstructure with machine learning. *Magn. Reson. Med.* 2021 doi: 10.1002/mrm.29014.
123. de Almeida Martins JP, Nilsson M, Lampinen B, et al. Neural networks for parameter estimation in microstructural MRI: Application to a diffusion-relaxation model of white matter. *Neuroimage* 2021;244 doi: 10.1016/j.neuroimage.2021.118601.
124. Golkov V, Dosovitskiy A, Sperl JJ, et al. q-Space Deep Learning: Twelve-Fold Shorter and Model-Free Diffusion MRI Scans. *IEEE Trans. Med. Imaging* 2016;35:1344–1351 doi: 10.1109/TMI.2016.2551324.
125. Nedjati-Gilani GL, Schneider T, Hall MG, et al. Machine learning based compartment models with permeability for white matter microstructure imaging. *Neuroimage* 2017;150:119–135 doi: 10.1016/j.neuroimage.2017.02.013.
126. Li Z, Gong T, Lin Z, et al. Fast and robust diffusion kurtosis parametric mapping using a three-dimensional convolutional neural network. *IEEE Access* 2019;7:71398–71411 doi: 10.1109/ACCESS.2019.2919241.
127. Bertleff M, Domsch S, Weingartner S, et al. Diffusion parameter mapping with the combined intravoxel incoherent motion and kurtosis model using artificial neural networks at 3 T. *NMR Biomed.* 2017;30 doi: <https://dx.doi.org/10.1002/nbm.3833>.

-
128. Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. *J. Imaging* 2020;6:1–18 doi: 10.3390/JIMAGING6060052.
 129. Nicholson Price W. Big data and black-box medical algorithms. *Sci. Transl. Med.* 2018;10 doi: 10.1126/scitranslmed.aao5333.
 130. Abdar M, Pourpanah F, Hussain S, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion* 2021;76:243–297 doi: 10.1016/j.inffus.2021.05.008.
 131. Gawlikowski J, Tassi CRN, Ali M, et al. A survey of uncertainty in deep neural networks. Springer Netherlands; 2023. doi: 10.1007/s10462-023-10562-9.
 132. Epstein SC, Bray TJP, Hall-Craggs M, Zhang H. Choice of training label matters: how to best use deep learning for quantitative MRI parameter estimation. 2022:1–20.
 133. Diao Y, Jelescu I. Parameter estimation for WMTI-Watson model of white matter using encoder–decoder recurrent neural network. *Magn. Reson. Med.* 2023;89:1193–1206 doi: 10.1002/mrm.29495.
 134. Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ArXiv Prepr. arXiv1610.02136* 2016.
 135. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. *Int. Conf. Mach. Learn.* 2017;3:2130–2143.
 136. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* 2017;2017-Decem:5575–5585.
 137. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *33rd Int. Conf. Mach. Learn. ICML 2016* 2016;3:1651–1660.
 138. Guerquin-Kern M, Lejeune L, Pruessmann KP, Unser M. Realistic analytical phantoms for parallel magnetic resonance imaging. *IEEE Trans. Med. Imaging* 2012;31:626–636 doi: 10.1109/TMI.2011.2174158.
 139. Didenko Vasylechko S, Warfield SK, Kurugol S, Afacan O. SynthMap: a generative model for synthesis of 3D datasets for quantitative MRI parameter mapping of myelin water fraction. *Proc. Mach. Learn. Res. Rev.* 2022:1–16.
 140. Van Reeth E, Tham IWK, Tan CH, Poh CL. Super-resolution in magnetic resonance imaging: A review. *Concepts Magn. Reson. Part A Bridg. Educ. Res.* 2012;40 A:306–325 doi: 10.1002/cmr.a.21249.
 141. Sumpf TJ, Uecker M, Boretius S, Frahm J. Model-based nonlinear inverse reconstruction for T2 mapping using highly undersampled spin-echo MRI. *J. Magn. Reson. Imaging* 2011;34:420–428 doi: 10.1002/jmri.22634.
 142. Maier O, Schoormans J, Schloegl M, et al. Rapid T1 quantification from high resolution 3D data with model-based reconstruction. *Magn. Reson. Med.* 2019;81:2072–2089 doi: 10.1002/mrm.27502.
 143. Uparkar O, Bharti J, Pateriya RK, Gupta RK, Sharma A. Vision Transformer Outperforms Deep Convolutional Neural Network-based Model in Classifying X-ray Images. *Procedia Comput. Sci.* 2022;218:2338–2349 doi: 10.1016/j.procs.2023.01.209.
 144. Gao Y, Zhou M, Metaxas DN. Utnet: A Hybrid Transformer Architecture for Medical Image Segmentation. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 2021;12903 LNCS:61–71 doi: 10.1007/978-3-030-87199-4_6.

-
145. Zheng S, Lu J, Zhao H, et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2021:6877–6886 doi: 10.1109/CVPR46437.2021.00681.
 146. Hatamizadeh A, Tang Y, Nath V, et al. UNETR: Transformers for 3D Medical Image Segmentation. *Proc. - 2022 IEEE/CVF Winter Conf. Appl. Comput. Vision, WACV 2022* 2022:1748–1758 doi: 10.1109/WACV51458.2022.00181.
 147. Ji Y, Zhang R, Wang H, et al. Multi-compound Transformer for Accurate Biomedical Image Segmentation. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 2021;12901 LNCS:326–336 doi: 10.1007/978-3-030-87193-2_31.
 148. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable Detr: Deformable Transformers for End-To-End Object Detection. *ICLR 2021 - 9th Int. Conf. Learn. Represent.* 2021:1–16.
 149. Naseer M, Ranasinghe K, Khan S, Hayat M, Khan FS, Yang MH. Intriguing Properties of Vision Transformers. *Adv. Neural Inf. Process. Syst.* 2021;28:23296–23308.
 150. Geirhos R, Narayanappa K, Mitzkus B, et al. Partial success in closing the gap between human and machine vision. *Adv. Neural Inf. Process. Syst.* 2021;29:23885–23899.
 151. Tuli S, Dasgupta I, Grant E, Griffiths TL. Are Convolutional Neural Networks or Transformers more like human vision? *Proc. 43rd Annu. Meet. Cogn. Sci. Soc. Comp. Cogn. Anim. Minds, CogSci 2021* 2021;2:1844–1850.
 152. Chen J, Lu Y, Yu Q, et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv Prepr. arXiv2102.04306* 2021.
 153. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conf. Comput. Vis. Pattern Recognit.* 2009:248–255 doi: 10.1109/cvpr.2009.5206848.
 154. Matsoukas C, Haslum JF, Söderberg M, Smith K. Is it Time to Replace CNNs with Transformers for Medical Images? *arXiv Prepr. arXiv2108.09038* 2021.
 155. Xie Y, Zhang J, Xia Y, Wu Q. UniMiSS: Universal Medical Self-supervised Learning via Breaking Dimensionality Barrier. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 2022;13681 LNCS:558–575 doi: 10.1007/978-3-031-19803-8_33.
 156. Lefebvre T, Hébert M, Bilodeau L, et al. Intravoxel incoherent motion diffusion-weighted MRI for the characterization of inflammation in chronic liver disease. *Eur. Radiol.* 2021;31:1347–1358 doi: 10.1007/s00330-020-07203-y.
 157. Zhang Y, Liu H, Hu Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 2021;12901 LNCS:14–24 doi: 10.1007/978-3-030-87193-2_2.
 158. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 2022;12962 LNCS:272–284 doi: 10.1007/978-3-031-08999-2_22.
 159. Tang Y, Yang D, Li W, et al. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2022;2022-June:20698–20708 doi: 10.1109/CVPR52688.2022.02007.
 160. Bernard O, Lalonde A, Zotti C, et al. Deep Learning Techniques for Automatic MRI

Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Trans. Med. Imaging* 2018;37:2514–2525 doi: 10.1109/TMI.2018.2837502.

161. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *JMLR Workshop and Conference Proceedings*. Vol. 9. ; 2010. pp. 249–256.

162. Koopman T, Martens R, Gurney-Champion OJ, et al. Repeatability of IVIM biomarkers from diffusion-weighted MRI in head and neck: Bayesian probability versus neural network. *Magn. Reson. Med.* 2021;85:3394–3402 doi: <https://dx.doi.org/10.1002/mrm.28671>.

163. Smith LN. Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision (WACV). ; 2017. pp. 464–472. doi: 10.1109/WACV.2017.58.

164. Kaandorp MPT, Barbieri S, Klaassen R, et al. Improved unsupervised physics-informed deep learning for intravoxel incoherent motion modeling and evaluation in pancreatic cancer patients. *Magn. Reson. Med.* 2021;86:2250–2265 doi: <https://dx.doi.org/10.1002/mrm.28852>.

165. Kaandorp MPT, Zijlstra F, Federau C, While PT. Deep learning intravoxel incoherent motion modeling: Exploring the impact of training features and learning strategies. *Magn. Reson. Med.* 2023;312–328 doi: 10.1002/mrm.29628.

166. Ianus A, Mulcahy C, Drobnjak I. Diffusion MRI; From Research to Clinic (2022 ISMRM Workshop). *Magn. Reson. Med.* 2022.

167. Kaandorp MPT, Gurney-Champion OJ. IVIMNET. GitHub code. <https://github.com/oliverchampion/IVIMNET>. Published 2021.

168. Kaandorp MPT. Explore Deep Learning IVIM. GitHub code. https://github.com/Mishakaandorp/Explore_Deep_Learning_IVIM. Published 2023.

169. Kaandorp MPT. Incorporating spatial information in deep learning parameter estimation. GitHub code. https://github.com/Mishakaandorp/Incorporating_spatial_information_deep_learning_parameter_estimation. Published 2023.

170. Varkey B. Principles of Clinical Ethics and Their Application to Practice. *Med. Princ. Pract.* 2021;30:17–28 doi: 10.1159/000509119.

171. Mittelstadt BD, Floridi L. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Sci. Eng. Ethics* 2016;22:303–341 doi: 10.1007/s11948-015-9652-2.






172. Willeminck MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology* 2020;295:4–15 doi: 10.1148/radiol.2020192224.

173. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of using and sharing clinical imaging data for artificial intelligence: A proposed framework. *Radiology* 2020;295:675–682 doi: 10.1148/radiol.2020192536.

174. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. *INC*; 2020. doi: 10.1016/B978-0-12-818438-7.00012-5.

Paper I

Improved unsupervised physics-informed deep learning for intravoxel incoherent motion modeling and evaluation in pancreatic cancer patients

Misha P. T. Kaandorp^{1,2,3}  | Sebastiano Barbieri⁴  | Remy Klaassen⁵ |
Hanneke W. M. van Laarhoven⁵  | Hans Crezee¹ | Peter T. While^{2,3}  |
Aart J. Nederveen¹ | Oliver J. Gurney-Champion¹ 

¹Department of Radiology and Nuclear Medicine, Cancer Center Amsterdam, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands

²Department of Radiology and Nuclear Medicine, St. Olav's University Hospital, Trondheim, Norway

³Department of Circulation and Medical Imaging, NTNU – Norwegian University of Science and Technology, Trondheim, Norway

⁴Centre for Big Data Research in Health, UNSW, Sydney, Australia

⁵Department of Medical Oncology, Cancer Center Amsterdam, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands

Correspondence

Misha Pieter Thijs Kaandorp, Department of Radiology and Nuclear Medicine, St. Olav's University Hospital, Postbox 3250 Torgarden, 7006 Trondheim, Norway.
Email: mpkaando@stud.ntnu.no

Funding information

KWF Kankerbestrijding, Grant/Award Number: KWF-UVA 18410 and KWF-UVA 2013.5932; Norges Forskningsråd, Grant/Award Number: FRIPRO Researcher Project 302624

Purpose: Earlier work showed that IVIM-NET_{orig}, an unsupervised physics-informed deep neural network, was faster and more accurate than other state-of-the-art intravoxel-incoherent motion (IVIM) fitting approaches to diffusion-weighted imaging (DWI). This study presents a substantially improved version, IVIM-NET_{optim}, and characterizes its superior performance in pancreatic cancer patients.

Method: In simulations (signal-to-noise ratio [SNR] = 20), the accuracy, independence, and consistency of IVIM-NET were evaluated for combinations of hyperparameters (fit S_0 , constraints, network architecture, number of hidden layers, dropout, batch normalization, learning rate), by calculating the normalized root-mean-square error (NRMSE), Spearman's ρ , and the coefficient of variation (CV_{NET}), respectively. The best performing network, IVIM-NET_{optim} was compared to least squares (LS) and a Bayesian approach at different SNRs. IVIM-NET_{optim}'s performance was evaluated in an independent dataset of 23 patients with pancreatic ductal adenocarcinoma. Fourteen of the patients received no treatment between two repeated scan sessions and nine received chemoradiotherapy between the repeated sessions. Intersession within-subject standard deviations (wSD) and treatment-induced changes were assessed.

Results: In simulations (SNR = 20), IVIM-NET_{optim} outperformed IVIM-NET_{orig} in accuracy (NRMSE(D) = 0.177 vs 0.196; NMRSE(f) = 0.220 vs 0.267; NRMSE(D^*) = 0.386 vs 0.393), independence ($\rho(D^*, f)$ = 0.22 vs 0.74), and consistency ($CV_{NET}(D)$ = 0.013 vs 0.104; $CV_{NET}(f)$ = 0.020 vs 0.054; $CV_{NET}(D^*)$ = 0.036 vs 0.110).

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Magnetic Resonance in Medicine* published by Wiley Periodicals LLC on behalf of International Society for Magnetic Resonance in Medicine

IVIM-NET_{optim} showed superior performance to the LS and Bayesian approaches at SNRs < 50. In vivo, IVIM-NET_{optim} showed significantly less noisy parameter maps with lower wSD for D and f than the alternatives. In the treated cohort, IVIM-NET_{optim} detected the most individual patients with significant parameter changes compared to day-to-day variations.

Conclusion: IVIM-NET_{optim} is recommended for accurate, informative, and consistent IVIM fitting to DWI data.

KEYWORDS

deep neural network, diffusion-weighted magnetic resonance imaging, intravoxel incoherent motion, IVIM, pancreatic cancer, unsupervised physics-informed deep learning

1 | INTRODUCTION

The intravoxel incoherent motion (IVIM) model¹ for diffusion-weighted imaging (DWI) shows great potential for estimating predictive and prognostic cancer imaging biomarkers.²⁻⁵ In the IVIM model, DWI signal is described by a bi-exponential decay, of which one component is attributed to conventional molecular diffusion and the other to the incoherent bulk motion of water molecules, typically credited to capillary blood flow. Hence, IVIM simultaneously provides information on diffusion (D ; diffusion coefficient), capillary microcirculation (D^* ; pseudo-diffusion coefficient), and the perfusion fraction (f) without the use of a contrast agent.⁶⁻⁸ However, despite IVIM's great potential,²⁻⁵ it is rarely used clinically. Two major hurdles preventing routine clinical use of IVIM are its poor image quality and the long fitting time.⁹⁻¹¹ Tackling these shortcomings will help toward wider use of IVIM.¹²

Currently, IVIM is often fitted using the conventional least squares (LS) algorithm. However, more accurate alternative approaches have been suggested.⁹ Until recently, Bayesian algorithms for IVIM fitting¹³ were most promising regarding inter-subject variability,⁹ precision, accuracy,¹⁴ and smooth parameter maps, suggesting less noise.¹⁵ Conversely, Bayesian approaches are substantially slower (9×10^{-2} s/vox¹¹) than the already slow LS approach (8×10^{-3} s/vox¹¹). Furthermore, Bayesian approaches may lead to biased perfusion estimates of the IVIM model.¹⁶

Recently, a promising alternative for IVIM fitting was introduced: estimating IVIM parameters with deep neural networks (DNNs). Initially, Bertleff et al¹⁷ introduced a supervised DNN for IVIM parameter estimation, in which the network was trained on simulated data for which the underlying parameters were known. However, the strong assumption of simulated training and test data being identically distributed could limit the network's performance in vivo, where noise behaves less ordered. We solved this shortcoming in earlier work,¹¹ where we used unsupervised physics-informed

deep neural networks (PI-DNNs).^{18,19} PI-DNNs formulate a physics-informed-loss-function that finds learned parameters through an iterative process. In this case, the PI-DNN used consistency between the predicted signal from the IVIM model and the measured signal as a loss term in the DNN. This resulted in an unsupervised PI-DNN capable of training directly on patient data with no ground truth: IVIM-NET_{orig}. We demonstrated in both simulations and volunteer analysis that IVIM-NET_{orig} is superior to the conventional LS approach and even performs (marginally) better than the Bayesian approach. Furthermore, IVIM-NET_{orig}'s fitting times were substantially lower (4×10^{-6} s/vox¹¹) than the LS and Bayesian approaches. However, that proof of principle IVIM-NET study did not explore many hyperparameters and focused on volunteer data.

In this work, we hypothesize that IVIM-NET_{orig} can be improved by exploring the architecture of the network, its training features and other hyperparameters. Hence, we characterized the performance of IVIM-NET for different hyperparameter settings by assessing the accuracy, independence, and consistency of the estimated IVIM parameters in simulated IVIM data. Finally, we compared the performance of our optimized IVIM-NET to the LS approach and a Bayesian approach in patients with pancreatic ductal adenocarcinoma (PDAC) receiving neoadjuvant chemoradiotherapy (CRT) in terms of image quality, parameter to noise ratio, test-retest reproducibility, and sensitivity to treatment effects.

2 | METHODS

2.1 | IVIM-NET

We initially implemented the original PI-DNN (IVIM-NET_{orig})¹¹ in Python 3.8 using PyTorch 0.4.1.²⁰ The input layer consisted of neurons that took the normalized DWI signal $S(b)/S(b=0)$ as input, where $S(b)$ is the measured signal at diffusion weighting b (b value). The input layer was

followed by three fully connected hidden layers. Each hidden layer had several neurons equal to the number of measurements (b values and the number of repeated measures) and each neuron, in turn, contained an exponential linear unit activation function.²¹ The output layer of the network consisted of the three IVIM parameters (D, f, D^*). To enforce the output layer to predict these IVIM parameters, two steps were taken. First, the absolute activation function was taken of the neuron's output (X) to constrain the predicted parameters, for example, to compute D :

$$D = |X[1]|. \quad (1)$$

Second, a physics-based loss function was introduced that computed the mean squared error between the measured input signal, $S(b)$, and the predicted IVIM signal $S_{net}(b)$, which was obtained by inserting the predicted output parameters into the normalized IVIM model. Hence:

$$L = \frac{1}{|B|} \sum_{b \in B} \left(\frac{S(b)}{S(b=0)} - S_{net}(b) \right)^2, \quad (2)$$

with

$$S_{net}(b) = fe^{-bD^*} + (1-f)e^{-bD}, \quad (3)$$

where B is the total number of image acquisitions.

Next, we evaluated whether seven novel hyperparameters (Table 1; Figure 1) of IVIM-NET improved fitting results. First, instead of fixing S_0 , we added S_0 as an additional output parameter, to allow the system to correct for noise in $S(b=0)$. Second, to restrict parameter values to physiologically plausible ranges, scaled sigmoid activation functions instead of absolute activation functions were used to constrain the predicted parameters (Table 1), for example, to compute D :

$$D = D_{min} + \text{sigmoid}(X[1]) * (D_{max} - D_{min}), \quad (4)$$

where D_{min} and D_{max} are the fit boundaries. Bound intervals of the sigmoid activation functions were chosen 60%

wider (30% at each side) than the fit boundaries of the LS and Bayesian approaches (specified in section 2.2 below) to compensate for decreasing gradients at the asymptotes of the sigmoid function. Third, we varied the number of hidden layers between one and nine. Fourth, we used dropout regularization²² in all hidden layers except for the last one. Dropout randomly removes a set percentage of network-weights each iteration during training. Fifth, we used batch normalization,²³ which normalizes the input by re-centering and re-scaling, and, consequently, preserves the representation ability of the network. Sixth, to reduce unwanted correlation between estimated parameter values, we implemented an alternative network architecture in which parameter values were predicted, in parallel, by independent sub-networks (Table 1; Figure 1). Furthermore, we evaluated different learning rates (LR) of the Adam optimizer,²⁴ ranging from 1×10^{-5} to 3×10^{-2} , and with constant $\beta = (0.9, 0.999)$.

In traditional deep learning, training and evaluation are done on separate datasets, but as this is an unsupervised DNN approach, training was done on the same data as evaluation.^{11,25} So, for simulations, these were simulated data, and in vivo, these were in vivo data. 90% of the data were used for training, and 10% of the data were used for validation. Early stopping occurred when the validation loss did not improve over 10 consecutive training epochs. Given the large amount of training data and the limited number of network parameters, each epoch consisted of only 500 random batches. So, effectively the network saw 500×128 IVIM curves in between validations.

2.2 | Simulations: characterization and optimization

100,000 IVIM curves were simulated to investigate the effects of different hyperparameters on the accuracy, independence, and consistency of the estimated IVIM parameters. DWI signals were simulated based on Equation 3 with $S_0 = 1$, 11 b values ($b = 0, 5, 10, 20, 30, 40, 60, 150, 300, 500$, and 700 s/mm^2), and pseudorandom uniformly sampled values

TABLE 1 Hyperparameter settings for training IVIM-NET, including the settings for IVIM-NET_{orig} and IVIM-NET_{optim}

Hyperparameter	Values	IVIM-NET _{orig}	IVIM-NET _{optim}
Fit S_0	True, False	False	True
Constraints	Sigmoid, Absolute	Absolute	Sigmoid
Parallel networks	True, False	False	True
Number of hidden layers	1, 2, 3, 4, 5, 6, 7, 8, 9	3	2
Dropout regularization	0%, 10%, 20%, 30%	0%	10%
Batch normalization	True, False	False	True
Learning rate	$1 \times 10^{-5}, 3 \times 10^{-5}, 1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}, 1 \times 10^{-2}, 3 \times 10^{-2}$	1×10^{-3}	3×10^{-5}

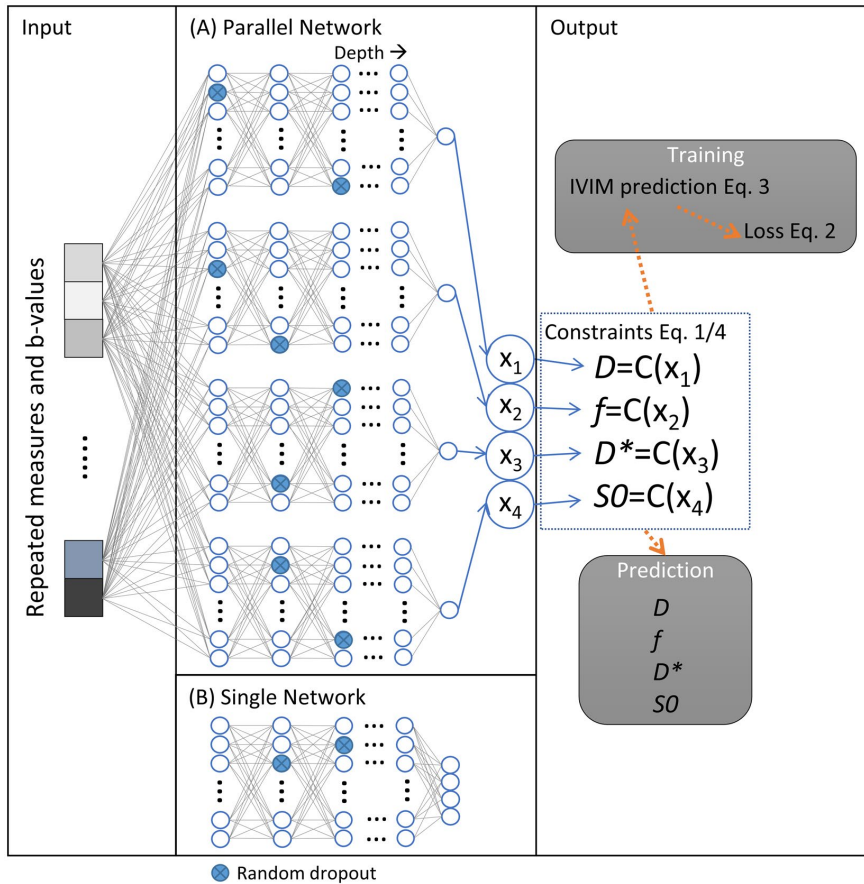


FIGURE 1 Representation of the PI-DNN with different hyperparameter options (Table 1). In this example, the input signal, consisting of the measured DWI signal, is fedforward either through a parallel network design where each parameter is predicted by a separate fully connected set of hidden layers (A) or the original single fully connected network design (B). The blue circles indicate an example of randomly selected neurons for dropout. In this example, the output layer consists of four neurons with either absolute (Equation 1) or sigmoid activation functions (Equation 4) whose values correspond to the IVIM parameters. Subsequently, the network predicts the IVIM signal (Equation 3), which is used to compute the loss function (Equation 2). With the loss function, the network trains the PI-DNN to give good estimates of the IVIM parameters

of D : 0.5×10^{-3} to 3×10^{-3} mm^2/s , f : 5 to 55%, and D^* : 10×10^{-3} to 100×10^{-3} mm^2/s . These ranges were slightly broader than the typical values found in abdominal IVIM.²⁶ Random Rician noise in the form of complex Gaussian noise was added to the curves with predefined signal-to-noise ratio (SNR) levels (constant noise amplitude over b values; SNR defined at $b = 0$ as $S(b = 0)/\sigma$, with σ the SD of the Gaussians).²⁷

Accuracy was assessed as the normalized root-mean-square error (NRMSE) between the ground truth parameter values and the estimated IVIM parameters.

Independence of the parameter estimates was assessed by the Spearman rank correlation coefficients (ρ) between all parameter pairs. As the simulated data were independent

and random, a ρ should be 0. The absolute value of ρ was taken, as both positive and negative deviations from zero are equally undesirable. Some networks always returned the same value for D^* , independent of the input data (Supporting Information Figure S1, which is available online). For such cases, ρ is technically undefined. As these cases are undesirable ρ was set to 1.

As training a DNN is a stochastic process, training on the same dataset results in different final network-weights, and consequently, different predictions on the same data. To assess the consistency of estimated parameter values, each network variant was trained 50 times on identical data, where each repeat had a new random initialization, dropout, and batch selection. The normalized coefficient of variation per

parameter over the repeated trainings (CV_{NET}) was taken as a measure of the consistency:

$$CV_{NET} = \frac{1}{\bar{x}_{true}} \sqrt{\frac{1}{n \times m - 1} \sum_{i=1}^n \sum_{j=1}^m x_{i,j} - \bar{x}_i^2}, \quad (5)$$

where \bar{x}_{true} is the mean simulated IVIM parameter value, n the number of simulated curves, m the number of repeated trainings, $x_{i,j}$ is the j^{th} repeated prediction of the i^{th} simulated decay curve, \bar{x}_i is the mean over the repeated m predictions of the i^{th} simulated signal curve. As the LS and Bayesian approaches are deterministic, their CV_{NET} was zero.

As a result of the repeated training, we obtained 50 values for the NRMSEs and ρ 's. Therefore, the median and interquartile ranges were reported.

As a baseline for comparison, we evaluated the IVIM parameters (D , f , D^*) in IVIM-NET_{orig}, the LS and Bayesian approaches. We used the Levenberg-Marquardt non-linear algorithm for the LS fit.^{28,29} For the Bayesian approach, we used the algorithm from previous work.¹¹ For both the LS and Bayesian approaches, $S0$ was included as a fit parameter. The Bayesian approach used a data-driven lognormal prior for D and D^* , and a beta distribution for f and $S0$. The prior distributions were determined empirically by fitting these distributions to the results from the LS approach on the same dataset. The maximum a posteriori probability was used as an estimate of the IVIM parameters. The LS and Bayesian approaches were performed with fit boundaries of D : 0×10^{-3} to 5×10^{-3} mm²/s, f : 0 to 70%, D^* : 5×10^{-3} to 300×10^{-3} mm²/s, and $S0$: 0.7 to 1.3.

After baseline characterization, IVIM-NET was optimized by testing various combinations of the hyperparameters (Table 1; Figure 1). Previous studies reported reliable SNR values of IVIM in the abdomen between 10 and 40.³⁰⁻³³ So, to simulate reliable abdominal IVIM signals, an SNR of 20 was chosen for hyperparameter evaluation. We trained the network on the simulated signals using every combination of the following options: fit $S0$ parameters, absolute or sigmoid constraints, parallel network, dropout, and batch normalization - while fixing the number of hidden layers to three (used in IVIM-NET_{orig}, Table 1) and the LR to 1×10^{-4} . In an exploratory phase, we found that reducing the LR from 1×10^{-3} (IVIM-NET_{orig}) to 1×10^{-4} was essential for obtaining networks with improvements in accuracy, independence, and consistency. Each network (ie, combination of hyperparameters) received a ranking in each of the nine performance measures (three metrics for three parameters), and these nine ranks were summed. Hence, the best possible summed rank is 9. The best performing network was then chosen by selecting the network with the lowest summed rank.

With the best options for the fit $S0$ parameters, constraints, parallel network, dropout, and batch normalization,

we tested the performance of the network as a function of the LR and the number of hidden layers (Table 1). From those results, we finally selected the best performing optimized network by again selecting the lowest summed rank: IVIM-NET_{optim}. IVIM-NET_{optim}'s performance was then characterized and compared to the LS approach, Bayesian approach and IVIM-NET_{orig} for SNR values between 8 (low) and 100 (high).

2.3 | Verification in patients with PDAC

We used two IVIM datasets of patients with PDAC to validate IVIM-NET_{optim}'s performance in vivo: one dataset to assess test-retest reproducibility, and one to test whether we can detect treatment effects. Both studies were approved by our local medical ethics committee, and all patients gave written informed consent.

Both datasets (NCT01995240; NCT01989000) were published earlier.^{9,34,35} The first dataset consists of 14 patients with locally advanced or metastatic PDAC who underwent IVIM in two separate imaging sessions (average 4.5 days apart, range: 1-8 days) with no treatment in-between. The second dataset consisted of nine PDAC patients with (borderline) resectable PDAC who received CRT as part of the PREOPANC study³⁶ where patients were scanned before and after CRT.

MRI data were acquired using a 3T MRI scanner (Ingenia, Philips, Best, The Netherlands). A respiratory triggered (navigator on liver dome) 2D multi-slice diffusion-weighted echo-planar imaging was used with parameters: repetition time (TR) > 2200 ms (depending on respiration speed), echo time (TE) = 45 ms, flip angle (FA) = 90 deg, field of view (FOV) = 432×108 mm², acquisition matrix = 144×34 , 18 slices, slice thickness = 3.7 mm and 12 b values (directions): 0 (15), 10 (9), 20 (9), 30 (9), 40 (9), 50 (9), 75 (4), 100 (12), 150 (4), 250 (4), 400 (4), and 600 (16) mm²/s. Fat suppression was carried out with a gradient reversal during slice selection and spectral presaturation with inversion recovery. Diffusion gradient times were 10.1 ms with a delay between diffusion gradients onset of 22.6 ms.

DWI images were co-registered to a reference volume consisting of a mean DWI image over all b values using deformable image registration in Elastix.³⁷ A radiologist (10 y of experience in abdominal radiology) and researcher (4 y of experience in contouring pancreatic cancer) drew a region of interest (ROI) in the tumor in consensus. IVIM parameter maps of D , f and D^* were derived using the LS approach, Bayesian approach, and IVIM-NET_{optim}. Background voxels were removed automatically before fitting by removing voxels with $S(b=0) < 0.5 \times \text{median}(S(b=0))$. Fitting was done without averaging over the diffusion directions. IVIM-NET_{optim} was trained on all combined patient data. Values

under 0 for D , f , and D^* were considered not physiologic and set to 0, and for further statistics, values of D^* were masked where $f < 5\%$ as D^* only “exists” in perfused voxels. All computations were carried out on a single core of a conventional desktop computer (CPU: Intel Core i7-8700 CPU at 3.20 GHz). The average fitting time of each algorithm was recorded.

First, the parameter maps were compared qualitatively in terms of feature clarity, and by visually assessing consistency of fit to the IVIM signal in pairs of neighboring voxels. For a quantitative comparison, the parameter to noise ratio (PNR) of the parameter maps was estimated in a homogeneous 2D ROI (>20 voxels) in the liver. PNR was defined as $mean/STD$ of the homogenous ROIs and was calculated for each scan separately. We tested whether the difference in PNR between fit approaches was significant using paired t -tests.

To determine clinical usefulness of IVIM-NET, we investigated whether we could detect changes in parameter values throughout CRT by comparing patients receiving treatment to the baseline test-retest repeatability. This analysis was performed with the median parameter values from within the ROIs. To evaluate test-retest repeatability, intersession within-subject standard deviation (wSD)³⁸ was calculated for each IVIM parameter using the data from the patients with repeated baseline scans. Bland-Altman plots were plotted for patients from both cohorts. We calculated the 95% confidence intervals (CIs) from the patients with repeated scans at baseline (assuming zero offsets). In the treated cohort, we used a paired t -test to test whether parameters had significantly changed due to treatment within the cohort (significance level $\alpha = 0.05$). Furthermore, patients from the treatment cohort were added to the Bland-Altman plots and individual patients who had changes exceeding the 95% CIs were considered to have significant changes in tumor microstructure.³⁹

3 | RESULTS

3.1 | Simulations: characterization and optimization

The original network, IVIM-NET_{orig}, showed substantially lower NRMSE for all estimated parameters than the LS and Bayesian approaches. However, IVIM-NET_{orig} had strong correlations between D^* and f (high $\rho(D^*, f)$; Table 2 and Figure 2D), and had considerable CV_{NET}.

The NRMSE, ρ , and CV_{NET} for all hyperparameter combinations are shown in the Supporting Information Figures S2-S9. The summarizing sum of ranks (Supporting Information Figures S5 and S9) allowed us to determine IVIM-NET_{optim} (Table 1). IVIM-NET_{optim} resolved the high $\rho(D^*, f)$ found in IVIM-NET_{orig} (Table 2; Figure 2D,E) and substantially reduced the NRMSE and CV_{NET}. Single changes away from IVIM-NET_{optim} can lead to marginally better NRMSE, lower ρ or lower CV_{NET} (Figure 3), but only at a cost to the other two attributes. It is clear that the reduced $\rho(D^*, f)$ cannot be attributed to a single parameter, but was a result of the combination of sigmoid constraints and batch normalization (Supporting Information Figure S3). Adding dropout (10%), fitting $S0$ and using our parallel network design decreased the NRMSE, while still having a low $\rho(D^*, f)$ (Table 2). Increasing dropout in IVIM-NET_{optim} or using a single network architecture resulted in similar NRMSE, however, increased $\rho(D^*, f)$ (Figure 3; Supporting Information Figures S2-S4). Generally, increasing the number of hidden layers resulted in a marginally higher ρ , and lower NRMSE and CV_{NET}. A too high/low LR (Supporting Information Figures S5-S7) caused higher NRMSEs and less consistency.

IVIM-NET_{optim} was superior to the LS and Bayesian approaches for SNRs 8-33. Compared to IVIM-NET_{orig}, IVIM-NET_{optim} was associated with improved NRMSE for f and D at all SNRs (Figure 4). For D^* , the networks performed

	Least squares	Bayesian	IVIM-NET _{orig}	IVIM-NET _{optim}
NRMSE D [fraction]	0.279	0.233	0.196 (0.190-0.214)	0.177 (0.176-0.178)
NRMSE f [fraction]	0.387	0.281	0.267 (0.259-0.273)	0.220 (0.218-0.222)
NRMSE D^* [fraction]	0.805	0.575	0.393 (0.382-0.414)	0.386 (0.381-0.390)
$\rho(D, D^*)$	0.24	0.08	0.23 (0.17-0.28)	0.20 (0.19-0.21)
$\rho(D, f)$	0.18	0.03	0.04 (0.02-0.09)	0.01 (0.00-0.01)
$\rho(D^*, f)$	0.20	0.13	0.74 (0.64-0.80)	0.22 (0.23-0.2)
CV _{NET} D [fraction]	0	0	0.104	0.013
CV _{NET} f [fraction]	0	0	0.054	0.020
CV _{NET} D^* [fraction]	0	0	0.110	0.036

TABLE 2 NRMSE, ρ , and CV_{NET} of the LS approach, Bayesian approach, IVIM-NET_{orig}, and IVIM-NET_{optim} for the estimated parameters IVIM (D , f , D^*) in simulations at SNR 20 for 50 repeated trainings^a

^aValues of IVIM-NET: median (interquartile range).

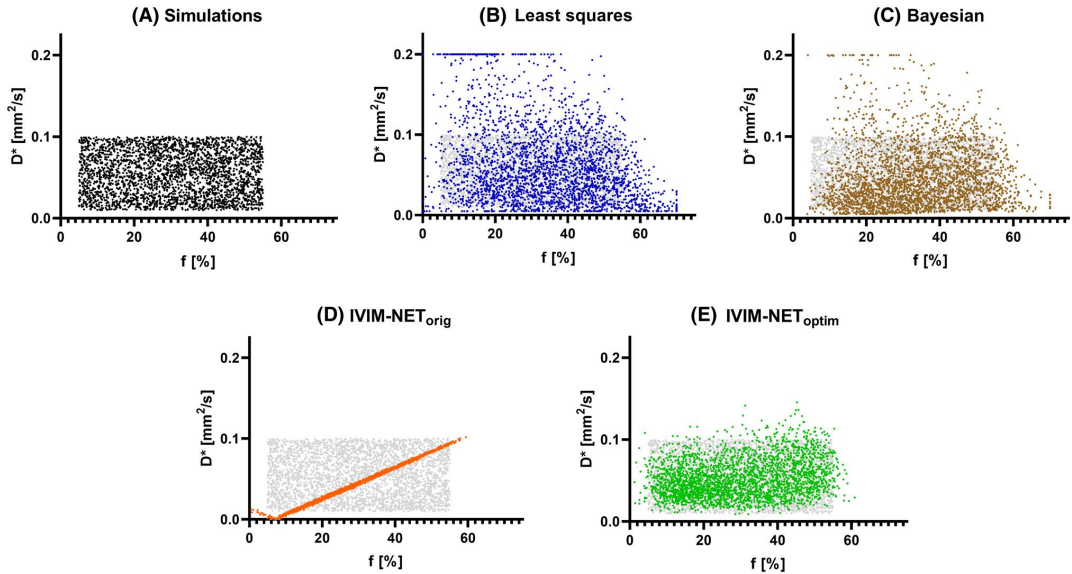


FIGURE 2 D^* plotted against f for simulations (A), LS (B), Bayesian (C), IVIM-NET_{orig} (D), and IVIM-NET_{optim} (E). In all plots, the values of the simulations are presented in gray. The apparent patterns in the LS approach (many predictions at $D^* = 0.2 \text{ mm}^2/\text{s}$) and IVIM-NET_{orig} (the line flips at $D^* = 0 \text{ mm}^2/\text{s}$) are a result of the fit constraints

similarly regarding NRMSE, with IVIM-NET_{optim} performing slightly better at SNRs > 20 and IVIM-NET_{orig} for SNRs < 20. IVIM-NET_{optim} had lower $\rho(D^*, f)$ than IVIM-NET_{orig} and improved CV_{NET} for all SNR levels.

3.2 | Verification in patients with PDAC

Examples of parameter maps computed with the LS approach, Bayesian approach, and IVIM-NET_{optim} together with two individual voxel fits of PDAC patients from both cohorts are presented in Figures 5 and 6. Additional parameter maps of 10 other PDAC patients are shown in the Supporting Information Figures S11-S20. Qualitatively comparing the parameter maps shows that IVIM-NET_{optim} has very similar voxel values as the LS and Bayesian approaches for most voxels. However, where the LS and Bayesian approaches sometimes show “noisy” voxels (ie, different from their neighbors) with substantially higher f (order of 1) and lower D (some voxels as low as $D = 0 \text{ mm}^2/\text{s}$; Supporting Information Figures S11, S12) and D^* (oftentimes to the lower bound of D^* ; Supporting Information Figures S12, S14-S16, S19), IVIM-NET_{optim} often sticks to sensible D , f , and D^* that are similar to the neighboring voxels resulting in more homogenous parameter maps. Note that IVIM-NET is fitted at a per-voxel level and is unaware of the voxel location. Quantitatively

evaluating the parameter maps shows that IVIM-NET_{optim} had significantly better PNR than both the LS and Bayesian approaches for D and D^* (Table 3) and significantly better PNR than the LS approach for f .

In the test-retest cohort, IVIM-NET_{optim} showed the lowest wSD for D and f (Table 3), while the Bayesian approach had the lowest wSD for D^* . When averaging IVIM parameters for the repeated patient scans, IVIM-NET_{optim} computed a higher D , lower f , and higher D^* than the LS and Bayesian approaches in the tumor (Table 3). The repeated scans are visualized as black x’s in the Bland-Altman plots, together with their 95% CIs in Figure 7.

When considering the CRT patients as a whole, IVIM-NET_{optim} found a significant increase in mean D and f after treatment, whereas the LS approach found only a significant increase in D after treatment in the tumor (Table 3). The Bayesian approach found no significant change in IVIM parameters.

Figure 7 shows the individual change in IVIM parameter values of patients receiving CRT compared to the 95% CIs of the test-retest cohort. With 10 significant changes, IVIM-NET_{optim} detected the most patients with significant parameter changes after CRT, with 4 individual patients with increased D , 3 patients with increased f , and 3 patients with changes in D^* . In comparison, the LS and Bayesian approaches detected only two and three significant parameter changes, respectively.

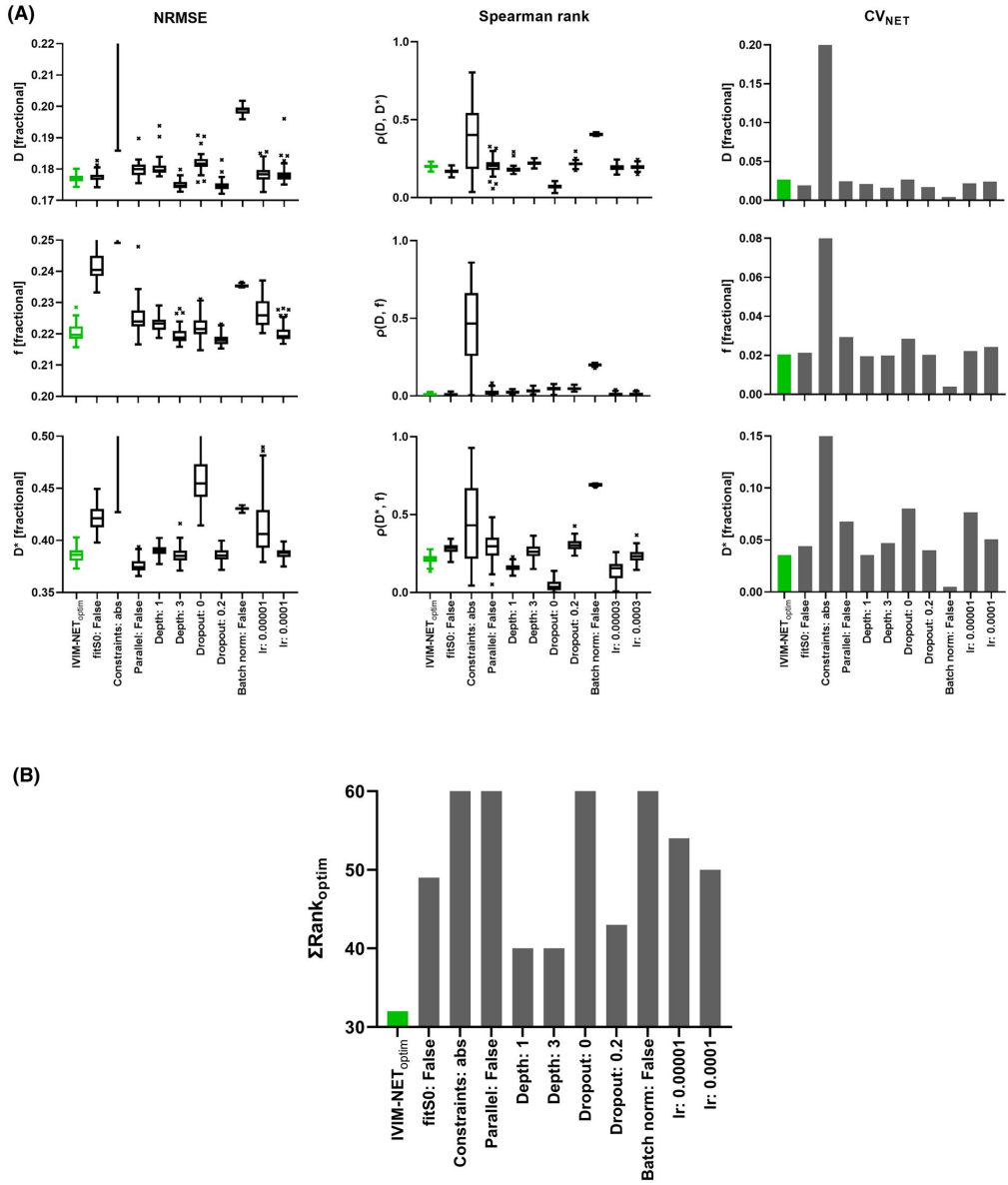


FIGURE 3 A, NRMSE (left), ρ (center), and CV_{NET} (right) plots of the estimated IVIM parameters (D , f , and D^*) with a single parameter change for IVIM-NET_{optim} (green) at SNR 20 for 50 repeated trainings. B, The ranked plot of IVIM-NET_{optim}

The average interference fitting time of IVIM-NET_{optim} after training was 3.0×10^{-5} s/vox, whereas the average fitting times of the LS and Bayesian approaches were 4.2×10^{-3} s and 1.0×10^{-1} s/vox, respectively. The median training time for IVIM-NET_{optim} (20 repeats) was 572 s with a range of 401 to 685 s, which for our dataset resulted in 2.9×10^{-4} s/vox training.

4 | DISCUSSION

This study is the first to show the potential clinical benefit of DNNs for IVIM fitting to DWI data in a patient cohort. We successfully developed and trained IVIM-NET_{optim}, an unsupervised PI-DNN IVIM fitting approach to DWI data that predicts accurate, independent, and consistent IVIM

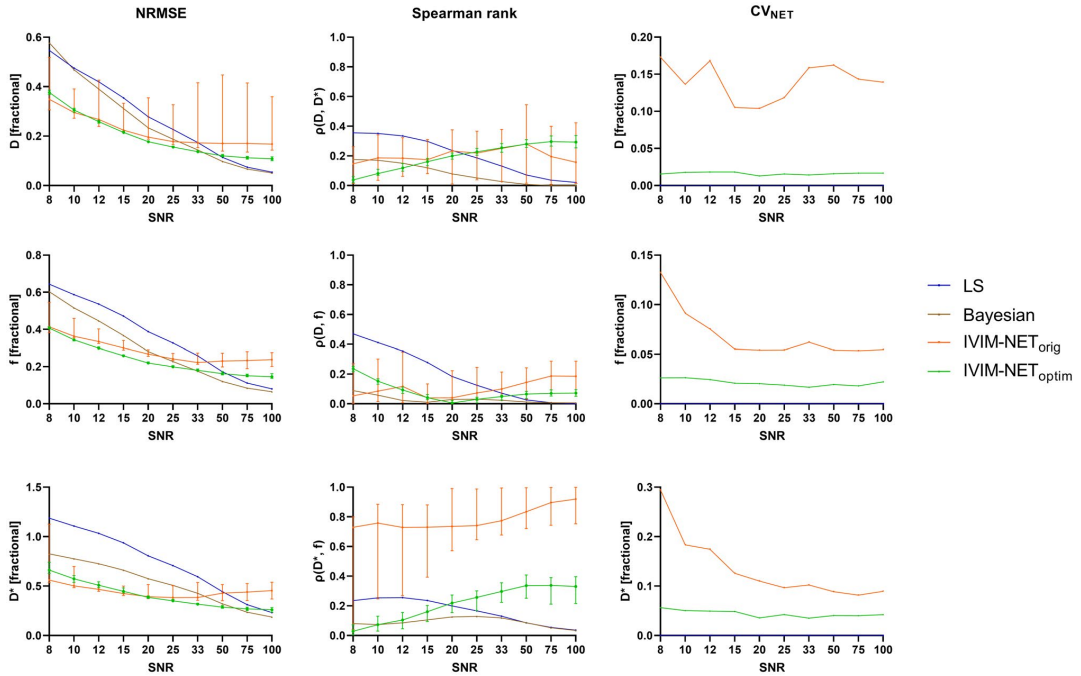


FIGURE 4 NRMSE (left), ρ (center), and CV_{NET} (right) plots of the estimated IVIM parameters (D , f , and D^*) vs SNR for the LS approach (blue), Bayesian approach (brown), $IVIM-NET_{orig}$ (orange), and $IVIM-NET_{optim}$ (green) approaches to IVIM fitting. The 5 to 95 percentiles of $IVIM-NET$ for 50 repeated trainings are plotted as error bars and show that $IVIM-NET_{orig}$ is highly inconsistent in producing IVIM parameters for multiple repeated trainings at all SNRs. $IVIM-NET_{optim}$ outperforms $IVIM-NET_{orig}$ for all SNRs. As the LS and Bayesian approaches are deterministic, their CV_{NET} was zero and not plotted. The LS and Bayesian approaches are superior at high SNRs

parameters in simulations and in vivo in patients with PDAC. $IVIM-NET_{optim}$ consisted of a parallel network architecture with two hidden layers, batch normalization, dropout of 10%, sigmoid constraints, and fitted S_0 . Optimized training was performed using an Adam optimizer with an LR of 3×10^{-5} . In simulations, $IVIM-NET_{optim}$ outperformed the original version, $IVIM-NET_{orig}$, by offering more accurate estimates of D , f and D^* , with substantially less correlation between the estimated parameters D^* and f , and more consistent parameter prediction. Furthermore, simulations demonstrated that $IVIM-NET_{optim}$ had substantially better accuracy than the conventional LS and state-of-the-art Bayesian approaches. Finally, in patients with PDAC, $IVIM-NET_{optim}$ also outperformed the alternatives. $IVIM-NET_{optim}$ showed the most detailed and significantly less noisy parameter maps, and a significant change in D and f for the whole cohort receiving CRT. Furthermore, $IVIM-NET_{optim}$ was associated with the best test-retest repeatability (smallest wSD) for D and f , which allowed it to detect the most patients with significant changes in all IVIM parameters after CRT.

$IVIM-NET_{optim}$ detected a significant positive trend in D and f for the whole cohort of patients receiving CRT, whereas the LS approach only found a significant positive

trend in D . Also, $IVIM-NET_{optim}$ detected four patients with a significant parameter increase for D , whereas the LS approach detected only one patient. These findings strongly suggest $IVIM-NET_{optim}$ as a good alternative for IVIM fitting in PDAC patients. Findings from other studies support this increase in D^{40} and f^{41} during CRT in PDAC patients. In general, PDACs tend to have lower diffusion due to the impeded water movement of compressing cells.⁴² Furthermore, PDACs are typically hypoperfused due to solid stress generated by the dense stroma and significant tumor sclerosis creating elevated interstitial pressure, which compresses tumor feeding vessels.^{41,43,44} An increase in diffusion is likely a consequence of a reduction in cell density due to necrosis as a result of treatment.^{45,46} An increase in perfusion can be explained two-fold: an effective treatment could reduce the amount of stroma and associated solid stress^{47,48}; furthermore, cell necrosis inside the tumor can reduce interstitial pressure.⁴⁷ Not all patients demonstrated a significant change induced by treatment. Therefore, using IVIM to discriminate between individual treatment effects may be feasible in the future. As treatment of these patients was part of induction therapy and patients received surgery directly after, overall survival cannot be attributed purely to CRT effects. Hence,

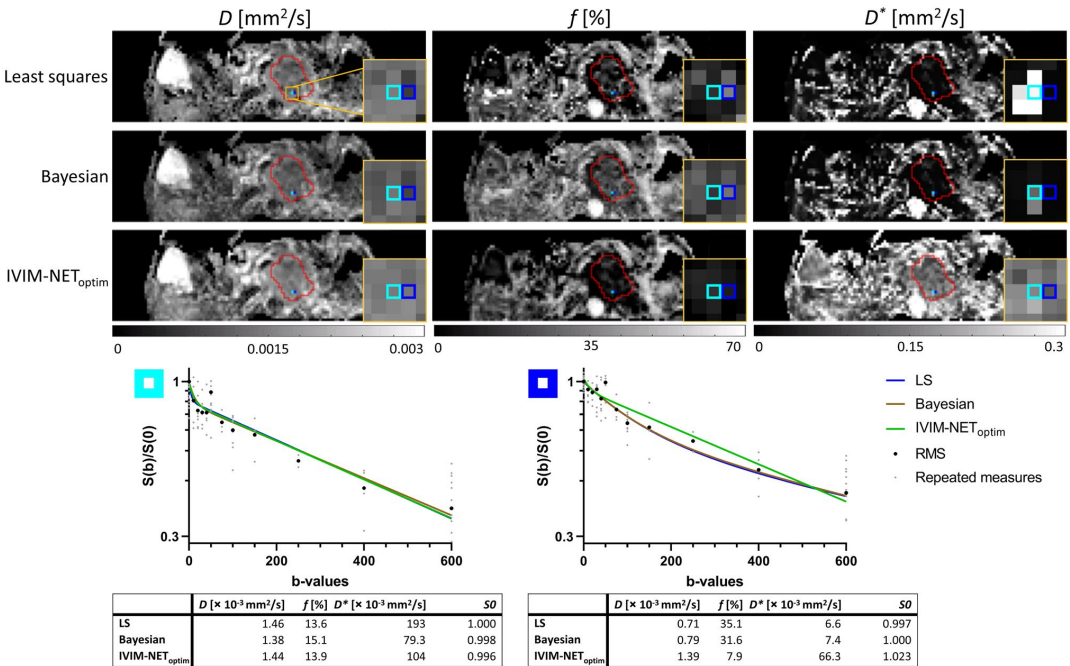


FIGURE 5 IVIM parameter maps (D , f , D^*) of the LS approach, Bayesian approach, and IVIM-NET_{optim} of a PDAC patient of the test-retest cohort. The red ROI represents the PDAC, the two highlighted blue regions correlate to the voxels from the log-plots below. The yellow square zooms in on the two highlighted voxels in the tumor. In the plots, the small light gray dots are the repeated measures, and the big black dots are the root-mean-squares of these repeated measures. The plot parameters are shown below. The light blue voxel (left plot) shows consistency in IVIM parameters for all three fitting approaches. Although the data are similar in the neighboring dark blue voxel (right plot) with a lower IVIM effect, the LS and Bayesian approaches compute a higher f , lower D , and very low D^* compared to their parameters in the light blue voxel. IVIM-NET_{optim} shows more consistency in IVIM parameters between the two neighboring voxels with a lower f for the dark blue voxel. In the parametric maps computed by IVIM-NET_{optim}, the tissues appear more homogeneous, whereas the LS approach shows noisier parameter maps, particularly around the tumor region

given the limited number of patients and the diluted treatment effect, we did not compare overall survival between patients that showed potential treatment effects and others.

Our previous work³⁵ showed that the LS approach for IVIM fitting was sensitive to individual treatment effects. However, the high wSD limited the study to detect individual treatment effects. Furthermore, that work³⁵ used denoised DWI b-images that substantially degraded image sharpness and tumor boundaries were harder to detect (eg, compare figures from this work to example figures from earlier work⁹). Conversely, our present study demonstrates that DNNs can estimate parameter maps directly from the noisy data resulting in sharp high-quality IVIM parameter maps.

For most voxels, IVIM-NET_{optim} produces very similar estimates to the LS and Bayesian approaches. However, within the tumor ROI, IVIM-NET_{optim} shows consistently different mean baseline parameters than the LS and Bayesian approaches (Table 3). We believe that there are two major contributors to this discrepancy in mean values. (1) The LS and Bayesian approaches have more noisy parameter maps

with some individual voxels showing extreme estimates. (2) IVIM-NET_{optim} is seemingly better at estimating parameters in regions of poor perfusion. The first observation is demonstrated by the individual voxel fits (Figures 5 and 6; Supporting Information Figures S11-S20), where the LS and Bayesian approaches occasionally compute noisy IVIM parameters with a substantially higher f , and lower D and D^* than IVIM-NET_{optim}. As the LS approach minimizes the sum of the residuals, this parameter combination could describe better the noisy data. However, inspecting neighboring voxels with, respectively, similar noisy data shows that the LS and Bayesian approaches are inconsistent in producing the same IVIM parameters, whereas IVIM-NET_{optim} is more consistent. The second note is especially interesting for PDACs, which are generally hypoperfused.^{41,43} Other studies reported an overestimation of perfusion parameters in poorly perfused tissue,⁴⁹⁻⁵¹ and indeed, the LS and Bayesian approaches show high and noisy perfusion fractions maps in the PDACs. Conversely, IVIM-NET_{optim} shows consistently low and less noisy perfusion in these regions. Another

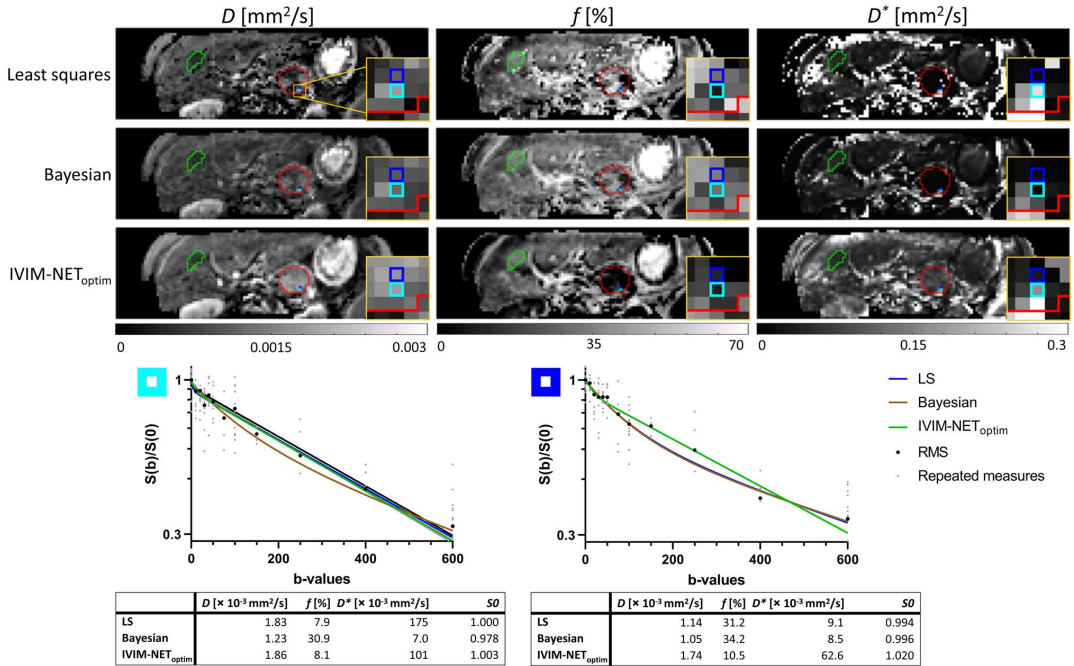


FIGURE 6 IVIM parameter maps (D, f, D^*) of the LS approach, Bayesian approach, and IVIM-NET_{optim} of a PDAC patient of the treated cohort before CRT. The red ROI represents the PDAC and the green ROI represents the 2D homogenous liver tissue ROI. The two highlighted blue regions correlate to the voxels from the log-plots below. The yellow square zooms in on the two highlighted voxels in the tumor. In the plots, the small light gray dots are the repeated measures, and the big black dots are the root-mean-squares of these repeated measures. Both plots show a low IVIM effect. The plot parameters are shown below. The light blue voxel (left plot) shows consistency in IVIM parameters between the LS approach and IVIM-NET_{optim} with low f and moderate D and D^* , while the Bayesian approach shows higher f , lower D , and very low D^* . Although the data are similar in the neighboring dark blue voxel (right plot), the LS and Bayesian approaches compute a higher f , lower D , and very low D^* compared to their parameters in the light blue voxel. IVIM-NET_{optim} shows more consistency in IVIM parameters between the two neighboring voxels with a consistent low f . In the parametric maps computed by IVIM-NET_{optim}, the tissues appear more homogeneous, particularly in the liver, the kidneys and around the tumor ROI

TABLE 3 PNR of homogenous liver tissue (top panel), wSD (middle panel), and mean IVIM parameters for the patients with treatment (bottom panel)

PNR	D		f		D^*	
LS	5.6		3.1		1.8	
Bayesian	6.3		4.0		2.3	
IVIM-NET _{optim}	8.1*		3.9		4.0*	
wSD	$D [\times 10^{-3} \text{ mm}^2/\text{s}]$		$f [\%]$		$D^* [\times 10^{-3} \text{ mm}^2/\text{s}]$	
LS	0.10		6.2		24.9	
Bayesian	0.09		4.9		5.1	
IVIM-NET _{optim}	0.06		2.4		15.8	
Mean treatment	$D [\times 10^{-3} \text{ mm}^2/\text{s}]$		$f [\%]$		$D^* [\times 10^{-3} \text{ mm}^2/\text{s}]$	
	Pre	Post	Pre	Post	Pre	Post
LS	1.35**	1.51**	13.1	16.3	52	48
Bayesian	1.18	1.30	20.4	23.4	8.6	17.7
IVIM-NET _{optim}	1.57**	1.68**	5.3**	9.1**	92	88

*Significantly ($P < .05$) better PNR compared to both of the other fitting approaches, determined by a two paired t-test, are printed bold.; **Significant ($P < .05$) changes between pre and post-treatment, determined by a paired t-test, are printed bold.

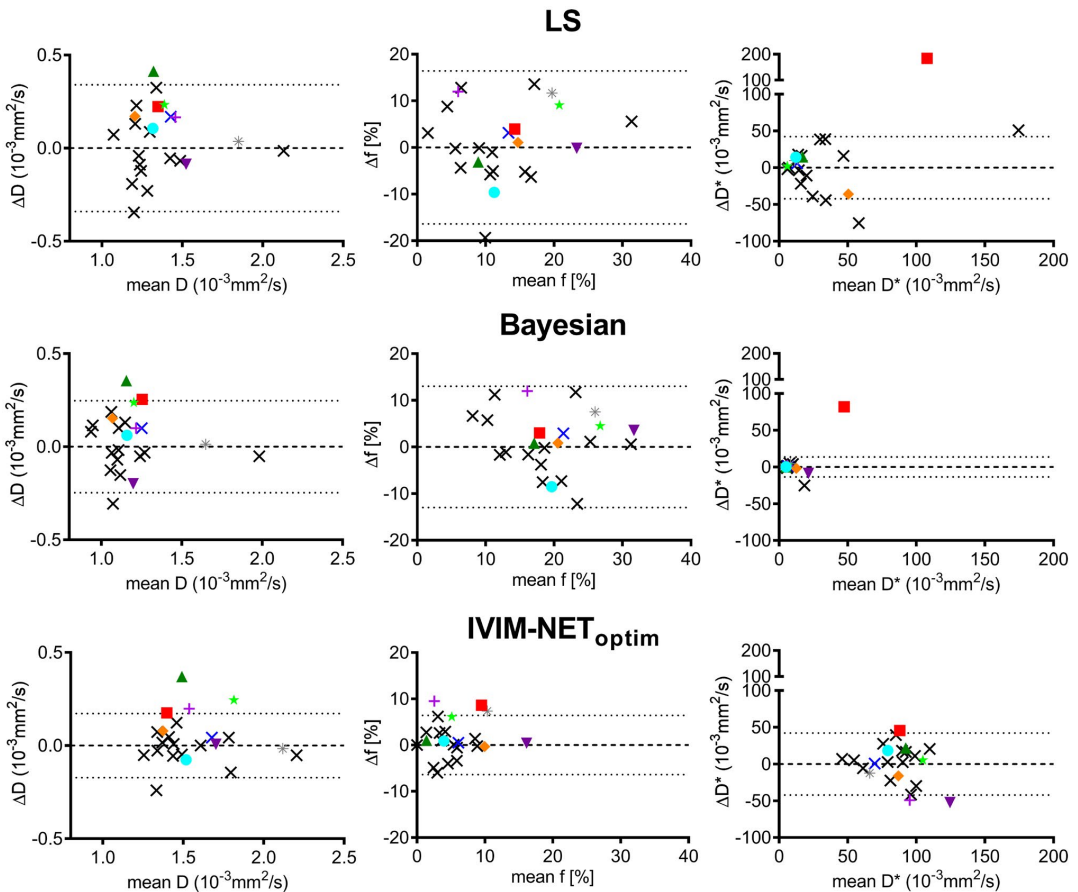


FIGURE 7 Bland-Altman plots of the LS, Bayesian, and IVIM-NET_{optim} approaches to IVIM fitting showing the mean and difference (Δ) between the inter-session repeatability patients (black crosses) and the mean and Δ between pre- and post-treatment patients (colored symbols), which represents the treatment effects. The dotted lines indicate the 95% CI of the test-retest data. Colored measurements that exceeded the 95% CI were considered significant to treatment response

interesting observation is that in Figure 6, IVIM-NET_{optim} shows a very similar f map to the LS approach for almost all tissue, suggesting again that there is no bias. Yet, contrary to the LS approach, IVIM-NET_{optim} has a homogeneously low perfusion map in the PDAC (ie, Figures 5 and 6; Supporting information Figures S11-S20). In the absence of ground truth in patients, we must rely on visual assessment of parameter map quality, the ability to detect treatment response, and the simulation results, in order to infer the best performing estimator. Given the evidence provided in this study, we argue that combined these factors rule in favor of IVIM-NET_{optim}.

Although IVIM-NET showed consistently better results both in simulations and in vivo, IVIM-NET predicts different IVIM parameters in repeated training. This causes a new sort of variability that, until now, was not an issue in fitting parameter maps. There may be methods to mitigate

this variability. First, when probing treatment response, we would advise using one network such that this additional effect is not different pre- and post-treatment. Second, to reduce the variation, one could consider taking the median prediction from 10 repeated trainings instead. We did so in an exploratory study where we formed 5 groups of 10 networks and showed that the median of 10 networks was substantially more consistent, with CV_{NET} values of 3.2×10^{-3} , 4.9×10^{-3} , and 9.3×10^{-3} for D , f , and D^* , respectively. Having a set of networks will also allow the user to estimate the variation on the predicted parameter. Finally, although we see this additional uncertainty, we would like to stress that it is secondary to the overall error of the LS approach, which is apparent from the fact that in the simulations, all 50 instances of IVIM-NET_{optim} had lower NRMSE than the LS approach.

IVIM-NET_{optim} outperformed IVIM-NET_{orig} at SNRs 8-100 and was superior to the LS and Bayesian approaches for SNRs 8-33 (Figure 4). However, at extremely high SNR (SNR = 100; Figure 4), the LS approach outperformed IVIM-NET. The Levenberg-Marquardt algorithm for the LS function is an iterative function that finds a minimum of the squared difference. For a relatively smooth loss landscape and high SNR signal, the LS algorithm is designed to find the correct parameter estimates. However, at low SNR, the LS approach has trouble finding the correct parameters. This occurs either because the loss landscape is no longer smooth and, hence, it gets stuck in a local minimum, or, what we believe is more common, the noise has changed the signal such that the global optimum no longer is nearby the ground truth parameters. Moreover, a DNN consists of a complex system that needs to encompass estimating the IVIM parameters for all voxels. It turns out that having been trained on all voxels enables better estimates for individual voxels at low SNR. We expect that DNNs focus on more consistent minima with parameter values that are more frequently observed. This might be similar to data-driven Bayesian fitting approaches.^{15,52} Conversely, IVIM-NET seems to reach a maximum accuracy at high SNR. Potentially more complex DNNs that are optimized with simulations done at high SNR could handle the subtle signal changes of the IVIM parameters at these SNRs. However, typical SNR values for IVIM data are <50. Therefore, our findings suggest that using IVIM-NET instead of the LS and Bayesian approaches for IVIM fitting would be beneficial in a clinical setting.

The choice of the hyperparameters for IVIM-NET_{optim} was based on an optimal combination of accuracy, independence, and consistency across all IVIM parameters. However, other hyperparameter options may be more appropriate when characterizing an individual IVIM parameter (eg, when an observer is only interested in D and IVIM is only used to correct for perfusion). Supporting Information Figures S2-S9 can help interested readers select the best network for their purposes.

The high dependency between D^* and f that appears in IVIM-NET_{orig} could not be attributed to a single cause. Initially, we expected that this dependency originated in the fully connected shared hidden layers of the original network. However, ρ remains substantial when adding the “parallel network architecture” to IVIM-NET_{orig} (Supporting Information Figure S10). Using IVIM-NET_{optim} and a single network architecture showed slightly worse performance in simulations regarding $\rho(D^*, f)$, but still had sufficient accuracy and consistency. The dependencies between the estimated IVIM parameters are not per se specific to unsupervised DNNs. For instance, similar dependencies between D^* and D or f were found in a different data-driven Bayesian fitting approach.⁹

For IVIM-NET_{optim}, these dependencies were small at clinical SNR values and similar to those of the LS approach.

Although simulation studies in parameter estimation are extremely valuable as the underlying parameter values are known, they also come with limitations. One limitation is that the noise characterization of real data can be diverse and hard to model. For instance, DWI artifacts caused by motion are not considered in simulations and may affect the results of fitting the IVIM model.⁵³ Another limitation is the underlying assumption that data are perfectly bi-exponential. In reality, the IVIM model is a simplification and real data will be more complex.

5 | CONCLUSIONS

We substantially improved the accuracy, independence, and consistency of both diffusion and perfusion parameters from IVIM-NET by changing the network architecture and tuning hyperparameters. Our new IVIM-NET_{optim} is considerably faster, and computes less noisy and more detailed parameter maps with substantially better test-retest repeatability for D and f than alternative state-of-the-art fitting methods. Furthermore, IVIM-NET_{optim} was able to detect the most individual patients with significant changes in the IVIM parameters throughout CRT. These results strongly suggest using IVIM-NET_{optim} for detection of treatment response in individual patients.


ACKNOWLEDGMENTS

The KWF Dutch Cancer Society supported this work under Grants No. KWF-UVA 18410 and KWF-UVA 2013.5932. M.P.T. Kaandorp and P.T. While gratefully acknowledge support from the Research Council of Norway under FRIPRO Researcher Project 302624.


DATA AVAILABILITY STATEMENT

To stimulate a wider clinical implementation of IVIM, we have made the code of IVIM-NET available on GitHub: <https://github.com/oliverchampion/IVIMNET>. This includes some simple introductory examples in simulations and additional volunteer data. We encourage our peers to use it in their research.


ORCID

Misha P. T. Kaandorp  <https://orcid.org/0000-0002-7340-8256>

Sebastiano Barbieri  <https://orcid.org/0000-0002-5919-372X>

Hanneke W. M. van Laarhoven  <https://orcid.org/0000-0003-3546-9709>

Peter T. While  <https://orcid.org/0000-0003-2602-0758>

Oliver J. Gurney-Champion  <https://orcid.org/0000-0003-1750-6617>

Oliver J. Gurney-Champion  <https://orcid.org/0000-0003-1750-6617>

Oliver J. Gurney-Champion  <https://orcid.org/0000-0003-1750-6617>

Oliver J. Gurney-Champion  <https://orcid.org/0000-0003-1750-6617>

Oliver J. Gurney-Champion  <https://orcid.org/0000-0003-1750-6617>

REFERENCES

- Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, Laval-Jeantet M. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. *Radiology*. 1988;168:497-505.
- Cho GY, Gennaro L, Sutton EJ, et al. Intravoxel incoherent motion (IVIM) histogram biomarkers for prediction of neoadjuvant treatment response in breast cancer patients. *Eur J Radiol Open*. 2017;4:101-107.
- Zhu LI, Wang H, Zhu L, et al. Predictive and prognostic value of intravoxel incoherent motion (IVIM) MR imaging in patients with advanced cervical cancers undergoing concurrent chemoradiotherapy. *Sci Rep*. 2017;7:1-9.
- Ma W, Zhang G, Ren J, et al. Quantitative parameters of intravoxel incoherent motion diffusion weighted imaging (IVIM-DWI): potential application in predicting pathological grades of pancreatic ductal adenocarcinoma. *Quant Imaging Med Surg*. 2018;8:301-310.
- Klaassen R, Steins A, Gurney-Champion OJ, et al. Pathological validation and prognostic potential of quantitative MRI in the characterization of pancreas cancer: preliminary experience. *Mol Oncol*. 2020;14:2176-2189.
- Paschoal AM, Leoni RF, dos Santos AC, Paiva FF. Intravoxel incoherent motion MRI in neurological and cerebrovascular diseases. *NeuroImage Clin*. 2018;20:705-714.
- Zhang Y-D, Wang Q, Wu C-J, et al. The histogram analysis of diffusion-weighted intravoxel incoherent motion (IVIM) imaging for differentiating the gleason grade of prostate cancer. *Eur Radiol*. 2015;25:994-1004.
- Li YT, Cercueil JP, Yuan J, Chen W, Loffroy R, Wang YXJ. Liver intravoxel incoherent motion (IVIM) magnetic resonance imaging: a comprehensive review of published data on normal values and applications for fibrosis and tumor evaluation. *Quant Imaging Med Surg*. 2017;7:59-78.
- Gurney-Champion OJ, Klaassen R, Froeling M, et al. Comparison of six fit algorithms for the intravoxel incoherent motion model of diffusion-weighted magnetic resonance imaging data of pancreatic cancer patients. *PLoS One*. 2018;13:1-18.
- Andreou A, Koh DM, Collins DJ, et al. Measurement reproducibility of perfusion fraction and pseudodiffusion coefficient derived by intravoxel incoherent motion diffusion-weighted MR imaging in normal liver and metastases. *Eur Radiol*. 2013;23:428-434.
- Barbieri S, Gurney-Champion OJ, Klaassen R, Thoeny HC. Deep learning how to fit an intravoxel incoherent motion model to diffusion-weighted MRI. *Magn Reson Med*. 2020;83:312-321.
- Keenan KE, Biller JR, Delfino JG, et al. Recommendations towards standards for quantitative MRI (qMRI) and outstanding needs. *J Magn Reson Imaging*. 2019;49:e26-e39.
- Neil JJ, Bretthorst GL. On the use of bayesian probability theory for analysis of exponential decay date: an example taken from intravoxel incoherent motion experiments. *Magn Reson Med*. 1993;29:642-647.
- Barbieri S, Donati OF, Froehlich JM, Thoeny HC. Impact of the calculation algorithm on biexponential fitting of diffusion-weighted MRI in upper abdominal organs. *Magn Reson Med*. 2016;75:2175-2184.
- Orton MR, Collins DJ, Koh DM, Leach MO. Improved intravoxel incoherent motion analysis of diffusion weighted imaging by data driven Bayesian modeling. *Magn Reson Med*. 2014;71:411-420.
- While PT. A comparative simulation study of Bayesian fitting approaches to intravoxel incoherent motion modeling in diffusion-weighted MRI. *Magn Reson Med*. 2017;78:2373-2387.
- Bertleff M, Domsch S, Weingärtner S, et al. Diffusion parameter mapping with the combined intravoxel incoherent motion and kurtosis model using artificial neural networks at 3 T. *NMR Biomed*. 2017;30:1-11.
- Hammernik K, Klatzer T, Kobler E, et al. Learning a variational network for reconstruction of accelerated MRI data. *Magn Reson Med*. 2018;79:3055-3071.
- Aggarwal HK, Mani MP, Jacob M. Modl-mussels: model-based deep learning for multishot sensitivity-encoded diffusion MRI. *IEEE Trans Med Imaging*. 2020;39:1268-1277.
- Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. *NIPS-W*. 2017;22:2-8.
- Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). 4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc. 2016, pp. 1-14.
- Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929-1958.
- Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *32nd Int Conf Mach Learn ICML 2015*. 2015;1:448-456.
- Kingma DP, Ba JL. Adam: a method for stochastic optimization. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. 2015, pp. 1-15.
- van Herten RLM, Chiribiri A, Breeuwer M, Veta M, Scannell CM. Physics-informed neural networks for myocardial perfusion MRI quantification. *arXiv*. 2020;2011.12844.
- Barbieri S, Donati OF, Froehlich JM, Thoeny HC. Comparison of intravoxel incoherent motion parameters across MR imagers and field strengths: evaluation in upper abdominal organs. *Radiology*. 2016;279:784-794.
- Koay CG, Basser PJ. Analytically exact correction scheme for signal extraction from noisy magnitude MR signals. *J Magn Reson*. 2006;179:317-322.
- Levenberg K. A method for the solution of certain non-linear problems in Least squares. *Q Appl Math*. 1944;2:164-168.
- Marquardt DW. An algorithm for least-squares estimation of non-linear parameters. *J Soc Ind Appl Math*. 1963;11:431-441.
- Cho GY, Moy L, Zhang JL, et al. Comparison of fitting methods and b-value sampling strategies for intravoxel incoherent motion in breast cancer. *Magn Reson Med*. 2015;74:1077-1085.
- Lemke A, Stieltjes B, Schad LR, Laun FB. Toward an optimal distribution of b values for intravoxel incoherent motion imaging. *Magn Reson Imaging*. 2011;29:766-776.
- Cohen AD, Schieke MC, Hohenwarter MD, Schmainda KM. The effect of low b-values on the intravoxel incoherent motion derived pseudodiffusion parameter in liver. *Magn Reson Med*. 2015;73:306-311.
- Jalnefjord O, Andersson M, Montelius M, et al. Comparison of methods for estimation of the intravoxel incoherent motion (IVIM) diffusion coefficient (D) and perfusion fraction (f). *Magn Reson Mater Phys Biol Med*. 2018;31:715-723.
- Klaassen R, Gurney-Champion OJ, Wilmink JW, et al. Repeatability and correlations of dynamic contrast enhanced and T2* MRI in patients with advanced pancreatic ductal adenocarcinoma. *Magn Reson Imaging*. 2018;50:1-9.
- Klaassen R, Gurney-Champion OJ, Engelbrecht MRW, et al. Evaluation of six diffusion-weighted MRI models for assessing effects of neoadjuvant chemoradiation in pancreatic cancer patients. *Int J Radiat Oncol Biol Phys*. 2018;102:1052-1062.

36. Versteijne E, Suker M, Groothuis K, et al. Original reports abstract preoperative chemoradiotherapy versus immediate surgery for resectable and borderline resectable pancreatic cancer: results of the dutch randomized phase III PREOPANC Trial. *J Clin Oncol*. 2020;38:1763-1774.
37. Klein S, Staring M, Murphy K, Viergever MA, Pluim J. A toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging*. 2010;29:196-205.
38. Barnhart HX, Barboriak DP. Applications of the repeatability of quantitative imaging biomarkers: a review of statistical analysis of repeat data sets. *Transl Oncol*. 2009;2:231-235.
39. Bland JM, Altman DG. Statistics notes: measurement error. *BMJ*. 1996;313:744.
40. Trajkovic-Arsic M, Heid I, Steiger K, et al. Apparent diffusion coefficient (ADC) predicts therapy response in pancreatic ductal adenocarcinoma. *Sci Rep*. 2017;7:1-9.
41. Kim H, Morgan DE, Schexnailder P, et al. Accurate therapeutic response assessment of pancreatic ductal adenocarcinoma using quantitative dynamic contrast-enhanced magnetic resonance imaging with a point-of-care perfusion phantom: a pilot study. *Invest Radiol*. 2019;54:16-22.
42. Robertis RDe, Martini PT, Demozzi E, et al. Diffusion-weighted imaging of pancreatic cancer. *World J Radiol*. 2015;7:319.
43. Sofuni A, Iijima H, Moriyasu F, et al. Differential diagnosis of pancreatic tumors using ultrasound contrast imaging. *J Gastroenterol*. 2005;40:518-525.
44. Stopa KB, Kusiak AA, Szopa MD, Ferdek PE, Jakubowska MA. Pancreatic cancer and its microenvironment—recent advances and current controversies. *Int J Mol Sci*. 2020;21:3218.
45. Zhu LI, Zhu L, Shi H, et al. Evaluating early response of cervical cancer under concurrent chemo-radiotherapy by intravoxel incoherent motion MR imaging. *BMC Cancer*. 2016;16:1-8.
46. Chiaradia M, Baranes L, Van Nhieu JT, et al. Intravoxel incoherent motion (IVIM) MR imaging of colorectal liver metastases: are we only looking at tumor necrosis? *J Magn Reson Imaging*. 2014;39:317-325.
47. Kim H, Samuel S, Lopez-Casas P, et al. SPARC-independent delivery of nab-paclitaxel without depleting tumor stroma in patient-derived pancreatic cancer xenografts. *Mol Cancer Ther*. 2016;15:680-688.
48. Hamdy A, Ichikawa Y, Toyomasu Y, et al. Perfusion CT to assess response to neoadjuvant chemotherapy and radiation therapy in pancreatic ductal adenocarcinoma: initial experience. *Radiology*. 2019;292:628-635.
49. Barbieri S, Donati OF, Froehlich JM, Thoeny HC. Impact of the calculation algorithm on biexponential fitting of diffusion-weighted MRI in upper abdominal organs. *Magn Reson Med*. 2016;75:2175-2184.
50. Meeus EM, Novak J, Withey SB, Zarinabad N, Dehghani H, Peet AC. Evaluation of intravoxel incoherent motion fitting methods in low-perfused tissue. *J Magn Reson Imaging*. 2017;45:1325-1334.
51. Gurney-Champion OJ, Froeling M, Klaassen R, et al. Minimizing the acquisition time for intravoxel incoherent motion magnetic resonance imaging acquisitions in the liver and pancreas. *Invest Radiol*. 2016;51:211-220.
52. Friedman N, Goldszmidt M, Wyner A. Data analysis with Bayesian networks: a bootstrap approach. *Uncertainty Artif Intell*. 1999:206-215.
53. Veraart J, Rajan J, Peeters RR, Leemans A, Sunaert S, Sijbers J. Comprehensive framework for accurate diffusion MRI parameter estimation. *Magn Reson Med*. 2013;70:972-984.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

FIGURE S1 Plots of the estimated IVIM parameters (D , f , D^*) where no Spearman rank correlation coefficient (ρ) can be determined and is set to a ρ of 1. In all plots, the values of the simulations are presented in grey

FIGURE S2 Normalized root-mean-square error (NRMSE) boxplots of the estimated IVIM parameters (D , f , D^*) that contain all hyperparameter combinations with a fixed learning rate set to 1×10^{-4} and a fixed number of hidden layers set to 3 at SNR 20 for 50 repeated trainings. Highlighted in green is the intermediate step of IVIM-NET_{optim}. Left of each plot shows the LS approach (blue), Bayesian approach (brown) and IVIM-NET_{orig} (orange; LR = 1×10^{-3})

FIGURE S3 Spearman rank correlation coefficient (ρ) boxplots of the estimated IVIM parameters (D , f , D^*) that contain all hyperparameter combinations with a fixed learning rate set to 1×10^{-4} and a fixed number of hidden layers set to 3 at SNR 20 for 50 repeated trainings. Highlighted in green is the intermediate step of IVIM-NET_{optim}. Left of each plot shows the LS approach (blue), Bayesian approach (brown) and IVIM-NET_{orig} (orange; LR = 1×10^{-3})

FIGURE S4 Normalized Coefficient of variation (CV_{NET}) plots of the estimated IVIM parameters (D , f , D^*) that contain all hyperparameter combinations with a fixed learning rate set to 1×10^{-4} and a fixed number of hidden layers set to 3 at SNR 20 for 50 repeated trainings. Highlighted in green is the intermediate step of IVIM-NET_{optim}. Left of each plot shows the LS approach (blue), Bayesian approach (brown) and IVIM-NET_{orig} (orange; LR = 1×10^{-3})

FIGURE S5 Ranked plots of the metrics (NRMSE, ρ and CV_{NET}) of evaluation 1 that contain all hyperparameter combinations with a fixed learning rate set to 1×10^{-4} and a fixed number of hidden layers set to 3 at SNR 20 for 50 repeated trainings. Highlighted in green is the intermediate step of IVIM-NET_{optim}. Left of each plot shows the LS approach (blue), Bayesian approach (brown) and IVIM-NET_{orig} (orange; LR = 1×10^{-3})

FIGURE S6 Normalized root-mean-square error (NRMSE) boxplots of the estimated IVIM parameters (D , f , D^*) of the second evaluation for different LR and number of hidden layers, with fixed hyperparameters of extra fitting parameter S_0 , sigmoid activation functions, a parallel network architecture, 10% dropout and batch normalization at SNR 20 for 50 repeated trainings. Highlighted in green is IVIM-NET_{optim}. Left of each plot shows the LS approach (blue) and Bayesian approach (brown) and IVIM-NET_{orig} (orange)

FIGURE S7 Spearman rank correlation coefficient (ρ) boxplots of the estimated IVIM parameters (D , f , D^*) of the second evaluation for different LR and number of hidden layers, with fixed hyperparameters of extra fitting parameter

S0, sigmoid activation functions, a parallel network architecture, 10% dropout and batch normalization at SNR 20 for 50 repeated trainings. Highlighted in green is IVIM-NET_{optim}. Left of each plot shows the LS approach (blue) and Bayesian approach (brown) and IVIM-NET_{orig} (orange)

FIGURE S8 Normalized coefficient of variation (CV_{NET}) plots of the estimated IVIM parameters (D, f, D^*) of the second evaluation for different LR and number of hidden layers, with fixed hyperparameters of extra fitting parameter S0, sigmoid activation functions, a parallel network architecture, 10% dropout and batch normalization at SNR 20 for 50 repeated trainings. Highlighted in green is IVIM-NET_{optim}. Left of each plot shows the LS approach (blue) and Bayesian approach (brown) and IVIM-NET_{orig} (orange)

FIGURE S9 Ranked plots of the metrics (NRMSE, ρ and CV_{NET}) of evaluation 2 for different LR and number of hidden layers, with fixed hyperparameters of extra fitting parameter S0, sigmoid activation functions, a parallel network architecture, 10% dropout and batch normalization at SNR 20 for 50 repeated trainings. Highlighted in green is IVIM-NET_{optim}. Left of each plot shows the LS approach (blue) and Bayesian approach (brown) and IVIM-NET_{orig} (orange)

FIGURE S10 Normalised root-mean-square error (NRMSE; left), Spearman rank correlation coefficient (ρ ; center) and normalized coefficient of variation (CV_{NET} ; right) plots of

the estimated IVIM parameters (D, f and D^*) with a single parameter change for IVIM-NET_{orig} (orange) at SNR 20 for 50 repeated trainings. The $\rho(D^*, f)$ remains substantial for single deviations from IVIM-NET_{orig}

FIGURE S11 See Table S1

FIGURE S12 See Table S1

FIGURE S13 See Table S1

FIGURE S14 See Table S1

FIGURE S15 See Table S1

FIGURE S16 See Table S1

FIGURE S17 See Table S1

FIGURE S18 See Table S1

FIGURE S19 See Table S1

FIGURE S20 See Table S1

TABLE S1 Overview of the parameter maps of Supporting Information Figures S11-S20

How to cite this article: Kaandorp MPT, Barbieri S, Klaassen R, et al. Improved unsupervised physics-informed deep learning for intravoxel incoherent motion modeling and evaluation in pancreatic cancer patients. *Magn Reson Med*. 2021;86:2250–2265. <https://doi.org/10.1002/mrm.28852>

Supporting Information 1: Simulations

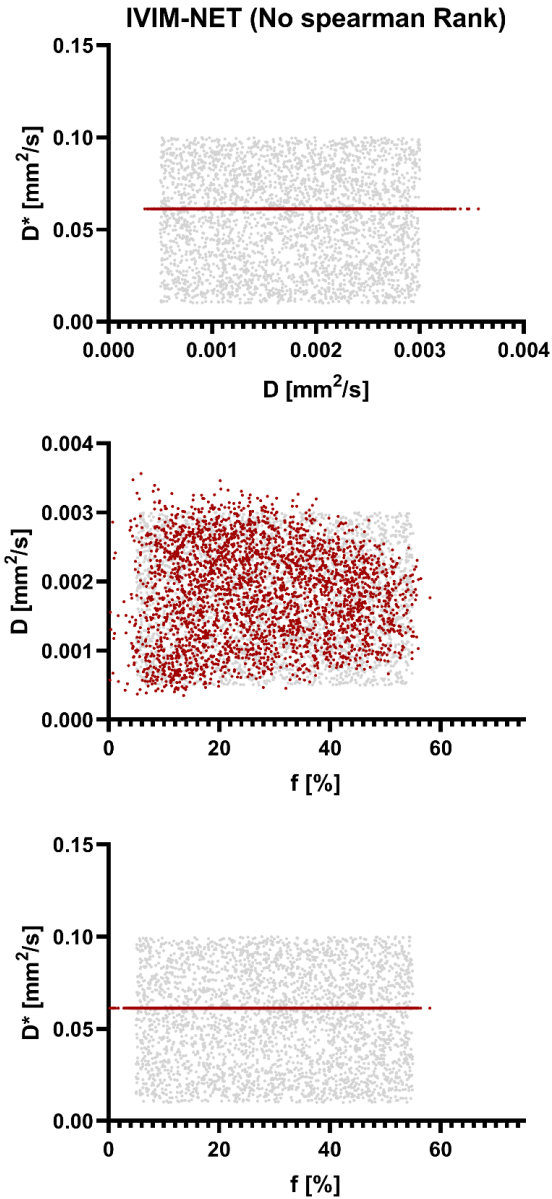


Figure S1: Plots of the estimated IVIM parameters (D , f , D^*) where no Spearman rank correlation coefficient (ρ) can be determined and is set to a ρ of 1. In all plots, the values of the simulations are presented in grey.

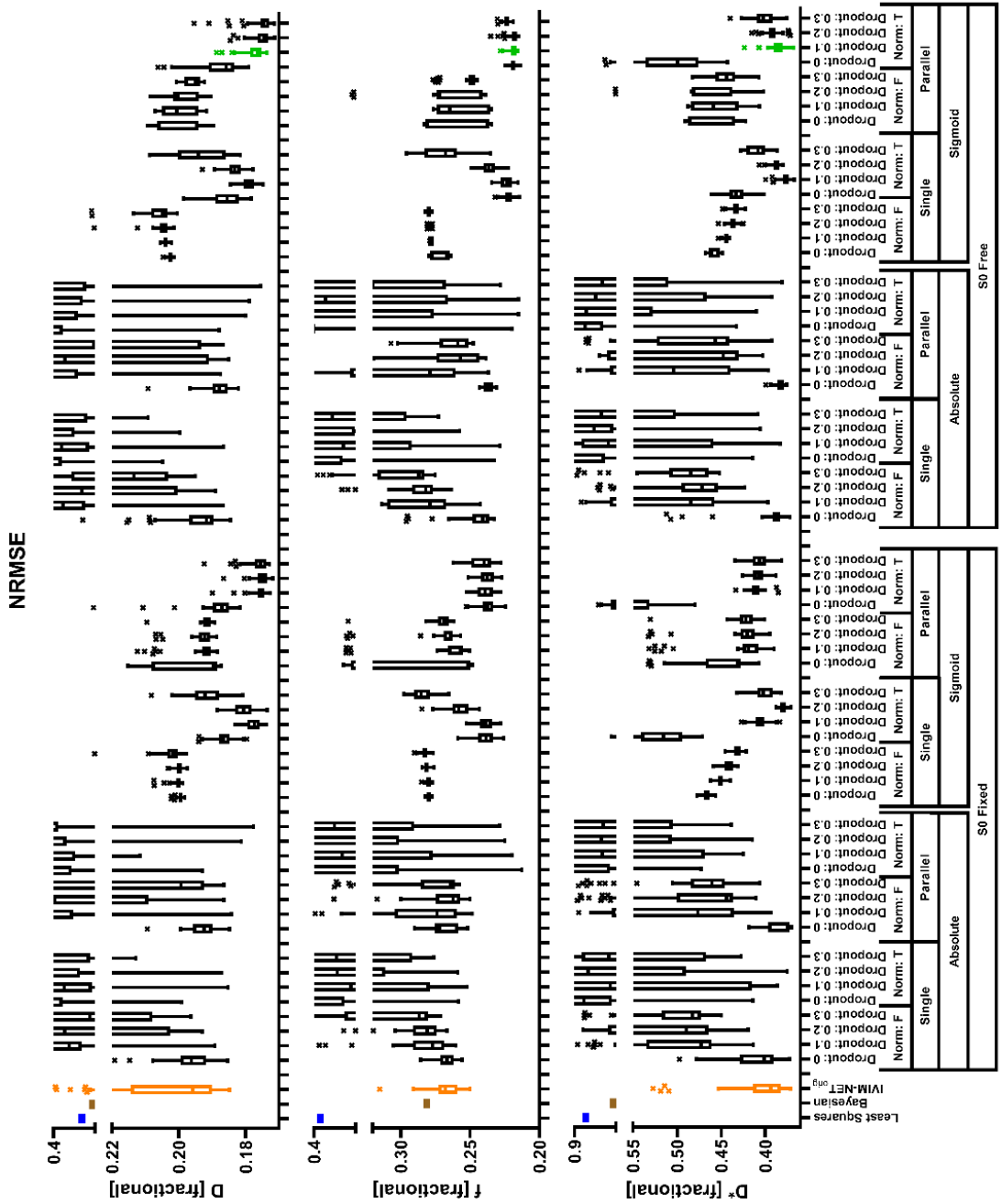


Figure S2: Normalized root-mean-square error (NRMSE) boxplots of the estimated IVIM parameters (D, f, D^*) that contain all hyperparameter combinations with a fixed learning rate set to 1×10^{-4} and a fixed number of hidden layers set to 3 at SNR 20 for 50 repeated trainings. Highlighted in green is the intermediate step of IVIM-NET_{optim}. Left of each plot shows the LS approach (blue), Bayesian approach (brown) and IVIM-NET_{orig} (orange; LR = 1×10^{-3}).

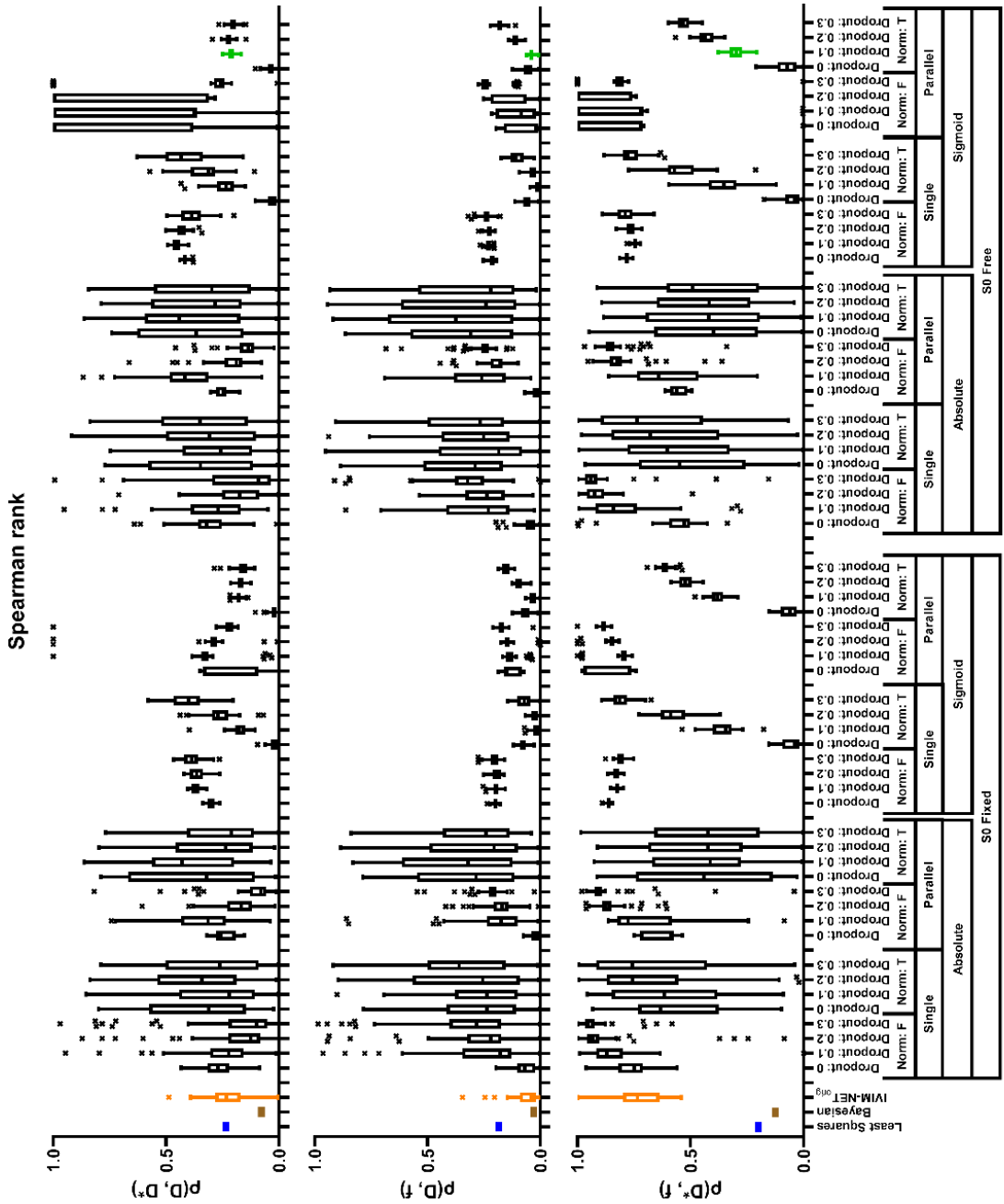


Figure S3: Spearman rank correlation coefficient (ρ) boxplots of the estimated IVIM parameters (D , f , D^*) that contain all hyperparameter combinations with a fixed learning rate set to 1×10^{-4} and a fixed number of hidden layers set to 3 at SNR 20 for 50 repeated trainings. Highlighted in green is the intermediate step of IVIM-NET_{optim}. Left of each plot shows the LS approach (blue), Bayesian approach (brown) and IVIM-NET_{orig} (orange; LR = 1×10^{-3}).

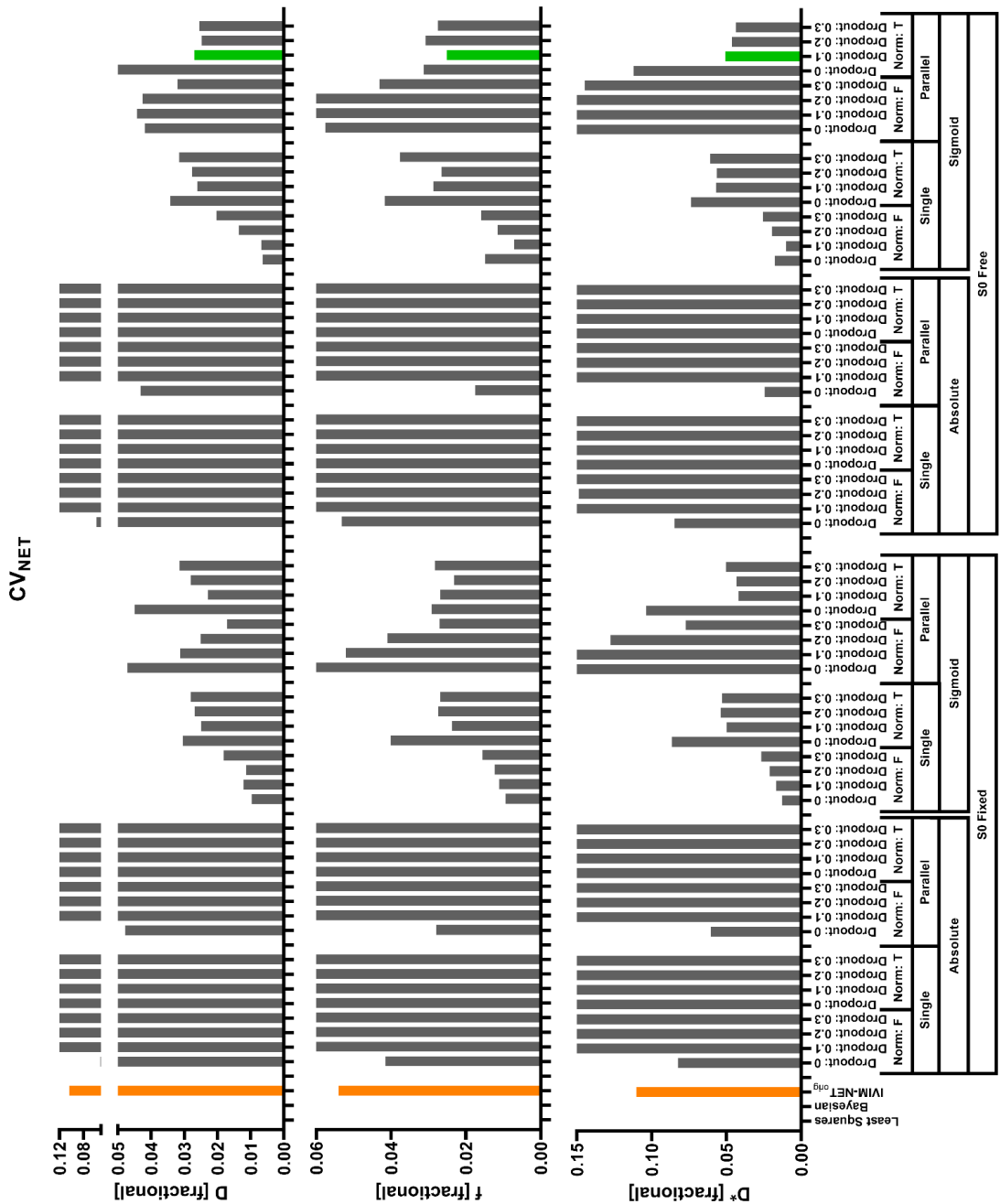


Figure S4: Normalized coefficient of variation (CV_{NET}) plots of the estimated IVIM parameters (D , f , D^*) that contain all hyperparameter combinations with a fixed learning rate set to 1×10^{-4} and a fixed number of hidden layers set to 3 at SNR 20 for 50 repeated trainings. Highlighted in green is the intermediate step of IVIM-NET_{optim}. Left of each plot shows the LS approach (blue), Bayesian approach (brown) and IVIM-NET_{orig} (orange; LR = 1×10^{-3}). As the LS and Bayesian approaches are deterministic, their CV_{NET} was zero and not plotted.

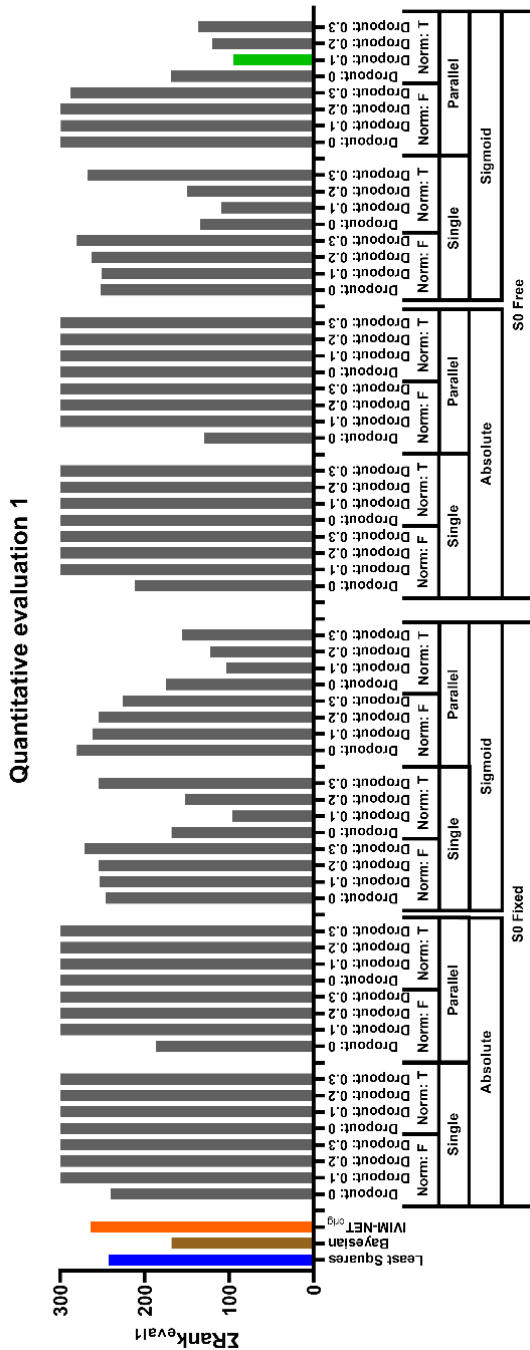


Figure S5: Ranked plots of the metrics (NRMSE, ρ and CV_{NET}) of evaluation 1 that contain all hyperparameter combinations with a fixed learning rate set to 1×10^{-4} and a fixed number of hidden layers set to 3 at SNR 20 for 50 repeated trainings. Highlighted in green is the intermediate step of IVIM-NET_{optim}. Left of each plot shows the LS approach (blue), Bayesian approach (brown) and IVIM-NET_{orig} (orange; LR = 1×10^{-3}).

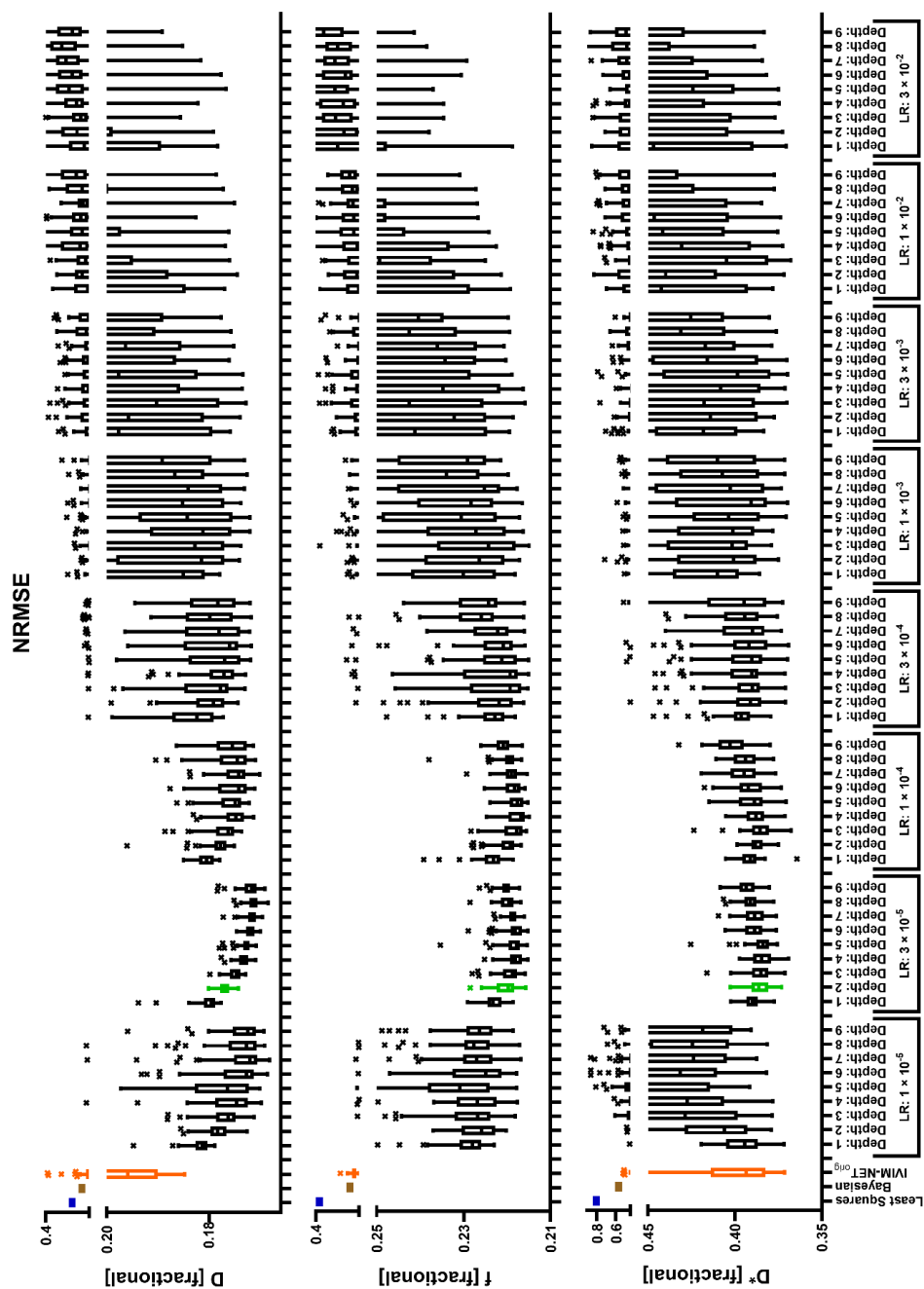


Figure S6: Normalized root-mean-square error (NRMSE) boxplots of the estimated IVIM parameters (D , f , D^*) of the second evaluation for different LR and number of hidden layers, with fixed hyperparameters of extra fitting parameter S_0 , sigmoid activation functions, a parallel network architecture, 10% dropout and batch normalization at SNR 20 for 50 repeated trainings. Highlighted in green is IVIM-NET_{optim}. Left of each plot shows the LS approach (blue) and Bayesian approach (brown) and IVIM-NET_{orig} (orange).

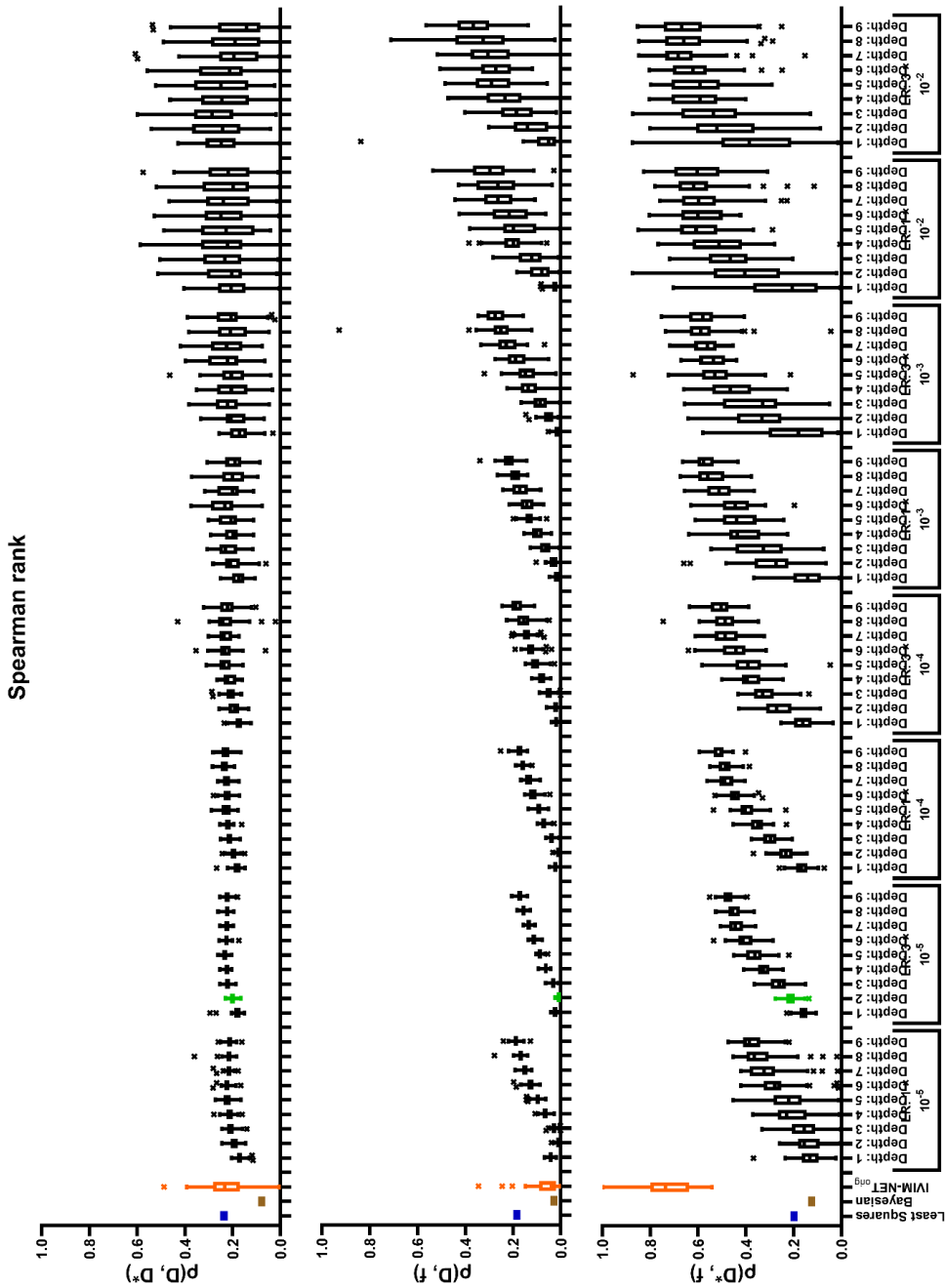


Figure S7: Spearman rank correlation coefficient (ρ) boxplots of the estimated IVIM parameters (D, f, D^*) of the second evaluation for different LR and number of hidden layers, with fixed hyperparameters of extra fitting parameter S_0 , sigmoid activation functions, a parallel network architecture, 10% dropout and batch normalization at SNR 20 for 50 repeated trainings. Highlighted in green is IVIM-NET_{optim}. Left of each plot shows the LS approach (blue) and Bayesian approach (brown) and IVIM-NET_{orig} (orange).

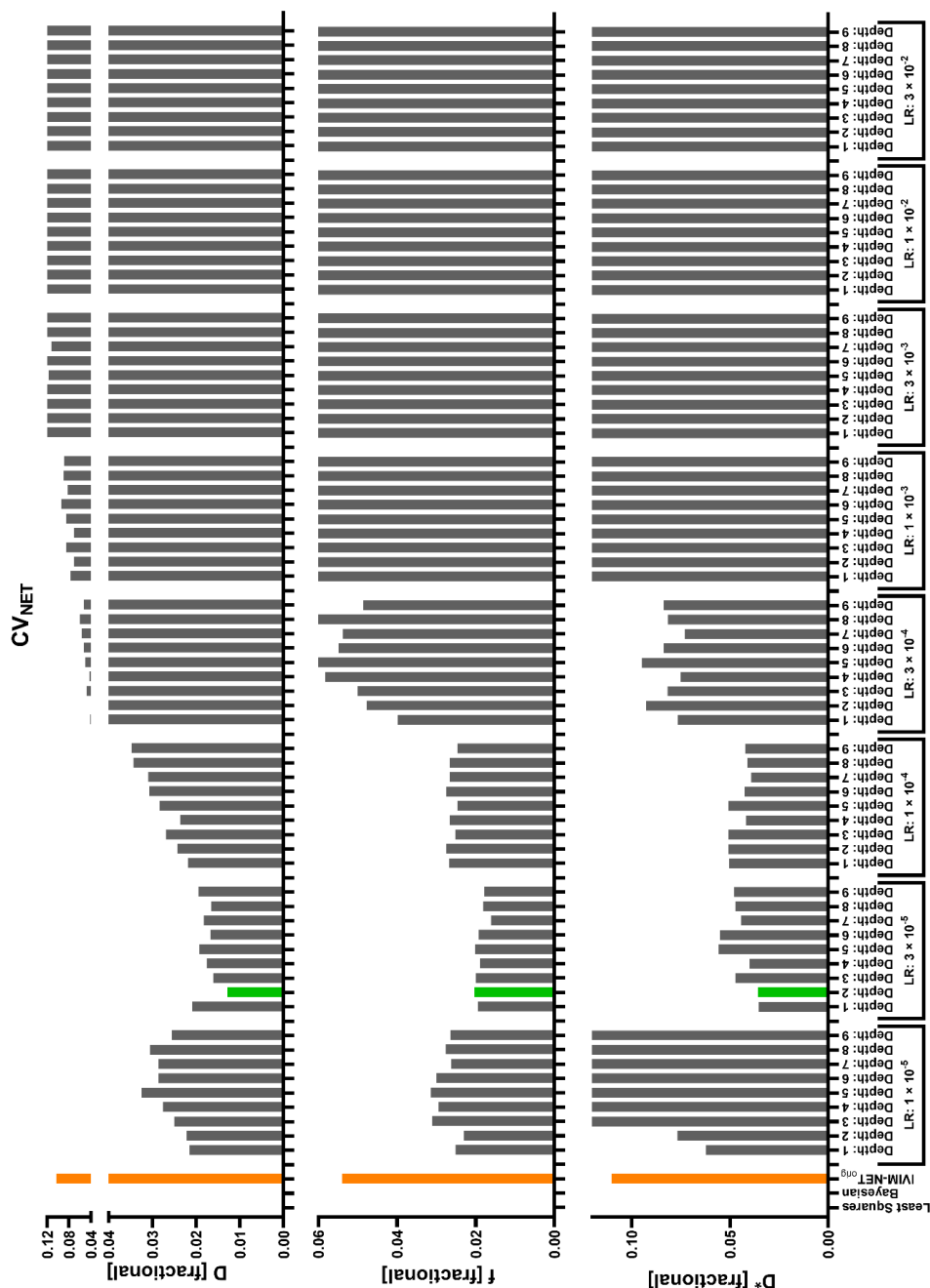


Figure S8: Normalized coefficient of variation (CV_{NET}) plots of the estimated IVIM parameters (D, f, D^*) of the second evaluation for different LR and number of hidden layers, with fixed hyperparameters of extra fitting parameter S_0 , sigmoid activation functions, a parallel network architecture, 10% dropout and batch normalization at SNR 20 for 50 repeated trainings. Highlighted in green is $IVIM-NET_{optim}$. Left of each plot shows the LS approach (blue) and Bayesian approach (brown) and $IVIM-NET_{orig}$ (orange). As the LS and Bayesian approaches are deterministic, their CV_{NET} was zero and not plotted.

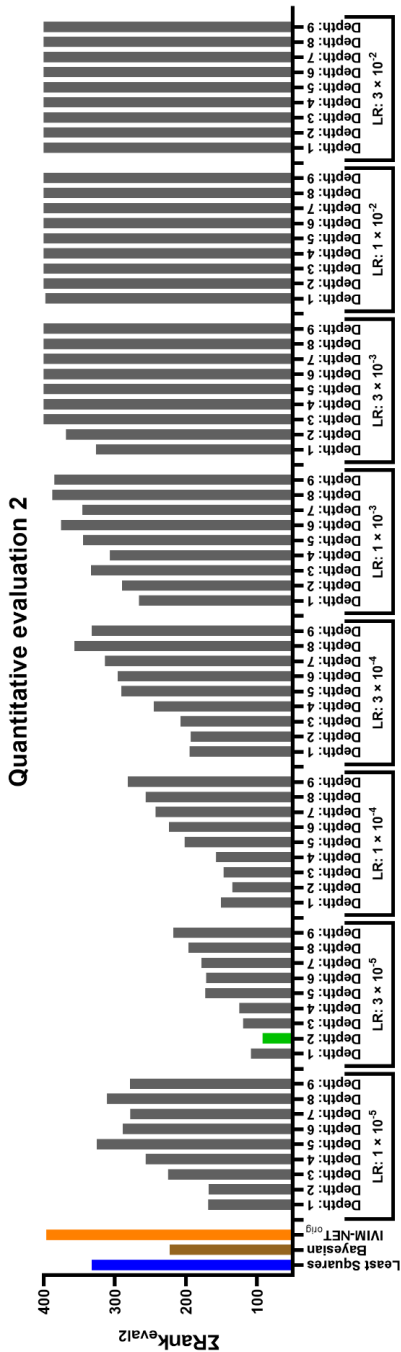


Figure S9: Ranked plots of the metrics (NRMSE, ρ and CV_{NET}) of evaluation 2 for different LR and number of hidden layers, with fixed hyperparameters of extra fitting parameter S_0 , sigmoid activation functions, a parallel network architecture, 10% dropout and batch normalization at SNR 20 for 50 repeated trainings. Highlighted in green is IVIM-NET_{optim}. Left of each plot shows the LS approach (blue) and Bayesian approach (brown) and IVIM-NET_{orig} (orange).

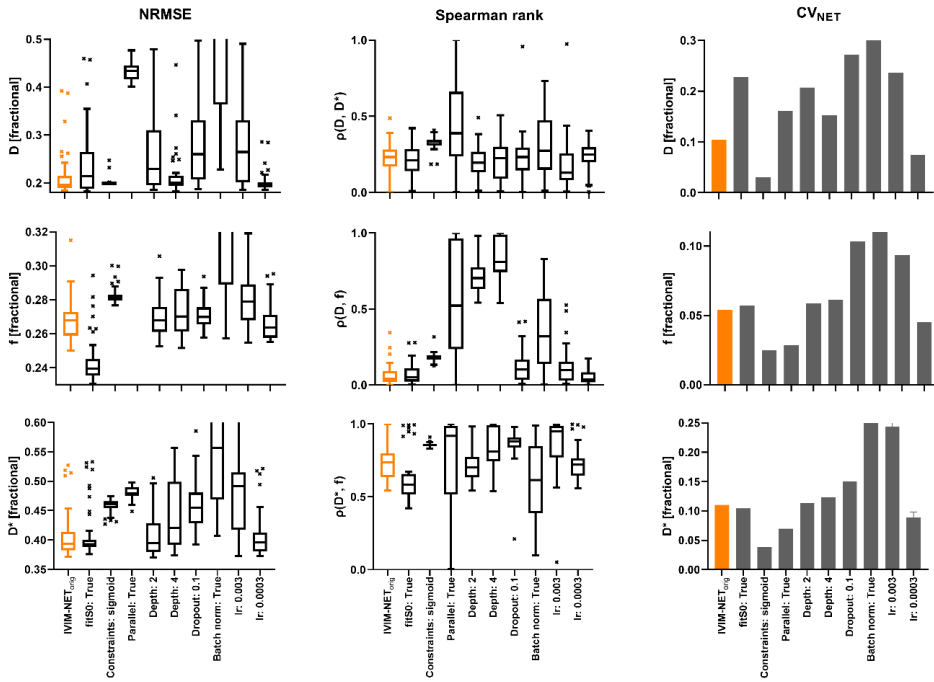


Figure S10: Normalised root-mean-square error (NRMSE; left), Spearman rank correlation coefficient (ρ ; center) and normalized coefficient of variation (CV_{NET}; right) plots of the estimated IVIM parameters (D , f and D^*) with a single parameter change for IVIM-NET_{orig} (orange) at SNR 20 for 50 repeated trainings. The $\rho(D^*, f)$ remains substantial for single deviations from IVIM-NET_{orig}.

Supporting Information 2: Verification in patients with PDAC

For simplicity, we have made a table overview (Table S1) of every parameter maps of the Supporting Information. As in Figures 5 and 6 of our manuscript, these Supporting Information Figures S11-S20 show IVIM parameter maps (D, f, D^*) of the LS approach, Bayesian approach and IVIM-NET_{optim} of a PDAC patient of the treated cohort before CRT or the test-retest cohort. The red ROI represents the PDAC and the green ROI represents homogenous 2D liver tissue ROI. The two highlighted blue regions correlate to the voxels from the plots below. The yellow square zooms in on the two highlighted voxels. In the plots, the small light grey dots are the repeated measures and the big black dots are the root-mean-squares of these repeated measures. The plot parameters are shown below.

Table S1: Overview of the parameter maps of Supporting Information Figures S11-S20.

Figure	Remark
S11	Highlighted voxels in the liver. The light blue voxel (left plot) shows consistency in IVIM parameters between the LS approach and IVIM-NET _{optim} with low f , and moderate D and D^* , while the Bayesian approach shows higher f , lower D and very low D^* . The neighboring dark blue voxel (right plot) shows no diffusion ($D = 0$ mm ² /s), a very high f and very low D^* for the LS and Bayesian approaches. IVIM-NET _{optim} shows more consistency in IVIM parameters between the two neighboring voxels. The LS and Bayesian approaches show noisier parameter maps, particularly in the liver and around the tumor region.
S12	Highlighted voxels in the tumor. The light blue voxel (left plot) shows indifferent IVIM parameters for all three fitting approaches. Although the data is similar in the neighboring dark blue voxel (right plot), the LS and Bayesian approaches compute a higher f , lower D (with $D = 0$ mm ² /s for the LS approach) and very low D^* (to the lower bound of $D^* = 5.0 \times 10^{-3}$ mm ² /s for the LS approach) compared to their parameters in the light blue voxel. IVIM-NET _{optim} shows more consistency in IVIM parameters between the two neighboring voxels. The LS and Bayesian approaches show noisier parameter maps, particularly in the liver and around the tumor region. Note that there is an artifact, which can be seen best in the middle part of the liver.
S13	Highlighted voxels in the liver. The light blue voxel (left plot) shows consistency in IVIM parameters for all three fitting approaches with a high IVIM effect. Although the data is similar in the neighboring dark blue voxel (right plot), the LS and Bayesian approaches compute a higher f , and lower D and D^* compared to their parameters in the light blue voxel. IVIM-NET _{optim} shows more consistency in IVIM parameters between the two neighboring voxels. The LS and Bayesian approaches show noisier parameter maps, particularly in the liver and around the tumor region.
S14	Highlighted voxels in the tumor. The light blue voxel (left plot) shows consistency in IVIM parameters for all three fitting approaches. Although the data is similar in the neighboring dark blue voxel (right plot), the LS and Bayesian approaches compute a higher f , and lower D and very low D^* (to the lower bound of $D^* = 5.0 \times 10^{-3}$ mm ² /s for the LS approach) compared to their parameters in the light blue voxel. IVIM-NET _{optim} shows more consistency in IVIM parameters between the two neighboring voxels. The LS and Bayesian approaches show noisier parameter maps, particularly in the liver, kidneys and around the tumor region.
S15	Highlighted voxels in the tumor. The light blue voxel (left plot) shows consistency in IVIM parameters between the LS approach and IVIM-NET _{optim} with low f , and moderate D and D^* , while the Bayesian approach shows higher f , lower D and very low D^* . Although the data is similar in the neighboring dark blue voxel (right plot), the LS and Bayesian approaches compute a higher f , and lower D and very low D^* (to the lower bound of $D^* = 5.0 \times 10^{-3}$ mm ² /s) compared to their parameters in the light blue voxel. IVIM-NET _{optim} shows more consistency in IVIM parameters between the two neighboring voxels. The LS and Bayesian approaches show noisier parameter maps, particularly in the kidneys and around the tumor region.
S16	Highlighted voxels in the tumor. The light blue voxel (left plot) shows consistency in IVIM parameters for all three fitting approaches. Although the data is similar in the neighboring dark blue voxel (right plot) with a lower IVIM effect, the LS and Bayesian approaches compute a higher f , and lower D and very low D^* (to the lower bound of $D^* = 5.0 \times 10^{-3}$ mm ² /s) compared to their parameters in the light blue voxel. IVIM-NET _{optim} shows more consistency in IVIM parameters between the two neighboring voxels with a lower f . The LS and Bayesian approaches show noisier parameter maps, particularly around the tumor region. Note that the LS approach has a very high D^* (to the upper bound of $D^* = 200 \times 10^{-3}$ mm ² /s) in the blue voxel, while in the neighboring dark blue voxel it has a very low D^* (to the lower bound of $D^* = 5.0 \times 10^{-3}$ mm ² /s).
S17	Highlighted voxels in the liver. The light blue voxel (left plot) shows consistency in IVIM parameters for all three fitting approaches with a high IVIM effect. Although the data is similar in the neighboring dark blue voxel (right plot) with a lower IVIM effect, the LS and Bayesian approaches compute a higher f , and lower D and D^* compared to their parameters in the light blue voxel. IVIM-NET _{optim} shows more consistency in IVIM parameters between the two neighboring voxels with a lower f . The LS and Bayesian approaches show noisier parameter maps, particularly in the liver.
S18	Highlighted voxels in the kidneys. The light blue voxel (left plot) shows consistency in IVIM parameters for all three fitting approaches. Although the data is similar in the neighboring dark blue voxel (right plot) with a lower IVIM effect, the Bayesian approaches compute a higher f , and lower D and very low D^* compared to its parameters in the light blue voxel. IVIM-NET _{optim} and the LS approach show more consistency in IVIM parameters between the two neighboring voxels with a lower f . The LS and Bayesian approaches show noisier parameter maps, particularly in the liver, kidneys and around the tumor region.
S19	Highlighted voxels in the liver. The light blue voxel (left plot) shows consistency in IVIM parameters between the LS approach and IVIM-NET _{optim} with low f , and moderate D and D^* , while the Bayesian approach shows higher f , lower D and very low D^* . Although the data is similar in the neighboring dark blue voxel (right plot), the LS and Bayesian approaches compute a higher f , and lower D and very low D^* compared to their parameters in the light blue voxel. IVIM-NET _{optim} shows more consistency in IVIM parameters between the two neighboring voxels. The LS and Bayesian approaches show noisier parameter maps, particularly in the liver.
S20	Highlighted voxels in the liver. The light blue voxel (left plot) shows consistency in IVIM parameters for all three fitting approaches. Although the data is similar in the neighboring dark blue voxel (right plot) with a lower IVIM effect, the Bayesian approaches compute a higher f , and lower D and very low D^* compared to its parameters in the light blue voxel. IVIM-NET _{optim} and the LS approach show more consistency in IVIM parameters between the two neighboring voxels with a lower f . The LS approach shows noisier parameter maps, particularly in the liver, kidneys and around the tumor region. Note that although the LS approach has a very high D^* (to the upper bound of $D^* = 200 \times 10^{-3}$ mm ² /s) in both voxels, the LS and IVIM-NET _{optim} show the same plots.

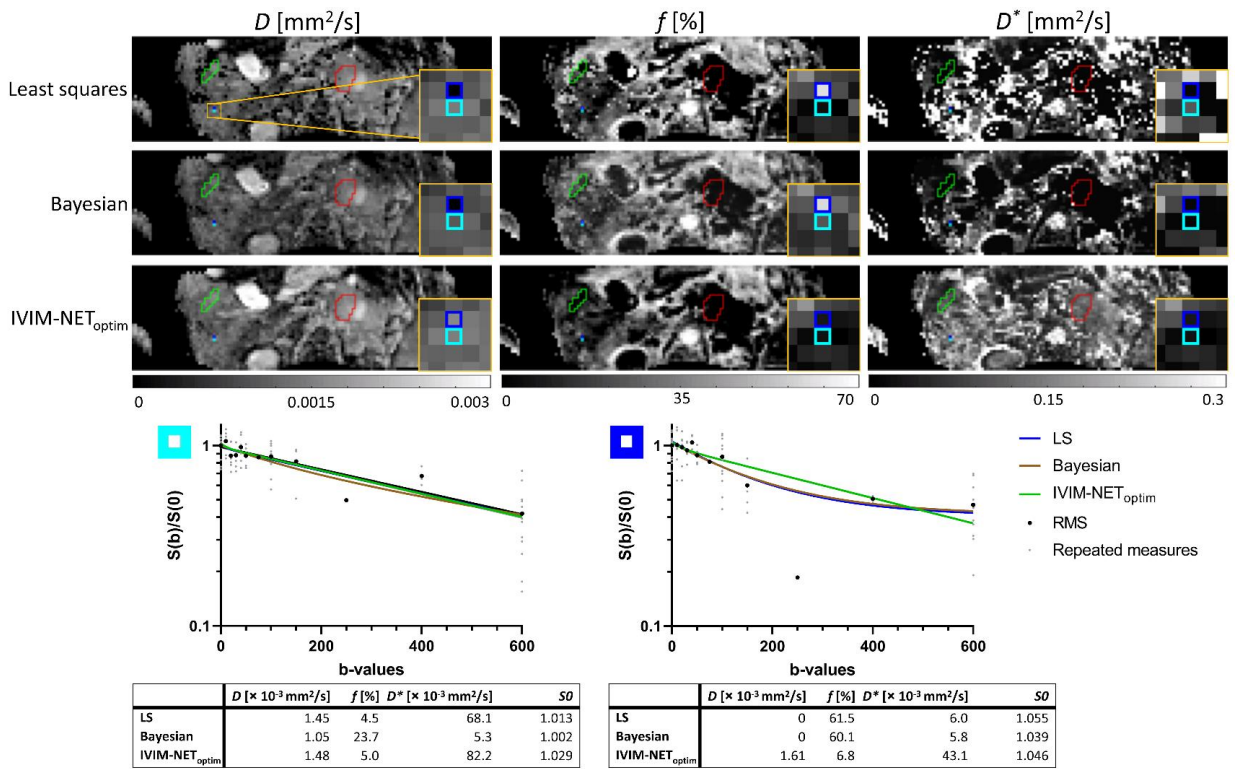


Figure S11: See Table S1.

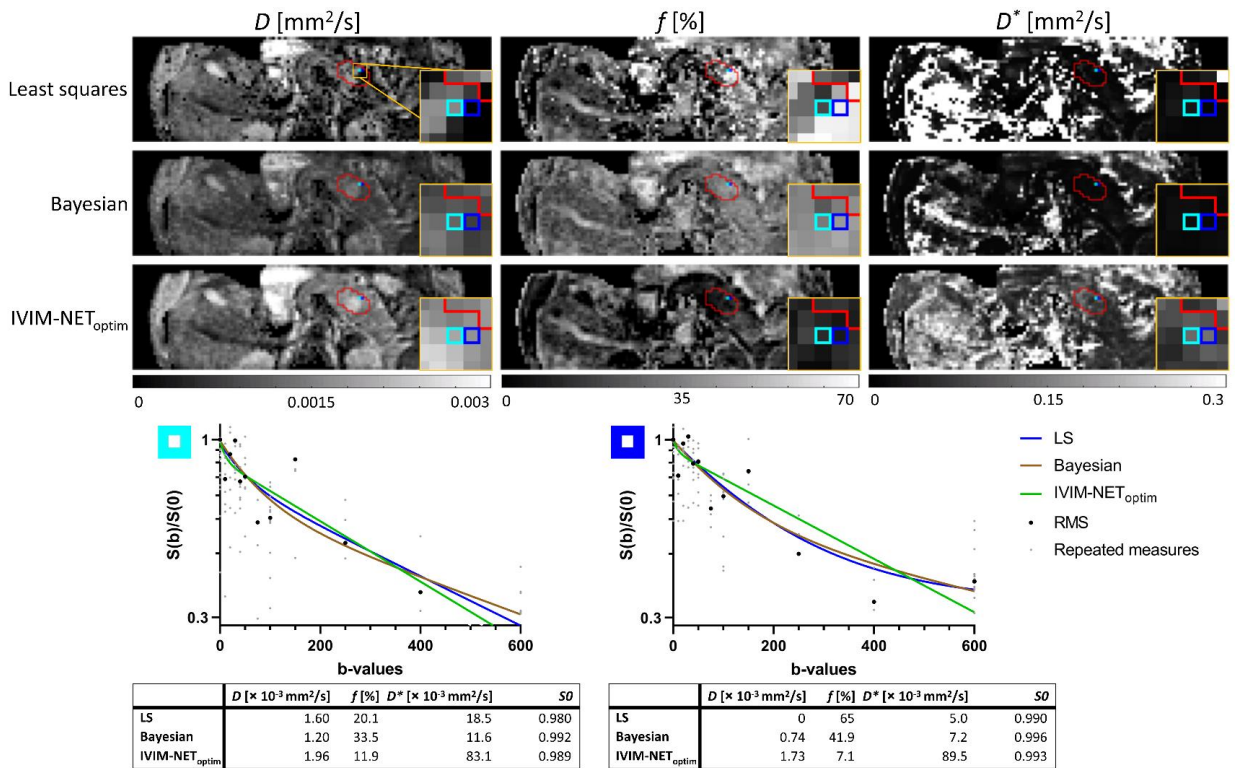


Figure S12: See Table S1.

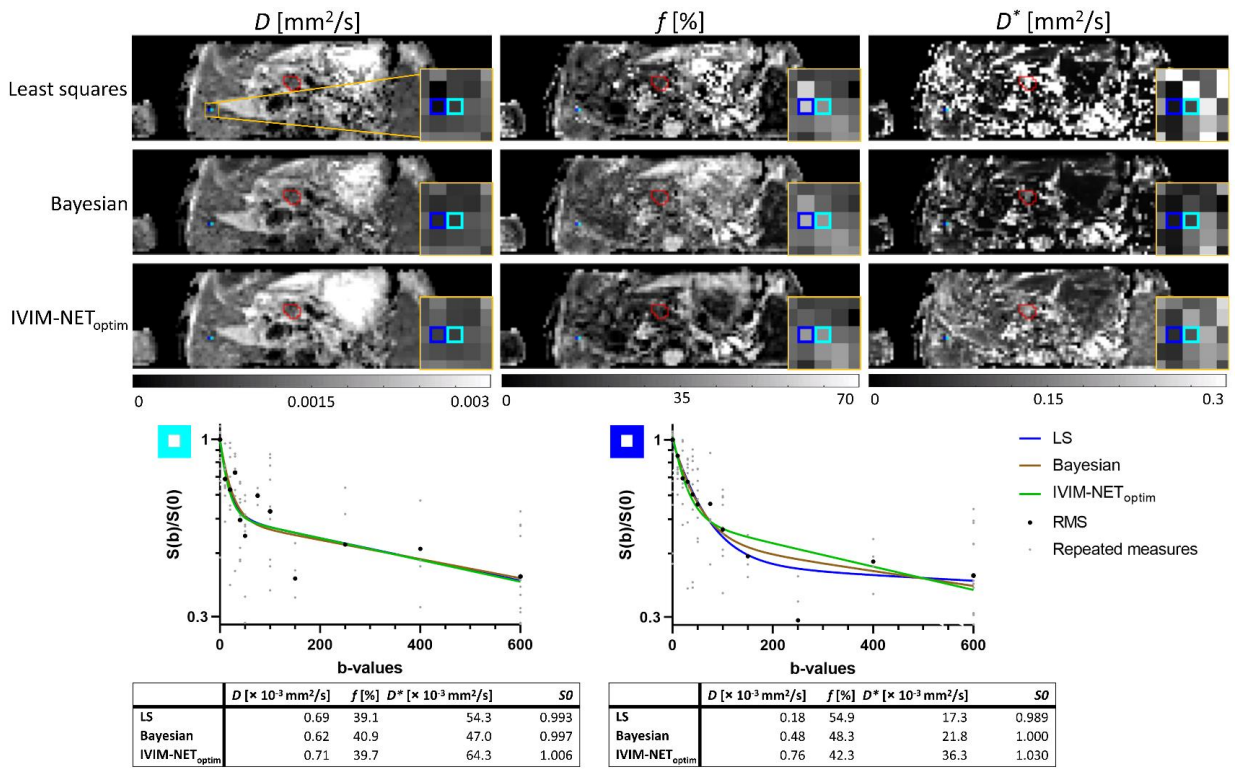


Figure S13: See Table S1.

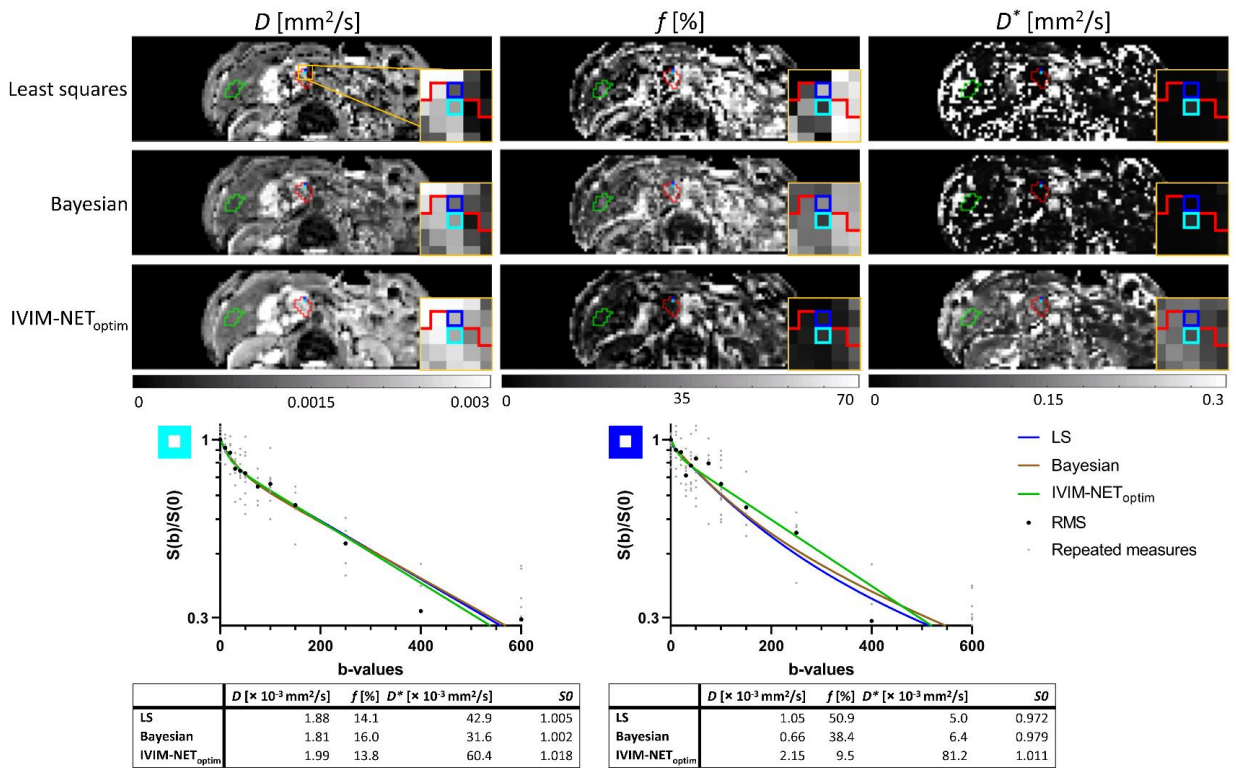


Figure S14: See Table S1.

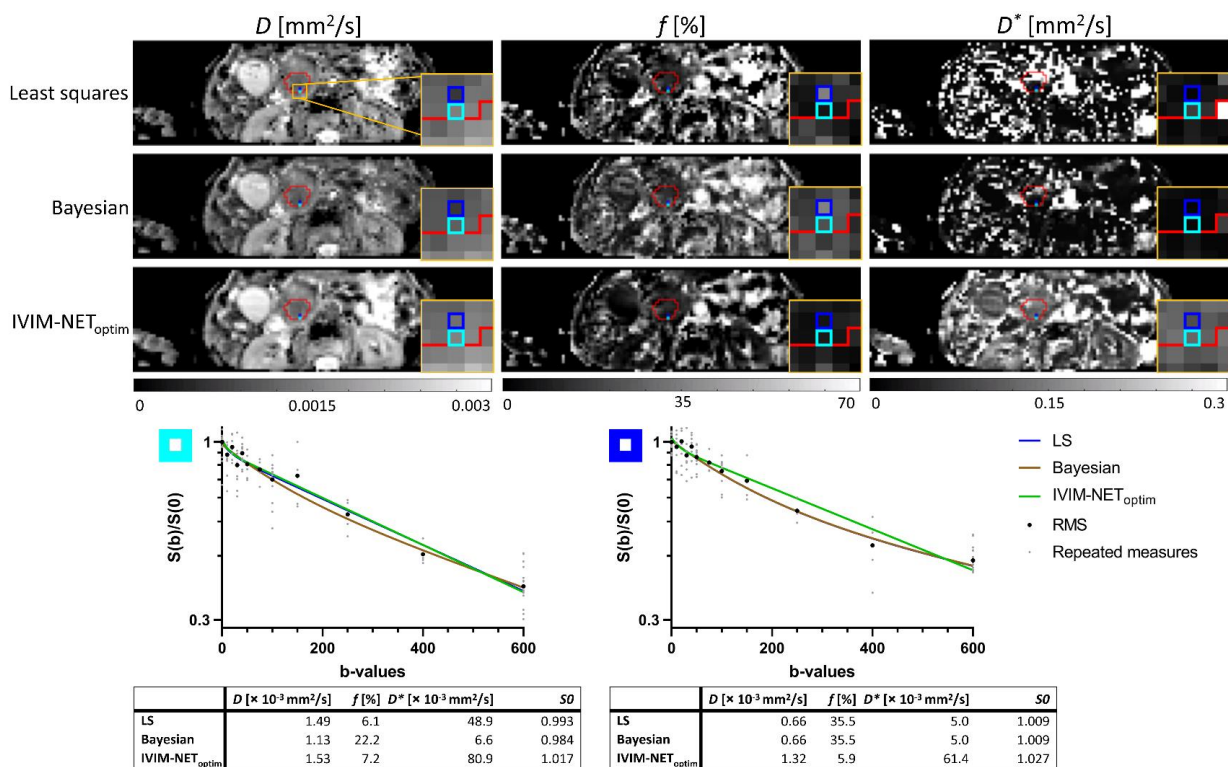


Figure S15: See Table S1.

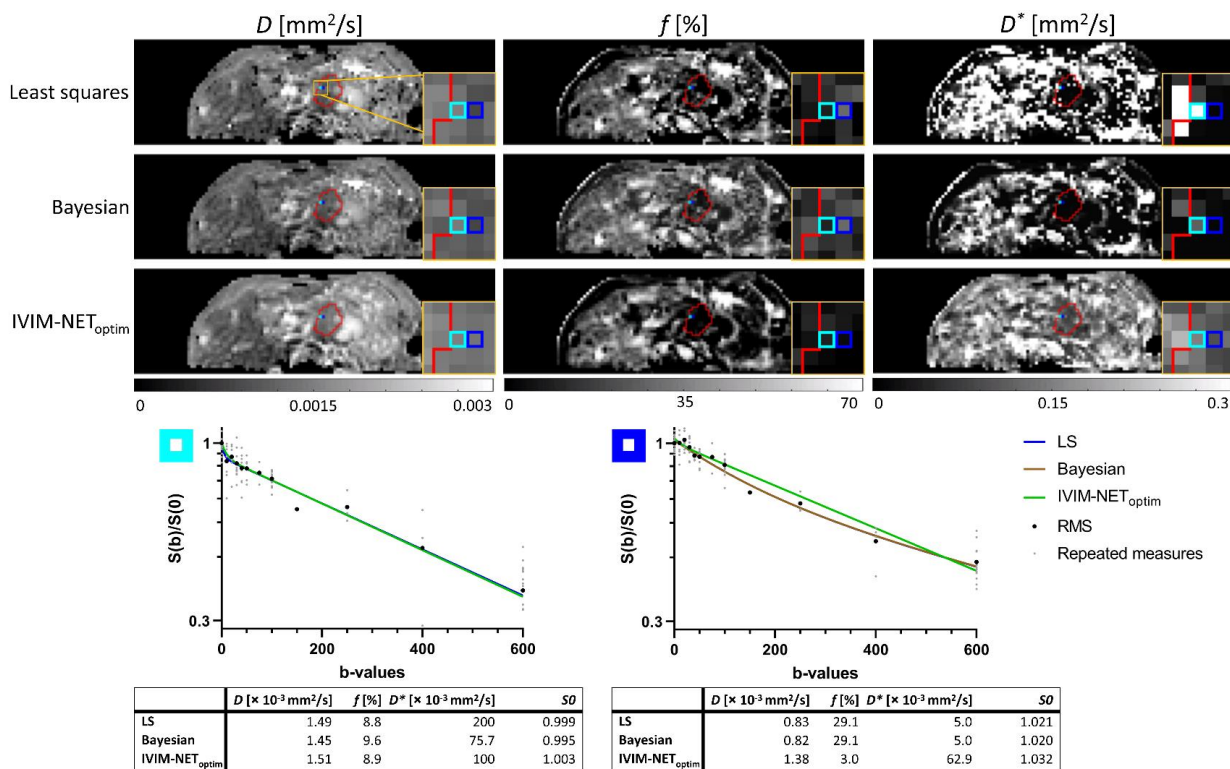


Figure S16: See Table S1.

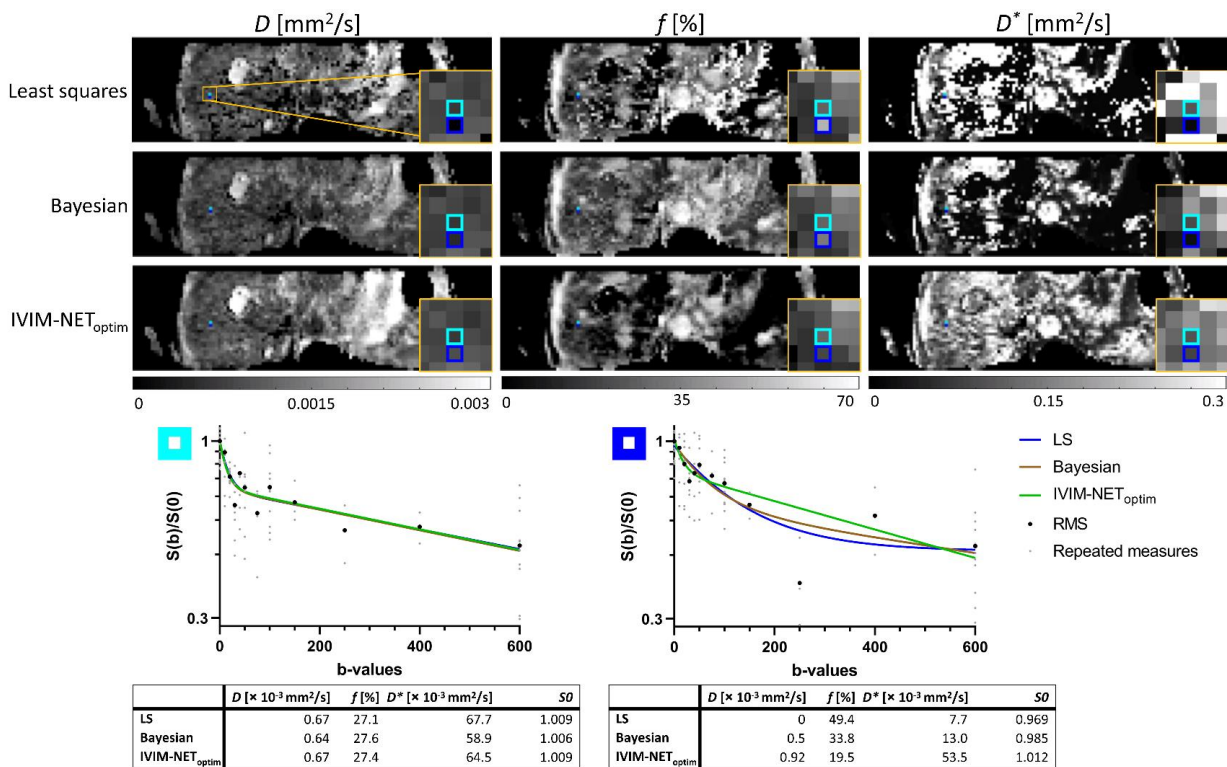


Figure S17: See table S1.

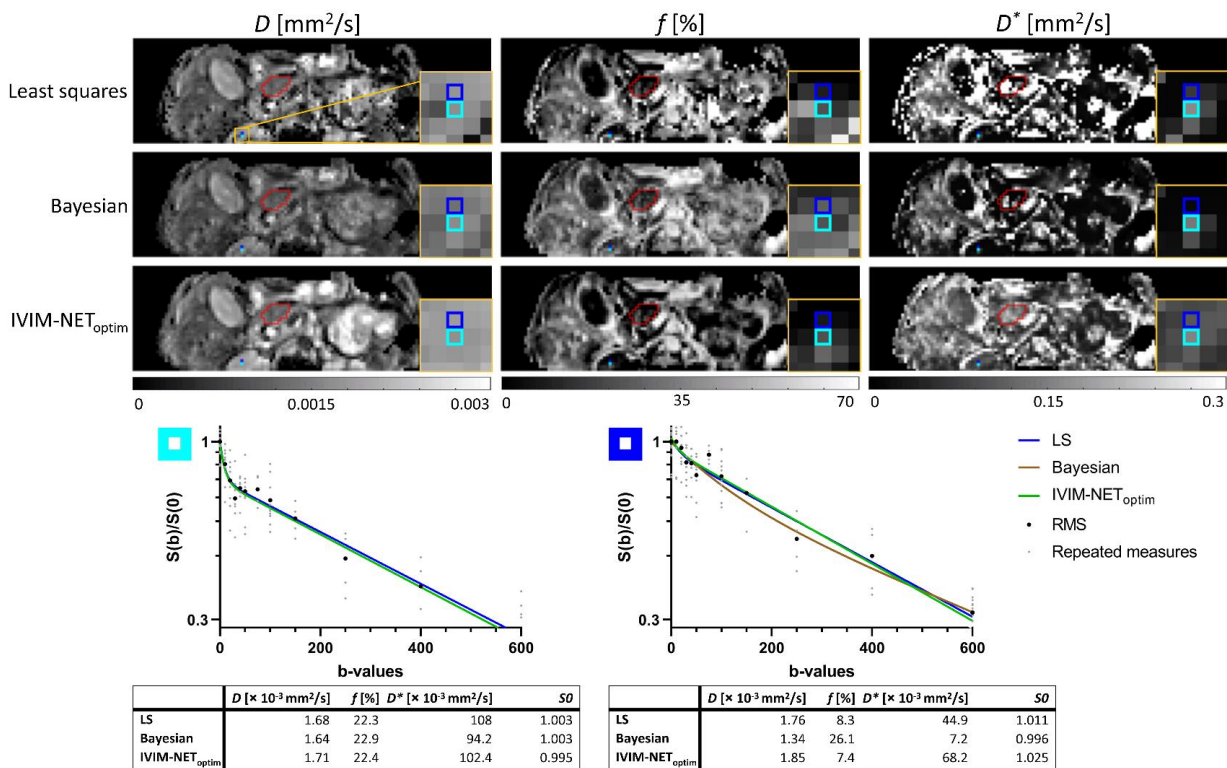


Figure S18: See Table S1.

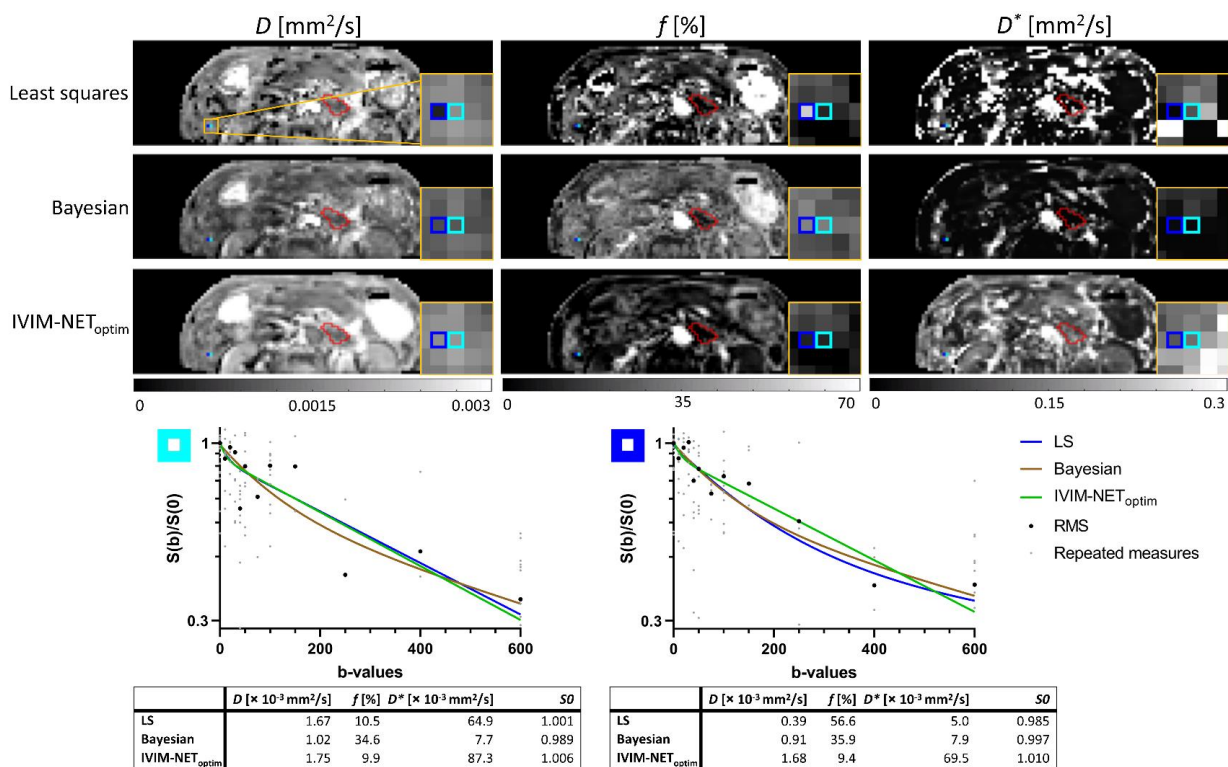


Figure S19: See Table S1.

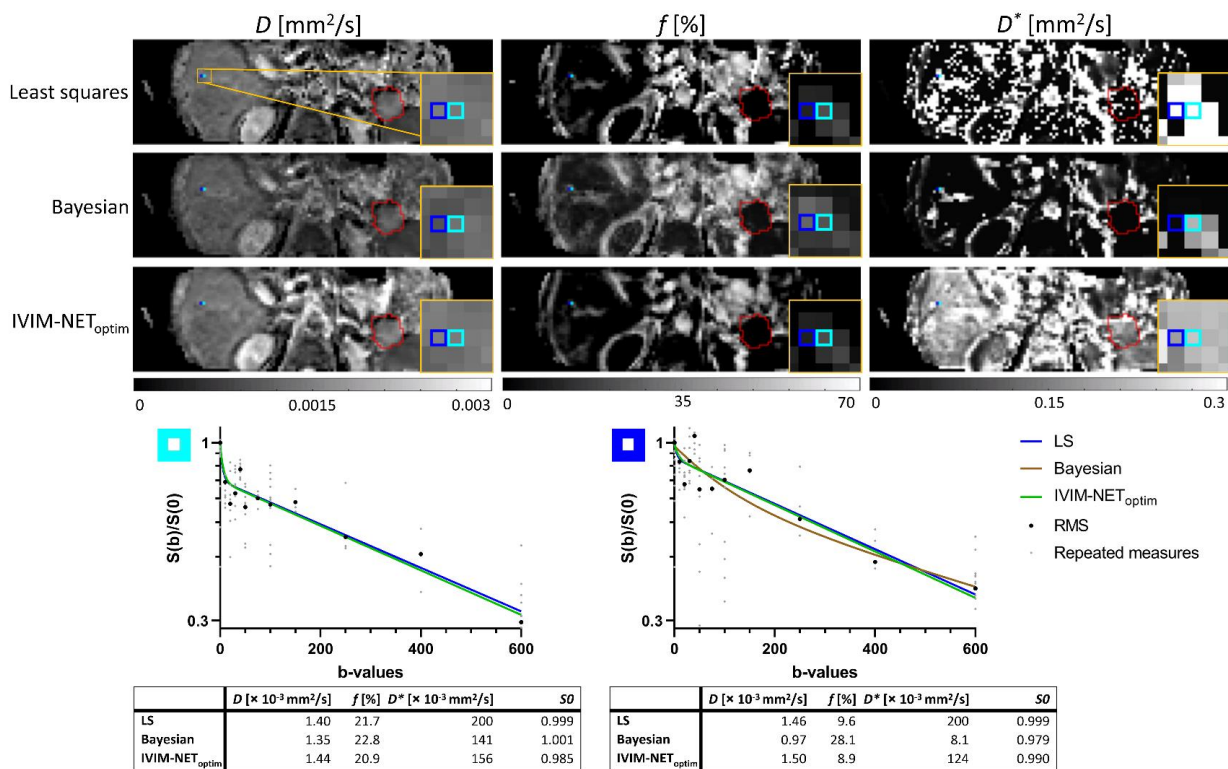


Figure S20: See Table S1.

Paper II

Deep learning intravoxel incoherent motion modeling: Exploring the impact of training features and learning strategies

Misha P. T. Kaandorp^{1,2}  | Frank Zijlstra^{1,2}  | Christian Federau^{3,4} | Peter T. While^{1,2} 

¹Department of Radiology and Nuclear Medicine, St. Olav's University Hospital, Trondheim, Norway

²Department of Circulation and Medical Imaging, NTNU – Norwegian University of Science and Technology, Trondheim, Norway

³Institute for Biomedical Engineering, University and ETH Zurich, Zurich, Switzerland

⁴AI Medical, Zurich, Switzerland

Correspondence

Misha P. T. Kaandorp, Department of Radiology and Nuclear Medicine, St. Olav's University Hospital, Trondheim, Norway.
Email: mpkaando@stud.ntnu.no

Funding information

Norges Forskningsråd, Grant/Award Number: Grant/Award Number: 302624

Purpose: The development of advanced estimators for intravoxel incoherent motion (IVIM) modeling is often motivated by a desire to produce smoother parameter maps than least squares (LSQ). Deep neural networks show promise to this end, yet performance may be conditional on a myriad of choices regarding the learning strategy. In this work, we have explored potential impacts of key training features in unsupervised and supervised learning for IVIM model fitting.

Methods: Two synthetic data sets and one in-vivo data set from glioma patients were used in training of unsupervised and supervised networks for assessing generalizability. Network stability for different learning rates and network sizes was assessed in terms of loss convergence. Accuracy, precision, and bias were assessed by comparing estimations against ground truth after using different training data (synthetic and in vivo).

Results: A high learning rate, small network size, and early stopping resulted in sub-optimal solutions and correlations in fitted IVIM parameters. Extending training beyond early stopping resolved these correlations and reduced parameter error. However, extensive training resulted in increased noise sensitivity, where unsupervised estimates displayed variability similar to LSQ. In contrast, supervised estimates demonstrated improved precision but were strongly biased toward the mean of the training distribution, resulting in relatively smooth, yet possibly deceptive parameter maps. Extensive training also reduced the impact of individual hyperparameters.

Conclusion: Voxel-wise deep learning for IVIM fitting demands sufficiently extensive training to minimize parameter correlation and bias for unsupervised learning, or demands a close correspondence between the training and test sets for supervised learning.

KEYWORDS

diffusion-weighted magnetic resonance imaging, gliomas, intravoxel incoherent motion, IVIM, supervised deep learning, unsupervised deep learning

1 | INTRODUCTION

The intravoxel incoherent motion (IVIM)¹ model for DWI shows great potential to be an alternative to dynamic susceptibility contrast (DSC) imaging for providing prognostic perfusion-based cancer imaging biomarkers, which can be used to characterize tumors and monitor treatment response without the use of a contrast agent.²⁻⁴ The IVIM model describes the attenuation of the MRI signal as a bi-exponential function of the diffusion weighting (b value). The IVIM model parameters of interest include a diffusion coefficient (D) associated with the slow diffusion of water molecules, a pseudo-diffusion coefficient (D^*) which is linked to the fast diffusion effect caused by capillary microcirculation, and a perfusion fraction (f) representing the signal contribution from the pseudo-diffusion component. The IVIM model has shown promise in several anatomies,^{3,5-7} including the brain.^{2,8,9} However, fitting IVIM parameters remains challenging in the in-vivo brain due to low SNR and low f .^{2,9,10}

Several data-processing algorithms have been proposed to improve the IVIM parameter estimation quality.¹¹ New IVIM fitting approaches are typically compared to conventional non-linear least squares (LSQ), or a two-step segmented approach, which has been shown in some studies to be less sensitive to noise than LSQ, particularly for the pseudo-diffusion estimates.^{10,12} More advanced Bayesian inference algorithms have been proposed, demonstrating improved precision and accuracy.^{13,14} However, these Bayesian approaches are typically very time-consuming, and may lead to biased parameter estimates or features disappearing due to over-smoothing.¹⁵

Recently, deep learning or deep neural networks (DNNs) were introduced as a promising alternative for IVIM fitting. These DNNs can either be trained unsupervised, where the network is optimized on the error between the estimated IVIM signal and the measured signal; or trained supervised, where the network is optimized on the error between the estimated parameters and the ground truth parameters. The approach of unsupervised learning, originally proposed by Barbieri et al.,¹⁶ allows the DNN to be trained directly on the in vivo signals, without the need for labeled data. Kaandorp et al.¹⁷ observed unexpected parameter correlations with this unsupervised network, but resolved these by optimizing various hyperparameters (IVIM-NET_{optim}). However, although IVIM-NET_{optim} showed promising results in the pancreas,¹⁷ applying it to brain data showed poor anatomy generalization and high D^* values.¹⁸

In contrast, the approach of supervised learning for IVIM fitting, originally proposed by Bertleff et al.,¹⁹ carries the assumption that the in vivo test data is well

represented by the training data, which could limit the network's performance in terms of generalizability. Gyori et al.²⁰ performed an in-depth investigation of the supervised approach for simple two-compartment and three-compartment diffusion models based on the spherical mean technique, and showed that these supervised approaches produce smooth parameter maps that are significantly influenced by the training distribution, which could possibly lead to deceptive parameter maps.

Based on these observations, we hypothesize that the performance of both unsupervised and supervised approaches for IVIM fitting is strongly influenced by the choice of both training data and network hyperparameters; hence, we identify the need to test and compare the limitations of both approaches. In this work, we explore the impact of learning rate, network size, training length, and training distribution on the convergence behavior of the unsupervised and supervised DNN loss terms, and on the accuracy of the parameter estimates. We also explore the impact of a suite of other hyperparameters on the final loss terms. We demonstrate the possible pitfalls associated with early stopping, training length, and different training data distributions, using both simulations and in-vivo data from glioma patients.

2 | METHODS

2.1 | In vivo data and preprocessing

We analyzed in vivo data from 28 glioma patients, which had been acquired and used in previous studies.^{2,18,21} Those studies were approved by the local ethics committee at the University of Lausanne, and patient consent was waived. Data were acquired using a 3T MRI scanner (Trio, Verio, or Skyra; Siemens, Erlangen, Germany). A Stejskal-Tanner diffusion-weighted spin-echo EPI pulse sequence was used with 16 b values: 0, 10, 20, 40, 80, 110, 140, 170, 200, 300, 400, 500, 600, 700, 800, 900 mm²/s in three orthogonal directions, and the corresponding trace was calculated (no averaging). A total of 20 patients were included from one study² with scan parameters: TR = 4000 ms, TE = 99 ms, FOV = 297 × 297 mm², acquisition matrix = 256 × 256, with varying slices (9-21), and receiver bandwidth = 1086 Hz/pixel. A total of eight patients were included from another study²¹ with scan parameters: TR = 4000 ms, TE = 93 ms, FOV = 270 × 270 mm², acquisition matrix = 225 × 225, with 20 slices, and receiver bandwidth = 1106 Hz/pixel. In addition, both studies used slice thickness = 4 mm, in-plane resolution of 1.2 × 1.2 mm², 75% partial Fourier encoding, parallel imaging with an acceleration factor of

2, and fat was suppressed with a spectrally selective saturation routine. The total acquisition time for both studies was 3 min and 7 s.

To pre-process the DWI data, we first removed background voxels by using both a manual threshold, as used in previous studies,^{2,21} and a Gaussian filter, which removed other background artifacts, including ghosting artifacts. Data were normalized to the SNR of the white matter (WM) of each patient, defined as the ratio of mean to standard deviation, estimated for a homogeneous region of interest (ROI), which had an SNR of approximately 30.

2.2 | Networks

We initially implemented the original DNN architecture of Barbieri et al.¹⁶ in PyTorch. The network architecture was a multi-layer perceptron with three hidden layers and exponential linear unit (ELU) activation functions.²² The network input consisted of the measured DWI signal $S(b)$ at each b value, and the network output consisted of the three IVIM parameters plus an extra parameter S_0 , the signal at $b = 0$. These parameters were further constrained by absolute value functions to avoid the possibility of exploding gradients due to negative output, and scaled (but not constrained) to appropriate physical ranges: $0 \leq S_0 \leq 1$, $0 \leq D \leq 3 \times 10^{-3}$ mm²/s, $0 \leq f \leq 50\%$, and $3 \times 10^{-3} \leq D^* \leq 100 \times 10^{-3}$ mm²/s (corresponding to network output 0–1). The upper scaling value of D is approximately equal to the self-diffusion coefficient of free water.²³

The networks were trained either unsupervised or supervised. The unsupervised networks were trained using a loss term equal to the mean-squared error (MSE) between the input signal $S(b)$ and the estimated IVIM signal $S_{\text{net}}(b)$, denoted in this manuscript as “signals-MSE”:

$$\text{signals-MSE} = \mathcal{L}(S(b), S_{\text{net}}(b)) = \sum_{b \in B} \|S(b) - S_{\text{net}}(b)\|^2 \quad (1)$$

where

$$S_{\text{net}}(b) = S_0 (f e^{-bD^*} + (1-f)e^{-bD}) \quad (2)$$

and B is the set of b values. The supervised networks were trained using a loss term equal to the MSE between the output parameters $\hat{\theta}_{\text{net}}$ of the network and the normalized IVIM parameters $\hat{\theta}$ (scaled between 0 and 1), denoted in this manuscript as “parameters-MSE”:

$$\text{parameters-MSE} = \mathcal{L}(\hat{\theta}, \hat{\theta}_{\text{net}}) = \sum_{S_0, D, F, D^*} \|\hat{\theta} - \hat{\theta}_{\text{net}}\|^2. \quad (3)$$

2.3 | Training and test data

In addition to the in vivo data described in Section 2.1, two synthetic data sets were considered for the training and testing of the networks. For the synthetic data, parameter combinations were drawn from the two distributions described below. IVIM signals were simulated using Eq. 2 for the same set of b values as the in vivo data. Rician noise was added to the signals such that when $S_0 = 1$ the SNR was 200. The following three data sets were used for training and testing:

- 1 Uniform distribution (synthetic): IVIM signals were simulated uniformly with the following parameter ranges: $0 \leq S_0 \leq 1$, $0 \leq D \leq 3 \times 10^{-3}$ mm²/s, $0 \leq f \leq 50\%$, and $3 \times 10^{-3} \leq D^* \leq 100 \times 10^{-3}$ mm²/s (i.e., identical to the scaling range of the output of the network). The uniform test set consisted of 100 000 randomly generated IVIM signals.
- 2 Patient distribution (synthetic): To provide a realistic patient distribution, we sampled IVIM parameter combinations obtained from conventional model fitting applied to the in vivo data. We performed a segmented fit using the Levenberg–Marquardt algorithm for each patient, where we first fitted D for $b > 200$ s/mm² and afterward fitted f and D^* using all b values while fixing D . S_0 was derived from the normalized $b = 0$ image. The upper bound for the segmented fit of f was 100%, while the upper bounds of S_0 , D and D^* were unconstrained. All parameters were constrained to a lower bound of 0. We then simulated IVIM signals using the parameter combinations obtained from the segmented fit. We only included parameter combinations for which $D^* \leq 100 \times 10^{-3}$ mm²/s. Segmented fitting was chosen over LSQ because it was found to be more robust to noise.
- 3 In vivo data from glioma patients, as described in Section 2.1.

For the patient-derived data ((2) and (3) above), we used 20 subjects for training and 8 subjects for testing. Figure 1 displays the parameter distributions for the synthetic test sets, which illustrates that the range and weighting of parameters differed considerably between the two synthetic test sets. Training was performed with 500 batches per epoch and batch size 128 using an Adam optimizer,²⁴ similar to previous work,¹⁷ for 50 000 epochs. Note that alternative hyperparameters were also considered, as described in Section 2.4.2 below. Training and testing were performed on a single core of an Intel Xeon Gold 6226R CPU, with each training of 50 000 taking approximately 16.5 h (1.18 s per epoch).

Normalized distributions of test sets

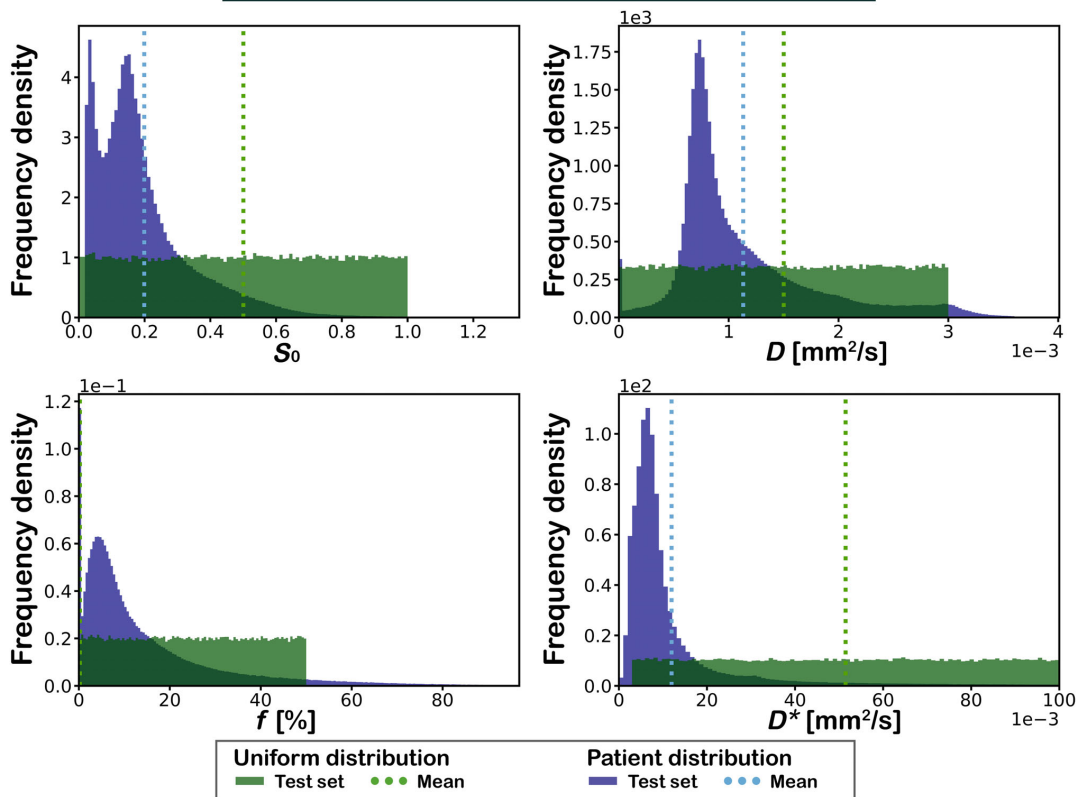


FIGURE 1 Normalized distributions for each IVIM parameter for the uniform distribution synthetic test set and the patient distribution synthetic test set. The histograms were normalized so that the area under the curve is 1 (i.e., density curve). For S_0 , D , and f , the upper 0.1% of the patient distribution test set is not displayed for visualization purposes. The dotted lines represent the parameter means for each test set.

2.4 | Evaluation

2.4.1 | Simulations: Learning rate and network size

We tested the impact of the learning rate and network size for the different strategies of unsupervised and supervised learning using the uniform distribution synthetic data set. We compared initially the performance of four networks by considering two different numbers of hidden units (16, 64) and two different learning rates (1×10^{-3} , 1×10^{-4}). For each network, we computed the signals-MSE (unsupervised loss), parameters-MSE (supervised loss) and Spearman's correlation coefficient between the pseudo-diffusion parameters ($\rho(D^*, f)$) at the end of each epoch. Previous work¹⁷ demonstrated that DNNs may predict erroneous correlations between the pseudo-diffusion parameters; hence, it is important to test

for correlations in assessing the performance of the networks. The stability of the networks was assessed in terms of the convergence of signals-MSE or parameters-MSE on the test set, and the hyperparameters corresponding to the most stable network were used in all subsequent analyses. For the most stable network, we evaluated the suitability of the early-stopping criterion used in previous approaches^{16,17} which is triggered when the test loss does not improve over 10 epochs.

2.4.2 | Simulations: Hyperparameters

We also investigated the influence of several other hyperparameters on the loss values at the early-stopping point and at the last epoch (Epoch-50000). We trained and tested networks that differed by one hyperparameter with respect to the baseline network that had been declared the most

stable (Section 2.4.1; see also Section 3.1). Here we assessed both learning strategies (unsupervised, supervised) using the uniform distribution synthetic data set. The different hyperparameters that were examined included (baseline in square parentheses): batch size ([128], 256, 512), activation function ([ELU], ReLu, sigmoid), number of hidden units (16, [64], 128), batch normalization ([without], with), dropout ([0%], 10%, 20%), number of hidden layers (2, [3], 4, 5, 6), optimizer ([Adam], SGD, RMSProp), and learning rate (1×10^{-3} , [1×10^{-4}], 1×10^{-5}).

2.4.3 | Simulations: Learning strategy and training distribution

Using the hyperparameters corresponding to the most stable network (Section 2.4.1), four further networks were trained by considering the two different learning strategies (unsupervised, supervised) and the two synthetic data sets (uniform distribution, patient distribution). These four networks were considered in all subsequent analyses. These networks were tested on the patient distribution test set to assess generalizability. For each network, we computed the signals-MSE, parameters-MSE, Spearman's $\rho(D^*, f)$, and normalized MSE per parameter at the end of each epoch.

2.4.4 | Simulations: Training length

The performance of the four networks was further evaluated at different validation points during training by comparing the estimated parameters with the ground truth for individual data points of the uniform distribution test set. To enable further qualitative assessment, parameter maps and root-mean-square error (RMSE) maps were generated for each network from the synthetic patient data. Here, the RMSE was calculated between the expected DWI signal for the estimated parameters and the ground truth DWI signal. Four different validation points were used: (i) early stopping; (ii) when D^* -MSE was at a minimum (Min(D^* -MSE)); (iii) Epoch-5000; (iv) the last epoch (Epoch-50000), representing extensive training. Comparisons were also made to parameter maps estimated by LSQ using the Levenberg–Marquardt algorithm, and a segmented approach as described in Section 2.3.

2.4.5 | In vivo

To demonstrate that our simulations are comparable to real in vivo data, we evaluated the performance of the four networks (i.e., trained on synthetic data) on the in vivo

glioma patient data, together with an additional unsupervised network that was directly trained on the in vivo signals. We computed the signals-MSE and $\rho(D^*, f)$ at the end of each epoch. In order to provide a fair assessment of the bulk of the data, we performed a correction for outliers by excluding voxels lying within the top 0.7% of signals-MSE, chosen to mimic classical outlier detection.²⁵ These voxels contained signals that were poorly represented by a bi-exponential function, and therefore represented signal behavior not seen by the four networks during training.

Performance was also assessed qualitatively for the in vivo glioma patient data by generating parameter maps and RMSE maps at the early-stopping point, Epoch-5000 and Epoch-50000. These maps were compared to parameter maps estimated by LSQ, the segmented approach, the four networks from Section 2.4.3 at Epoch-50000, and IVIM-NET_{optim}.¹⁷ IVIM-NET_{optim} was applied as in the original article,¹⁷ with the exception that the range of the S_0 parameter was 0–2 instead of 0.7–1.3 to encompass the signal scaling in this study.

An additional important consideration is to assess the impact that the individual fitting approaches may have on the quantitative values of each parameter for certain tissue types of interest. Hence, for each IVIM parameter, we pooled the voxel-wise estimates for in vivo tumor regions of all patients (one patient was excluded as no tumor mask was available) and compared the distribution of parameter estimates for each fitting approach at the early-stopping point and at Epoch-50000.

3 | RESULTS

3.1 | Simulations: Learning rate and network size

Figure 2 shows the loss curves of the first 5000 epochs for both the unsupervised and supervised learning strategies trained and tested on the uniform distribution using different network size and learning rate. For both learning strategies, using fewer hidden units reduced convergence speed, whereas higher learning rate resulted in spiky convergence of the loss term, which could result in sub-optimal solutions, particularly in unsupervised learning. Therefore, the network containing hidden units = 64 and learning rate = 1×10^{-4} was considered the most stable in both strategies, and was used in all subsequent analyses.

Early stopping (patience = 10 epochs) resulted in sub-optimal solutions prior to true convergence, at epoch 237 for unsupervised learning and epoch 118 for supervised learning (Figure 2). Furthermore, the

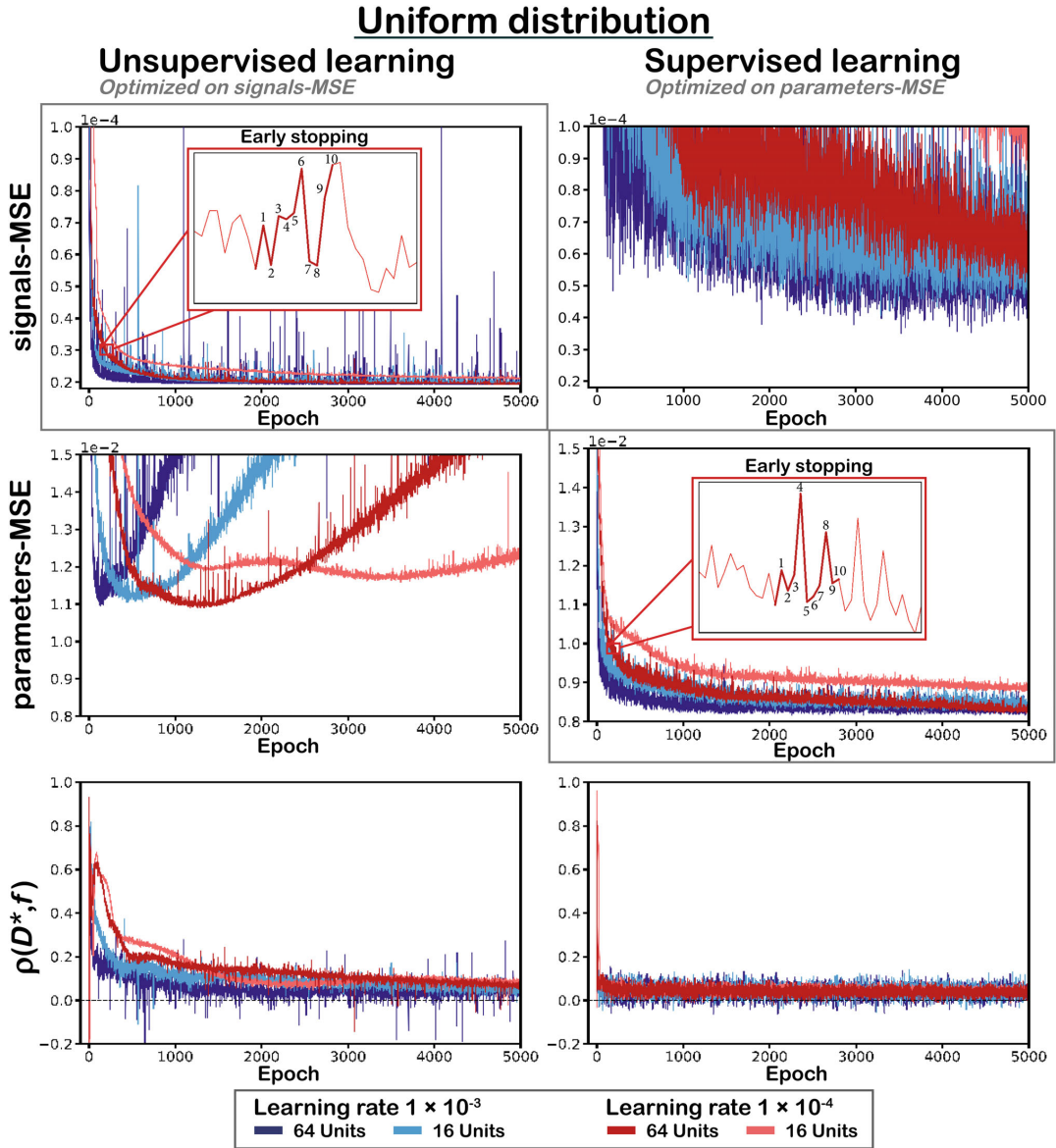


FIGURE 2 Test curves showing the metrics signals-MSE, parameters-MSE, and Spearman's $\rho(D^*,f)$ for unsupervised (left) and supervised (right) networks with different number of hidden units (16, 64) and learning rate (1×10^{-3} , 1×10^{-4}), trained and tested on synthetic data from the uniform distribution, and evaluated over 5000 epochs. For the most stable network, the red square highlights where the test loss did not improve over 10 consecutive epochs and the early-stopping criterion used in prior work^{16,17} would have ended the training, which results in sub-optimal convergence, as well as correlated parameters for the unsupervised case at epoch 223 ($\rho(D^*,f) = 0.62$).

pseudo-diffusion parameters were correlated for unsupervised learning ($\rho(D^*,f) = 0.62$). Extending training resolved these correlations and reduced parameters-MSE. However, training substantially longer resulted in increased parameters-MSE. For supervised learning, both

parameters-MSE and signals-MSE decreased with longer training length. As expected, unsupervised learning yielded lower signals-MSE (unsupervised loss), whereas supervised learning yielded lower parameters-MSE (supervised loss) in all cases.

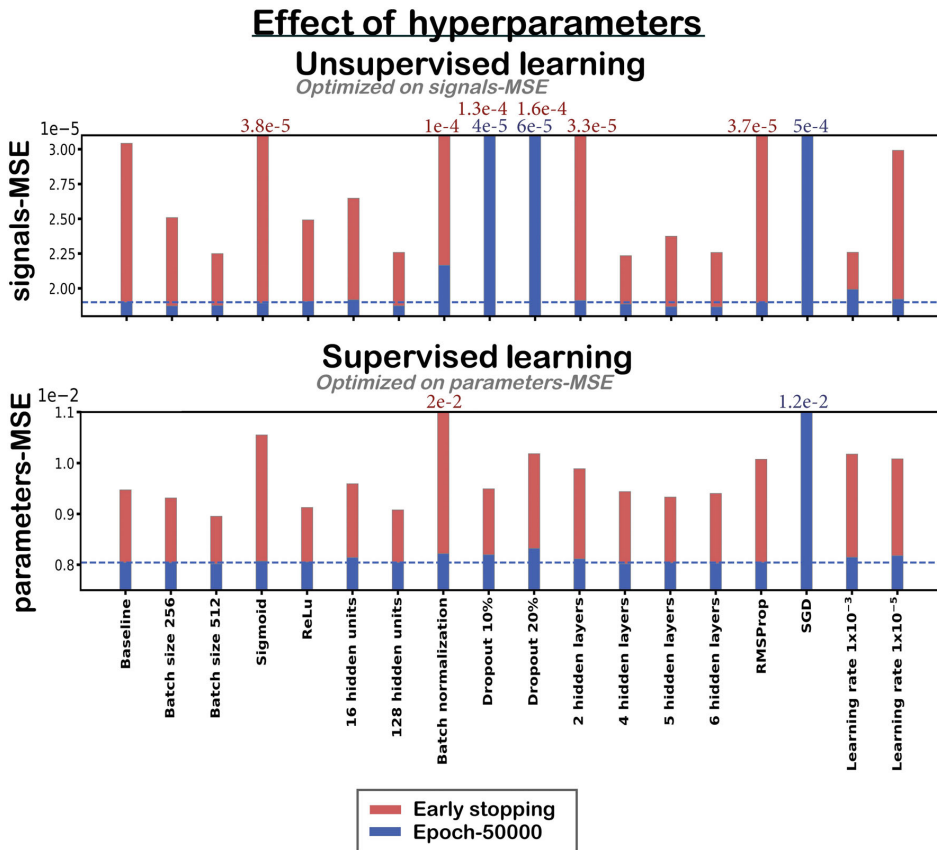


FIGURE 3 Performance of the networks that differed by one hyperparameter with respect to the baseline network that had been declared the most stable in the previous analysis in Figure 2 (see Sections 2.4.1 and 3.1). Results are shown for both learning strategies (unsupervised and supervised) when trained on the uniform distribution and evaluated at the early-stopping point and after training for 50 000 epochs (Epoch-50000). Details about the hyperparameters considered and the baseline set can be found in Section 2.4.2. The blue dashed line represents the loss for the baseline network at Epoch-50000. The values printed above the plots are the loss values at the early-stopping point (red) and at Epoch-50000 (blue) that could not be displayed within the chosen range of the plot. The SGD optimizer had no early-stopping point.

3.2 | Simulations: Hyperparameters

Figure 3 displays loss values at two different validation points during training for the networks that differed by one hyperparameter from the baseline network determined in Section 3.1. Every network performed substantially better when training was extended beyond early stopping. The networks with greater learning capacity (i.e., more hidden layers or more hidden units) showed a marginal improvement in the final loss value. Adding batch normalization resulted in a marginally higher loss. Adding dropout resulted in a substantially higher loss for unsupervised learning, whereas for supervised learning the loss was only marginally higher. Increasing batch size resulted

in a marginally lower loss. The choice of activation function had little impact on the loss after extensive training, whereas the choice of optimizer may be an important consideration given the high loss for SGD. In general, the variability in loss values between networks was far greater at the early-stopping point than after extensive training.

3.3 | Simulations: Learning strategy and training distribution

In Figure 4, we show the loss curves of the four networks, as described in Section 2.4.3, tested on the patient distribution test set. Figure 4A shows that for both

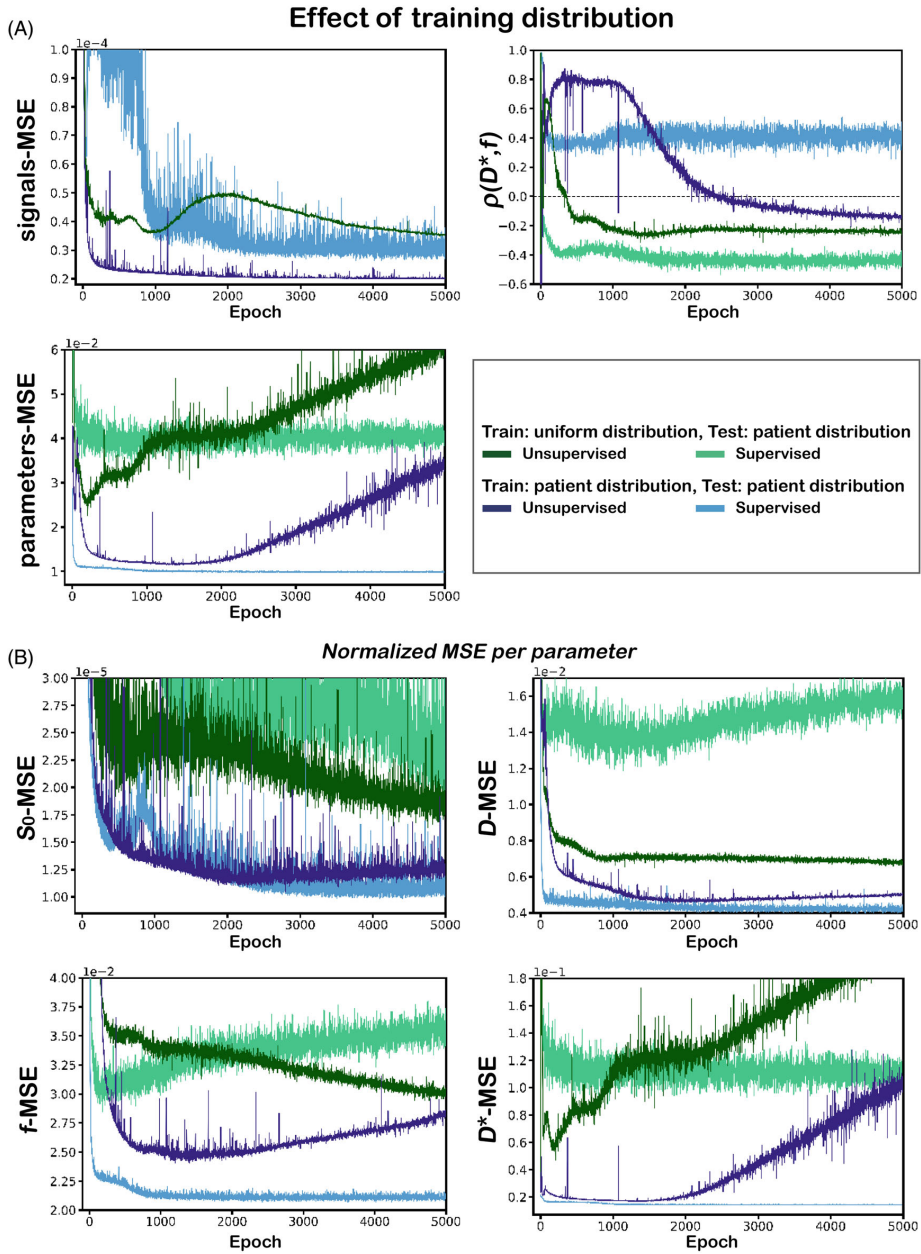


FIGURE 4 (A) Test curves showing the metrics signals-MSE, parameters-MSE, and Spearman's $\rho(D^*, f)$ for the four networks described in Section 2.4.3, that trained either unsupervised or supervised (hidden units = 64 and learning rate = 1×10^{-4}) on synthetic data from either the uniform distribution or patient distribution, and tested on the patient distribution test set over 5000 epochs. (B) Test curves showing the normalized MSE per parameter for each network.

unsupervised and supervised learning, training and testing on the same distribution yields lower loss and faster convergence compared to training and testing on different distributions. For the unsupervised networks, training on the uniform distribution (i.e., different to test set) shows an increase in test loss (signals-MSE) as training is extended (e.g., between epoch 1000 and 2000). Moreover, for unsupervised learning, the undesirable parameter correlation persists longer when training on the patient distribution compared to training on the uniform distribution. Here, early stopping occurred at epoch 181 with $\rho(D^*, f) = 0.73$ when trained on the patient distribution. The increase in parameters-MSE observed in unsupervised learning is mainly due to an increase in errors in D^* , as evidenced by the similarity in plots between D^* -MSE (Figure 4B) and parameters-MSE (Figure 4A), although increases in errors for S_0 , D , and f also occur. Note that $\text{Min}(D^*\text{-MSE})$ occurs at epoch 208 when training on the uniform distribution and epoch 1552 for the patient distribution. Conversely, in supervised learning, the signals-MSE, parameters-MSE, and $\rho(D^*, f)$ gradually decrease with training length, but parameters-MSE is substantially greater when the testing and training distributions are different.

3.4 | Simulations: Training length

Figure 5 displays scatter plots comparing the estimations with the known ground truth values for the four networks at different validation points during training, when applied to the uniform distribution test set. This figure clearly shows the impact of early stopping, especially when training on the patient distribution, which results in poor accuracy. Early in training, both strategies estimate parameters that are apparently biased toward the mean of the training distributions (see Figure 1), particularly for D^* . As training progresses in unsupervised learning, the estimates corresponding to low S_0 signals (i.e., low SNR) exhibit higher variability and display a distribution tending toward that of LSQ. Conversely, as training progresses in supervised learning, the bias toward the mean of the distributions persists for a greater proportion of data points. These data correspond to either low S_0 (i.e., low SNR), where a bias occurs for every IVIM parameter; low f , which results in uncertainty in D^* ; or low D^* , which results in uncertainty in f .

In addition, Figure 5 shows that early in training the parameter values that are overrepresented in the training distribution (see Figure 1) are estimated more accurately than underrepresented parameter values. Furthermore, there is poor-fitting behavior for low D values with high S_0 when training on the patient distribution. This is particularly evident for supervised learning, and is due to a lack

of these parameter combinations occurring in the patient data.

Figure 6 displays parameter maps and RMSE maps at different validation points during training of the four networks in a synthetic patient slice. At the early-stopping point, the parameter maps are relatively inaccurate, particularly for D^* when trained on the uniform distribution, and appear smooth and homogeneous. As found in Figure 5, early in training, the parameter estimates in voxels corresponding to low SNR signals (e.g., WM voxels) are biased toward the mean of the synthetic training distribution. As training progresses in unsupervised learning, the parameter estimates in WM voxels exhibit a distribution tending toward that of LSQ, which is also visualized by comparable extremities in D^* and RMSE maps, although f maps are generally more comparable to the segmented fit. There are increases in the RMSE at Epoch-50000 when training on the uniform distribution (red arrows), which are not present when training on the patient distribution.

For supervised learning on the uniform distribution, the estimates for D^* in the WM voxels are biased toward the mean of the training distribution, even after extensive training, which is consistent with Figure 5 for data with low S_0 . In addition, the voxels that represent the tumor in this synthetic patient slice (purple arrow in $b = 0$ image) possess very low f values, such that the simulated signals contain little information for estimating accurate D^* . Hence, the supervised network estimates inaccurate D^* close to the mean of this distribution, resulting in high residuals. In contrast, the supervised network trained on the patient distribution shows lower residuals, which are more comparable to the results of the unsupervised networks, because the bias toward the mean of the patient distribution yields values that are closer to the actual ground truth.

3.5 | In vivo

Figure 7 displays loss curves and parameter maps for the four networks applied to a representative slice of the in vivo glioma patient data, as well as corresponding results for the additional unsupervised network trained directly on the in vivo data. These results are in broad agreement with those displayed in Figures 4 and 6 for the simulations. Similar to Figure 4, training and testing on the same dataset yielded the lowest loss term and fastest convergence (Figure 7A). Furthermore, early stopping resulted in sub-optimal solutions at epoch 312 prior to true convergence, where parameters were strongly correlated ($\rho(D^*, f) = 0.94$). Similar to Figure 6, early stopping resulted in seemingly smooth and homogeneous parameter maps. As training was extended, parameter maps tended toward

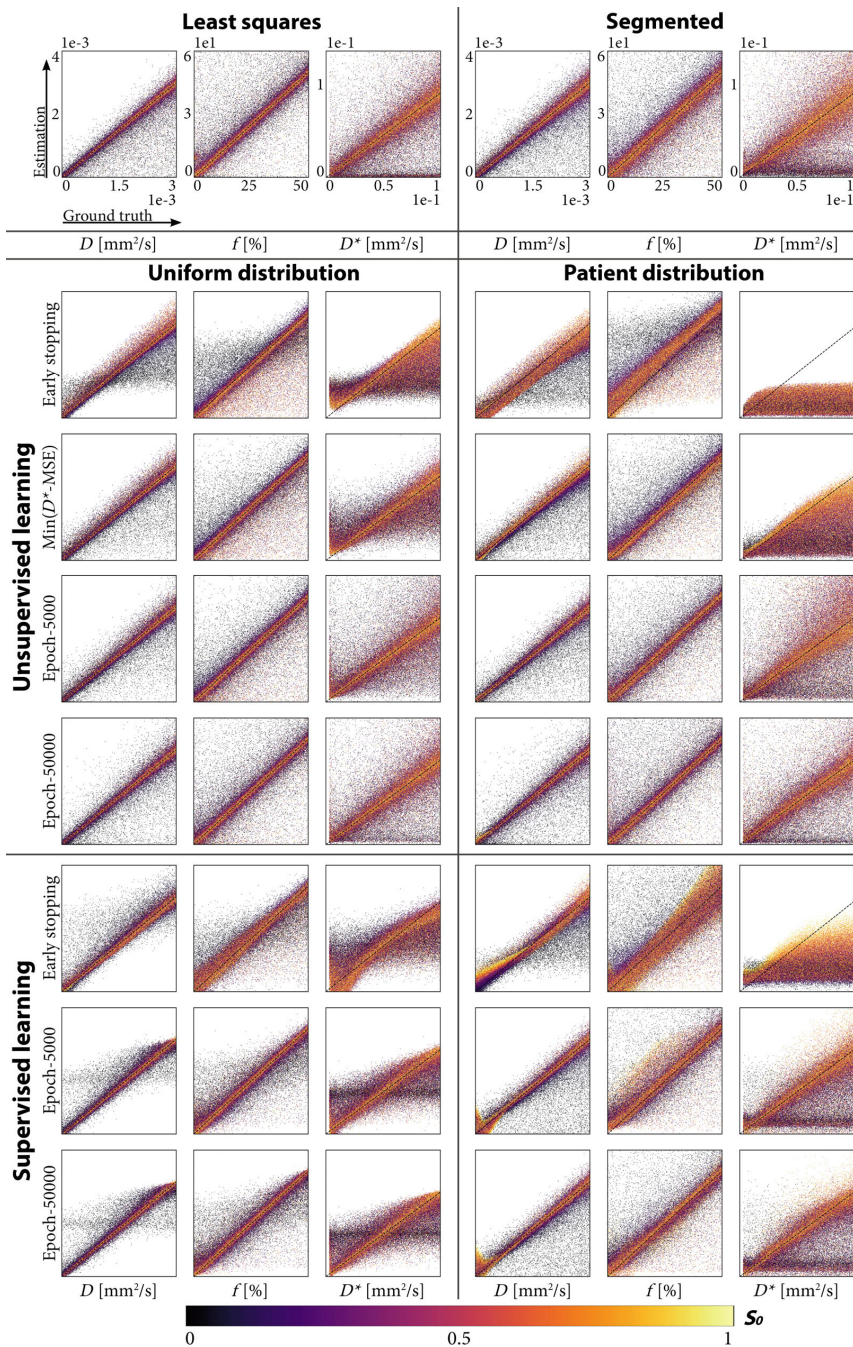


FIGURE 5 Scatter plots of estimated parameter values against ground truth at each of the four validation points (Early stopping, $\text{Min}(D^*-\text{MSE})$, Epoch-5000, and Epoch-50000) for the four networks described in Section 2.4.3 (i.e., trained either unsupervised or supervised on synthetic data from either the uniform distribution or patient distribution), and tested on the uniform distribution test set. Corresponding plots for least squares and the segmented approach are also shown. All data points are colored by their S_0 -value, where $S_0 = 0$ (black) corresponds to SNR = 0 and $S_0 = 1$ (bright yellow) corresponds to SNR = 200.

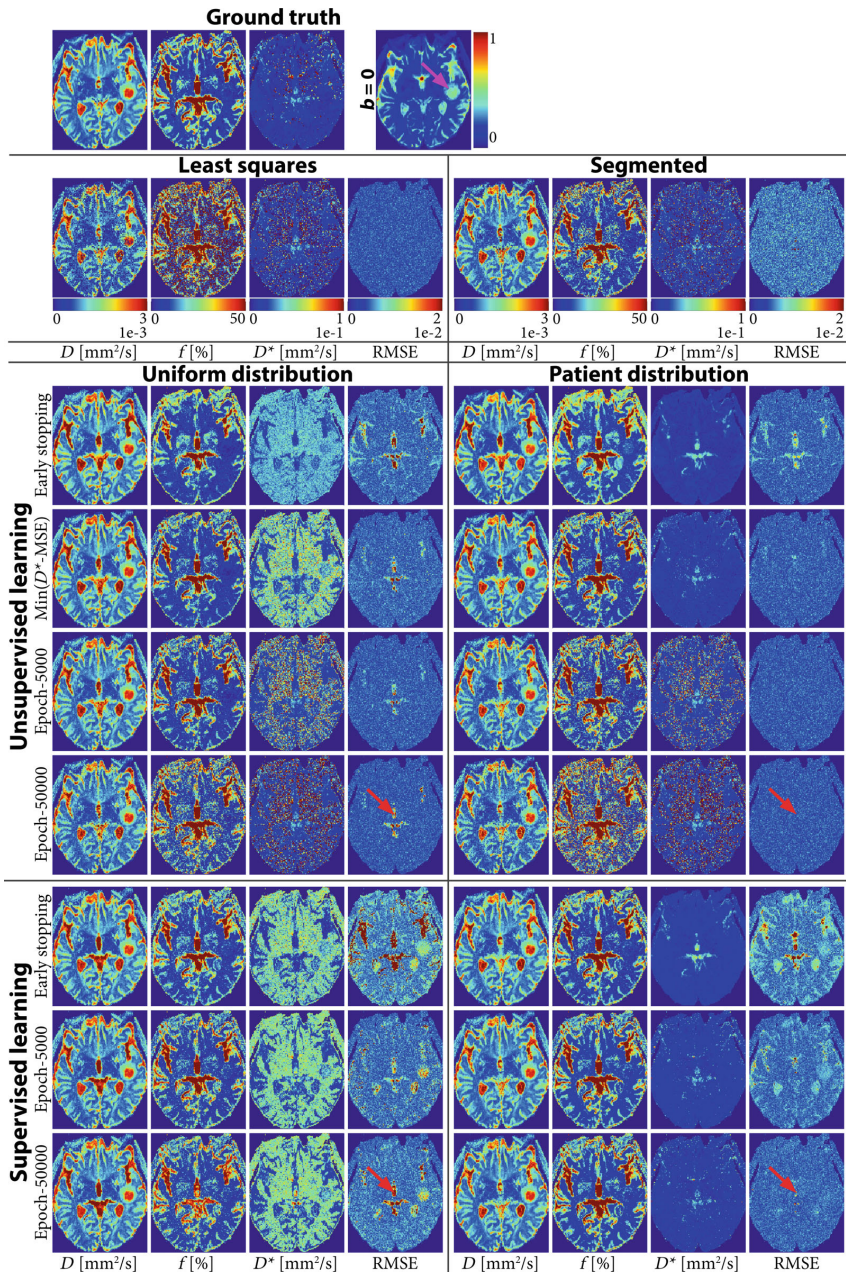


FIGURE 6 IVIM parameter maps and RMSE maps generated for a representative slice of a synthetic glioma patient. Maps are displayed at each of the four stopping points (Early stopping, $\text{Min}(D^*-\text{MSE})$, Epoch-5000, and Epoch-50000) for the four networks described in Section 2.4.3 (i.e., trained either unsupervised or supervised on synthetic data from either the uniform distribution or patient distribution). The ground truth parameter maps are generated by applying the segmented approach on the real patient data. The red arrows in the Epoch-50000 RMSE maps indicate out-of-distribution data for the uniform distribution, where $D > 3 \times 10^{-3} \text{ mm}^2/\text{s}$ or $f > 50\%$, which lie inside of the domain of the patient distribution, but outside of the domain of the uniform distribution. Corresponding maps for least squares and the segmented approach fitted on the synthetic glioma patient are also shown. The purple arrow in the ground truth $b = 0$ image indicates the location of a tumor.

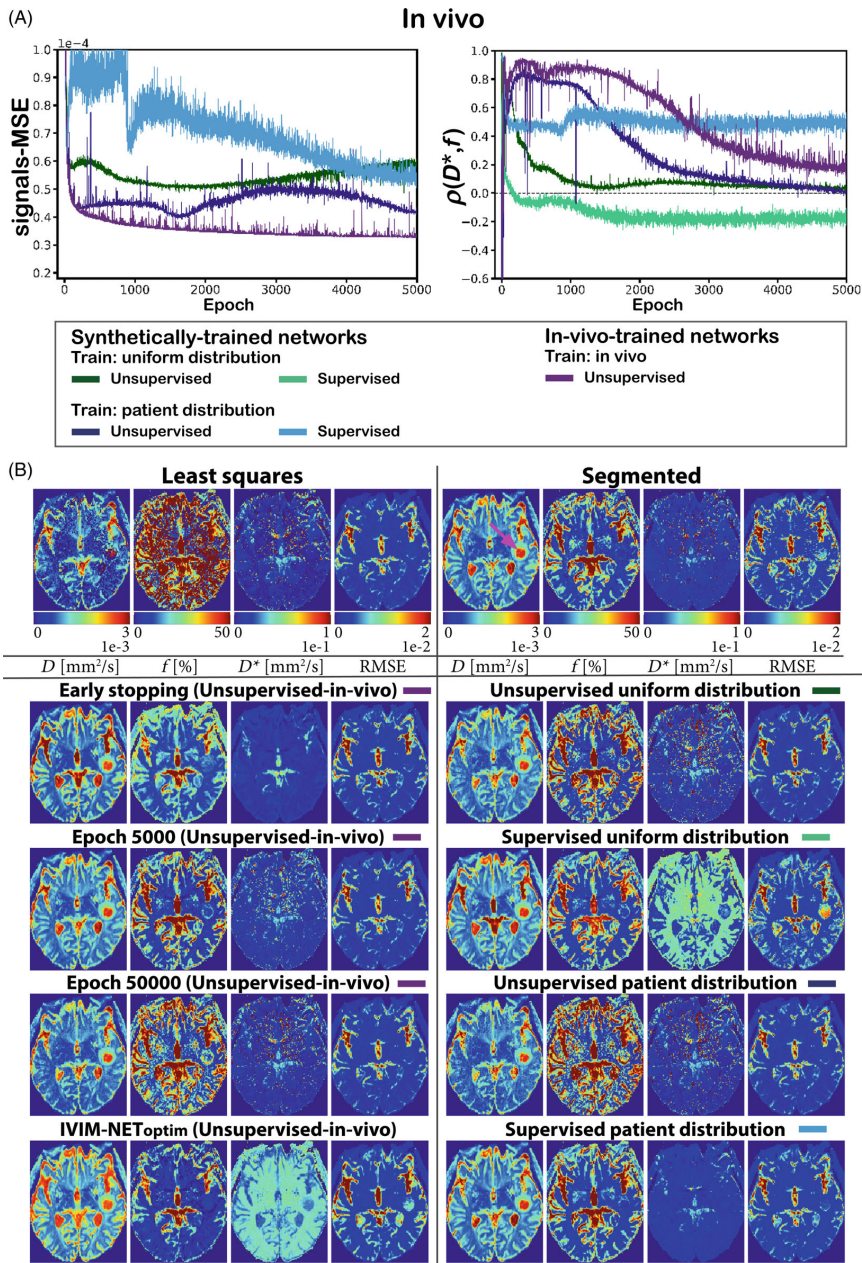


FIGURE 7 (A) Test curves showing the metrics signals-MSE and Spearman's $\rho(D^*, f)$ for the four networks described in Section 2.4.3 (i.e., trained either unsupervised or supervised on synthetic data from either the uniform distribution or patient distribution), and tested on the in vivo data over 5000 epochs. In addition, test curves for an unsupervised network trained directly on the in vivo signals are displayed. (B) IVIM parameter maps and RMSE maps for a representative slice from the in vivo glioma patient data for each of the networks described in (A). For the unsupervised network trained directly on the in vivo data (left), maps corresponding to three stopping points are displayed (Early stopping, Epoch-5000, and Epoch-50000), whereas the maps shown for the four synthetically-trained networks (right) are after training for 50 000 epochs. Corresponding maps are also shown for least squares (top left), the segmented approach (top right) and IVIM-NET_{optim} (bottom left). The purple arrow in the D map for the segmented approach indicates the location of a tumor.

IVIM estimates for the in vivo tumor

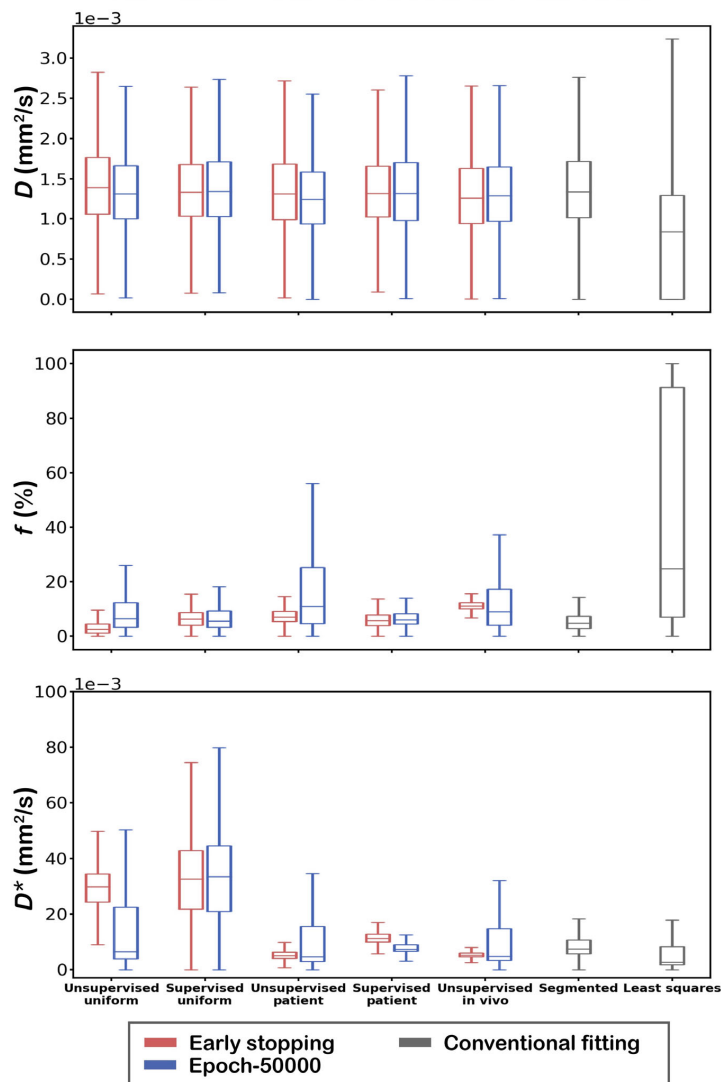


FIGURE 8 Boxplots of the voxel-wise IVIM parameter estimates for in vivo tumor regions of all patients for each of the networks described in Figure 7A, at the early-stopping point and after training for 50 000 epochs (Epoch-50000). Corresponding boxplots are also shown for least squares and the segmented approach. The outliers of the boxplots (i.e., those data exceeding 1.5 times the interquartile range above the upper quartile and below the lower quartile) are not displayed for visualization purposes.

those of LSQ, particularly for D^* . Generally, the residual maps of Figure 7B displayed higher RMSE than those of Figure 6. The parameter maps and RMSE maps of the unsupervised networks that were trained on either the uniform or patient distributions were similar to those for the unsupervised network trained directly on the in vivo data. In contrast, IVIM-NET_{optim} displayed inferior RMSE and high D^* (cf.¹⁸), and parameter estimates were also considerably different for voxels in the tumor region.

For the supervised networks, a bias toward the mean of the training distribution occurred, which resulted in high

RMSE for the network trained on the uniform distribution. In contrast, the parameter maps and RMSE maps of the supervised network trained on the patient distribution are comparable to those for the unsupervised networks. Despite the differences in parameter estimates for the different unsupervised and supervised approaches, RMSE maps were generally similar for low SNR regions (e.g., WM regions).

In Figure 8, we show boxplots of all voxel-wise IVIM parameter estimates for in vivo tumor regions of all patients for each fitting approach. For D , all of the

approaches yield a similar distribution of estimates, with the exception of LSQ, which tends to estimate lower values. Greater variability between approaches is observed for f , and especially for D^* , where at the early-stopping point, the networks trained on synthetic data (unsupervised and supervised) estimate the bulk of the parameter values toward the mean of the respective synthetic training distribution (see Figure 1). After extensive training (Epoch-50000), the unsupervised networks display broader distributions of parameter estimates, but apparently with less bias, given the observed similarity between the distributions for each network. Conversely, the bias observed for the supervised networks appears to persist even after extensive training, which is consistent with the results from the simulations. Note also in Figure 8 that LSQ yields a substantially broader distribution of f values in the tumor, which is also reflected in Figure 7 with considerably higher estimates for f , whereas the segmented approach produces a much narrower distribution of lower f values.

4 | DISCUSSION

This work highlights the impact of learning rate, network size, training length and training distribution on unsupervised and supervised DNN IVIM fitting to DWI data. In simulations, we showed that high learning rate and early stopping may lead to correlated parameter estimates and sub-optimal model fitting. We showed that a network with more hidden units increases convergence speed, a lower learning rate results in stable convergence, and extending training beyond early stopping leads to reduced parameter correlations and parameter error. However, extensive training resulted in an increased sensitivity to noise. For unsupervised learning, parameter estimates in regions with low SNR (or low f), that is, with underlying uncertainty, displayed variability similar to LSQ estimates. For supervised learning, parameter estimates with underlying uncertainty were affected strongly by the distribution of the training data, showing biases toward the mean of the training distribution, which resulted in smooth, yet possibly deceptive parameter maps. Moreover, we showed that early in training the networks were more accurate at estimating parameter combinations that were well represented in the training set, compared with underrepresented combinations, but the accuracy of the latter improved with extensive training. The in vivo results were in broad agreement with the simulations, and fitting residuals were almost identical between approaches, particularly for the unsupervised networks. However, the pseudo-diffusion maps varied considerably, demonstrating the difficulty of fitting D^* in these regions.

Previous work demonstrated that the use of DNNs for IVIM fitting can lead to erroneous correlation between estimates for the pseudo-diffusion coefficient (D^*) and the perfusion fraction (f).¹⁷ Whereas that work resolved those correlations by exhaustive hyperparameter optimization, the present work has demonstrated that extending training beyond early stopping is a simpler alternative, which provides consistent results across many hyperparameters (Figure 3). Of course, this begs the question of how much training is sufficient or optimal. In the approach of Barbieri et al.,¹⁶ the early stopping criterion resulted in sub-optimal convergence at approximately 50 epochs, whereas we trained for 50 000 epochs (i.e., 1000 times longer) to ensure adequate convergence. In the simulations, extensive training in unsupervised learning resulted in an increase in MSE per parameter. Therefore, it could be argued that the point where D^* -MSE is lowest might be a good choice of stopping point in unsupervised learning (Figure 4). However, the minimum D^* -MSE observed in the simulations is really only an artifact of the persisting bias in the parameter estimates at that point in training (Figures 5, 6), coupled with the chosen metric for gauging accuracy. For unsupervised learning, extensive training was found to ameliorate this bias and also further reduce the fitting residuals, and may therefore be favored in spite of the higher parameter variability. Indeed, there were still improvements after training for 5000 epochs (Figure 5).

In contrast, for supervised learning, bias toward the mean of the training distribution was found to persist even after extensive training, for data with low SNR (or low f). While this approach produced smooth parameter maps, little confidence can be placed in the estimated maps for D^* (Figures 6, 7). Similar bias was demonstrated by Gyori et al. for the spherical mean technique diffusion model.²⁰ Since parameter accuracy for supervised learning is highly dependent on the choice of training data (Figure 4), it may be of benefit to derive ground truth parameter combinations from conventional model fitting, as done for the synthetic patient distribution described in this study, such that the degree of bias is reduced. Nevertheless, it is important to remain aware that this inherent bias, although reduced, will ultimately still be present in regions of low SNR, in spite of the possibly alluring smoothness of the parameter maps.

We showed that the choice of learning rate strongly impacts the convergence behavior of the loss term, particularly for unsupervised learning. A higher learning rate was associated with spiky convergence, which has the potential for poor parameter estimates if the termination of training coincides with such a spike in the loss term. In contrast, a lower learning rate mitigates this spiky convergence at the cost of more training epochs. In this work, we have used a fixed learning rate. An alternative approach would

be to vary the learning rate over the training process, using a so-called learning rate schedule, which could speed up convergence while avoiding spiky behavior at the end of training.

We showed that a network with a higher number of hidden units in each layer resulted in increased convergence speed. Previous approaches^{16,17} set the hidden units equal to the number of measured b values (e.g., 16 in this study). Consequently, if a sparse protocol were used for acquiring DWI data, then the hidden units would be very low and could limit the learning capacity of the DNN. Therefore, we recommend using a fixed number of hidden units that is considerably higher than the number of b values (e.g., 64 in this study) in order to avoid a low learning capacity. Adding even more network parameters may make the network more powerful, but this would also extend computation time, and ultimately may not be beneficial if the network is already sufficiently powerful for the application.

In general, we found that altering hyperparameters of the network did not produce a considerable improvement in performance, provided that training was sufficiently long to ensure convergence (Figure 3). The marginally lower loss that resulted from a higher batch size is expected because the gradients calculated at each training step are more reliably defined. Furthermore, we found that using batch normalization, dropout, or the SGD optimizer could result in inferior performance, in particular for unsupervised learning. Therefore, use of these methods must be carefully evaluated. Performance of batch normalization may be improved by increasing batch size.

For the in vivo data (Figure 7), the parameter maps and RMSE maps for the unsupervised networks that were trained on synthetic data were very similar to those for the additional unsupervised network that was trained directly on the in vivo data (Epoch-50000). This suggests that a single unsupervised network may possibly be trained using synthetic data generated by a broad range of parameter combinations, and utilized for every anatomy, provided that the amount of training were sufficiently extensive. It is important to note that these parameter maps were also very similar to those obtained by conventional model fitting. As such, the practical benefit of these unsupervised networks may simply be a possible reduction in computation time. Correspondence between the results of the unsupervised approach and LSQ fitting is expected since both minimize the L2-norm on the signals (signals-MSE, Eq. 1). This is most evident for D^* , whereas for D and f there is perhaps greater similarity with the maps of the segmented approach. One supposition is that, similar to the segmented approach, an unsupervised network may first estimate adequate D , wherefrom it can learn adequate

f and subsequently adequate D^* . This may explain why f and D^* are correlated if training is terminated early.

Every network performed poorly on data that lay outside of the respective training distribution, so-called out-of-distribution (OOD) data. In Figure 6, the networks trained on the synthetic data from the patient distribution showed low RMSE for all voxels in the representative slice, especially after extensive training. Conversely, the networks trained on synthetic data from the uniform distribution showed high RMSE for the voxels associated with parameters that lay outside of the training distribution (i.e., $D > 3 \times 10^{-3}$ mm²/s and $f > 50\%$; red arrows). The impact of these OOD data in the test set was also observed in the plot of signals-MSE in Figure 4A, where the unsupervised network trained on the uniform distribution displayed an increase in signals-MSE as training progressed. However, if the patient distribution test set was instead truncated using the same bounds as the uniform distribution, the difference in signals-MSE between the two unsupervised networks was found to be minimal (data not shown). One further example of the consequences of OOD data is shown in Figure 5, where the networks trained on the synthetic patient distribution show inaccurate estimation of low D , despite high S_0 . This is due to the lack of certain parameter combinations occurring in the patient distribution, while being present in the uniform distribution test set, which were therefore not seen during training.

Note that the poorly-fitted OOD data in Figure 6 correspond to parameter values that are not physically realistic (i.e., $D > 3 \times 10^{-3}$ mm²/s and $f > 50\%$). This demonstrates one limitation of this study, which uses conventional model fitting to generate and provide synthetic patient data for training and testing. Specifically, we used the two-step segmented approach, since this was found to be more robust to noise than LSQ for the in vivo data considered in this study (Figures 7 and 8). These conventional methods are applied under the assumption that the in vivo data are well approximated by a bi-exponential function. However, the IVIM model does not account for other possible biophysical processes or imaging artifacts (e.g., subject motion) that may influence the signal. These factors make it difficult to generate synthetic data and corresponding ground truth parameters that provide an accurate representation of real data and the underlying biophysics. Future work should consider how to provide more appropriate ground truth data for training.

DNNs are promising for enhancing medical imaging technologies, yet for IVIM fitting, performance is strongly dependent on certain design choices, which can make results difficult to interpret, particularly in clinical practice. Clinical assessment is therefore necessary

to determine whether DNNs provide a clinical advantage over other advanced estimators. Comparison to histological characterization could be invaluable, where, for instance, gliomas have been shown to produce a pathological microvascular network in the periphery through neoangiogenesis, which results in hyperperfusion to provide nutrients for tumor growth, whereas the center of the tumor consists of dead tissue that is hypoperfused.^{2,26} A ring of high f values at the location of the tumor would be consistent with this histological finding, which in Figure 7 is most clearly shown for the unsupervised approaches at Epoch-50000. Tumor classification could be another important consideration of clinical assessment,²⁷ where the choice of fitting approach may be a key factor in the pursuit of quantitative biomarkers (Figure 8).

This work focused on voxel-wise fully-connected network implementations for IVIM fitting, but a detailed investigation of the impact of other network architectures has not been considered. For instance, Vasylechko et al.²⁸ proposed an unsupervised convolutional neural network based on the U-NET architecture.²⁹ It is expected that the incorporation of spatial information into the training of a neural network for IVIM fitting may improve performance, but further investigation of this is beyond the scope of the present article and has been left for future work. Lastly, this study mainly focused on the brain, which is a particularly challenging organ for IVIM fitting given the typically low f values. A comprehensive comparison study for other anatomies should be pursued in future work, although similar findings regarding extensive training and biases are expected.

5 | CONCLUSIONS

Using both simulations and in vivo data from glioma patients, we have explored the potential impact of key training features in unsupervised and supervised learning for IVIM model fitting. The main motivation of this work was not to develop a network that could improve IVIM parameter estimation. Rather, our aim was to investigate the possible limitations of voxel-wise deep learning IVIM fitting, by surveying a broad range of reasonable design choices. For both learning strategies, we demonstrated that a high learning rate, small network size, and early stopping can result in sub-optimal solutions and correlated parameters. We demonstrated that extending training beyond early stopping can resolve these correlations and reduce parameter error, providing an alternative to exhaustive hyperparameter optimization. However, extensive training resulted in increased sensitivity to noise,

where unsupervised parameter estimates displayed variability similar to estimates from conventional LSQ fitting. In contrast, supervised parameter estimates demonstrated improved precision, but showed strong biases towards the mean of the training distribution. Hence, visual assessment alone is not sufficient for evaluating the quality of the estimates. While the apparent sensitivity to noise in unsupervised learning may be undesirable, it could be argued that the corresponding variability observed in the parameter maps is indicative of the underlying uncertainty, which is indeed useful information. This uncertainty is exemplified by the contrasting D^* maps between approaches, despite the similar residual maps, and illustrates the difficulty in estimating D^* in the brain. Therefore, while deep-learning-based model fitting is a promising tool for IVIM parameter estimation, the impact of design choices on fitting performance and biases must be carefully evaluated.

ACKNOWLEDGMENTS

Misha P. T. Kaandorp, Frank Zijlstra, and Peter T. While gratefully acknowledge support from the Research Council of Norway under FRIPRO Researcher Project 302624.

DATA AVAILABILITY STATEMENT

To stimulate wider clinical implementation of IVIM, we have made the code for our unsupervised and supervised networkstrained on the uniform distribution synthetic data set available on GitHub: https://github.com/Mishakaandorp/Explore_Deep_Learning_IVIM.

ORCID

Misha P. T. Kaandorp  <https://orcid.org/0000-0002-7340-8256>

Frank Zijlstra  <https://orcid.org/0000-0002-9184-7666>

Peter T. While  <https://orcid.org/0000-0003-2602-0758>

REFERENCES

1. Le Bihan D, Breton E, Lallemand D, Aubin M, Vignaud JLM. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. *Radiology*. 1988;168:497-505.
2. Federau C, Meuli R, O'Brien K, Maeder P, Hagmann P. Perfusion measurement in brain gliomas with intravoxel incoherent motion MRI. *Am J Neuroradiol*. 2014;35:256-262. doi:10.3174/ajnr.A3686
3. Klaassen R, Steins A, Gurney-Champion OJ, et al. Pathological validation and prognostic potential of quantitative MRI in the characterization of pancreas cancer: preliminary experience. *Mol Oncol*. 2020;14:2176-2189. doi:10.1002/1878-0261.12688
4. Zhu L, Wang H, Zhu L, et al. Predictive and prognostic value of intravoxel incoherent motion (IVIM) MR imaging in patients with advanced cervical cancers undergoing concurrent chemo-radiotherapy. *Sci Rep*. 2017;7:1-9. doi:10.1038/s41598-017-11988-2

5. Lemke A, Laun FB, Klau M, et al. Differentiation of pancreas carcinoma from healthy pancreatic tissue using multiple *b*-values. *Invest Radiol*. 2009;44:769-775. doi:10.1097/rli.0b013e3181b62271
6. Notohamiprodjo M, Chandarana H, Mikheev A, et al. Combined intravoxel incoherent motion and diffusion tensor imaging of renal diffusion and flow anisotropy. *Magn Reson Med*. 2015;73:1526-1532. doi:10.1002/mrm.25245
7. Luciani A, Vignaud A, Cavet M, et al. Liver cirrhosis: Intravoxel incoherent motion MR imaging-pilot study. *Radiology*. 2008;249:891-899. doi:10.1148/radiol.2493080080
8. Federau C. Intravoxel incoherent motion MRI as a means to measure in vivo perfusion: a review of the evidence. *NMR Biomed*. 2017;30:1-15. doi:10.1002/nbm.3780
9. Federau C, Sumer S, Becce F, et al. Intravoxel incoherent motion perfusion imaging in acute stroke: initial clinical experience. *Neuroradiology*. 2014;56:629-635. doi:10.1007/s00234-014-1370-y
10. Pekar J, Moonen CTW, van Zijl PCM. On the precision of diffusion/perfusion imaging by gradient sensitization. *Magn Reson Med*. 1992;23:122-129. doi:10.1002/mrm.1910230113
11. While PT. Advanced methods for IVIM parameter estimation. In: Bihan D, Iima M, Federau C, Sigmund EE, eds. *Intravoxel incoherent motion (IVIM) MRI: Principles and applications*. Pan Stanford Publishing; 2018:449-484.
12. Meeus EM, Novak J, Withey SB, Zarinabad N, Dehghani H, Peet AC. Evaluation of intravoxel incoherent motion fitting methods in low-perfused tissue. *J Magn Reson Imaging*. 2017;45:1325-1334. doi:10.1002/jmri.25411
13. Barbieri S, Donati OF, Froehlich JM, Thoeny HC. Impact of the calculation algorithm on biexponential fitting of diffusion-weighted MRI in upper abdominal organs. *Magn Reson Med*. 2016;75:2175-2184. doi:10.1002/mrm.25765
14. Orton MR, Collins DJ, Koh DM, Leach MO. Improved intravoxel incoherent motion analysis of diffusion weighted imaging by data driven Bayesian modeling. *Magn Reson Med*. 2014;71:411-420. doi:10.1002/mrm.24649
15. While PT. A comparative simulation study of bayesian fitting approaches to intravoxel incoherent motion modeling in diffusion-weighted MRI. *Magn Reson Med*. 2017;78:2373-2387. doi:10.1002/mrm.26598
16. Barbieri S, Gurney-Champion OJ, Klaassen R, Thoeny HC. Deep learning how to fit an intravoxel incoherent motion model to diffusion-weighted MRI. *Magn Reson Med*. 2020;83:312-321. doi:10.1002/mrm.27910
17. Kaandorp MPT, Barbieri S, Klaassen R, et al. Improved unsupervised physics-informed deep learning for intravoxel incoherent motion modeling and evaluation in pancreatic cancer patients. *Magn Reson Med*. 2021;86:2250-2265. doi:10.1002/mrm.28852
18. Spinner GR, Federau C, Kozerke S. Bayesian inference using hierarchical and spatial priors for intravoxel incoherent motion MR imaging in the brain: analysis of cancer and acute stroke. *Med Image Anal*. 2021;73:102144. doi:10.1016/j.media.2021.102144
19. Bertleff M, Domsch S, Weingärtner S, et al. Diffusion parameter mapping with the combined intravoxel incoherent motion and kurtosis model using artificial neural networks at 3 T. *NMR Biomed*. 2017;30:1-11. doi:10.1002/nbm.3833
20. Gyori NG, Palombo M, Clark CA, Zhang H, Alexander DC. Training data distribution significantly impacts the estimation of tissue microstructure with machine learning. *Magn Reson Med*. 2021;87:932-947. doi:10.1002/mrm.29014
21. Federau C, O'Brien K. Increased brain perfusion contrast with T2-prepared intravoxel incoherent motion (T2prep IVIM) MRI. *NMR Biomed*. 2015;28:9-16. doi:10.1002/nbm.3223
22. Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings. 2016: 1-14.
23. Le Bihan D, Iima M. Diffusion magnetic resonance imaging: what water tells us about biological tissues. *PLoS Biol*. 2015;13:1-13. doi:10.1371/journal.pbio.1002203
24. Kingma DP, Ba JL. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015: 1-15.
25. Tukey JW. Exploratory data analysis. *Biometrics*. 1977;33:768.
26. Kumar V, Abbas AK, Fausto N, Aster JC. *Robbins and Cotran pathologic basis of disease*. Elsevier Health Sciences; 2014.
27. Vidić I, Jerome NP, Bathen TF, Goa PE, While PT. Accuracy of breast cancer lesion classification using intravoxel incoherent motion diffusion-weighted imaging is improved by the inclusion of global or local prior knowledge with bayesian methods. *J Magn Reson Imaging*. 2019;50:1478-1488. doi:10.1002/jmri.26772
28. Vasylechko SD, Warfield SK, Afacan O, Kurugol S. Self-supervised IVIM DWI parameter estimation with a physics based forward model. *Magn Reson Med*. 2022;87:904-914. doi:10.1002/mrm.28989
29. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Computer-assisted intervention. In: Navab N, Hornegger J, Wells WM, Frangi AF, Hutchison D, eds. *International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI*. Dimensions; 2015:234-241.

How to cite this article: Kaandorp MPT, Zijlstra F, Federau C, While PT. Deep learning intravoxel incoherent motion modeling: Exploring the impact of training features and learning strategies. *Magn Reson Med*. 2023;90:312-328. doi: 10.1002/mrm.29628

Paper III

Incorporating spatial information in deep learning parameter estimation with application to the intravoxel incoherent motion model in diffusion-weighted MRI

Misha P.T. Kaandorp^{1,2,3}, Frank Zijlstra^{1,2}, Davood Karimi³, Ali Gholipour³, Peter T. While^{1,2}

¹ Department of Radiology and Nuclear Medicine, St. Olav's University Hospital, Trondheim, Norway

² Department of Circulation and Medical Imaging, NTNU – Norwegian University of Science and Technology, Trondheim, Norway

³ Department of Radiology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

*** Corresponding Author**

Name: Misha Pieter Thijs Kaandorp

Department: Department of Radiology and Nuclear Medicine

Institute: St. Olav's University Hospital

Address: Ragnhilds gate 15, 7030 Trondheim, Norway

Email: mpkaando@stud.ntnu.no

Word count: 8932

This paper is submitted for publication and is therefore not included.

ISBN 978-82-326-7890-7 (printed ver.)
ISBN 978-82-326-7889-1 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)



NTNU

Norwegian University of
Science and Technology