Advancing Medical Imaging with Artificial Intelligence: PET Acquisition Enhancement and MRI-Based Brain Tumour Diagnosis

Milan Decuyper

Doctoral dissertation submitted to obtain the academic degree of
Doctor of Biomedical Engineering

Supervisors
Prof. Roel Van Holen, PhD - Prof. Stefaan Vandenberghe, PhD

Department of Electronics and Information Systems
Faculty of Engineering and Architecture, Ghent University

October 2021

GHENT
UNIVERSITY

Members of the Examination Board

Chair

Prof. Hennie De Schepper, PhD, Ghent University

Other members entitled to vote

Prof. Karel Deblaere, PhD, Ghent University

Prof. Roland Hustinx, PhD, Université de Liège

Prof. Jef Vandemeulebroucke, PhD, Vrije Universiteit Brussel

Prof. Christian Vanhove, PhD, Ghent University

Supervisors

Prof. Roel Van Holen, PhD, Ghent University

Prof. Stefaan Vandenberghe, PhD, Ghent University

Acknowledgements

> *" There's no learning without trying lots of ideas*
> *and failing lots of times."*
> Jonathan Ive

Four years of PhD research bundled into this book. It should, however, be pointed out that doing a PhD is more than just bringing forth a book. Apart from many research results that did not make it into this dissertation, a PhD is also a personal journey and learning experience. It teaches you how to set up your own project, integrity and perseverance, responsibility, to deal with ups and downs and moments of insight and excitement but also of frustration and discouragement. This book therefore only contains a fraction of what I have learned during these past four years. As this journey would not have been possible without the help of many people, I would like to start this book with expressing my gratitude to all of you.

First of all, I would like to thank my promotors, prof. dr. Roel Van Holen and prof. dr. Stefaan Vandenberghe for giving me the opportunity and freedom to perform this research. Roel, thank you for seeing in me the right candidate to start a PhD position at MEDISIP. Through the weekly meetings and your ever constructive feedback, I have learned a lot on how to manage and shape my project, ask the critical questions and writing and presentation skills. You were there to tell me when to stop optimising my networks and start writing my research down into papers and conference presentations. Without you, I would probably still be tuning the network hyperparameters of my first study. Stefaan, thank you for always being available for questions and feedback on my papers and book. You taught me a lot about medical imaging.
I would also like to thank my co-authors and jury members for reading

i

my articles and dissertation and providing valuable feedback that significantly enhanced the quality of my work. My gratitude also goes to prof. dr. Ingeborg Goethals, dr. Giorgio Hallaert, ir. Sam Donche and dr. ir. Stijn Bonte for collecting the Ghent University Hospital data.

Next to prof. dr. Roel Van Holen and prof. dr. Stefaan Vandenberghe, I am grateful to prof. dr. Christian Van Hove, prof. dr. Pieter Van Mierlo, prof. dr. Vincent Keereman, dr. Benedicte Descamps and dr. Lars Emil Larsen for making MEDISIP a multidisciplinary research group in which everybody can feel at home. In this regard, two people are also invaluable. Saskia and Inge, your welcome helped me to immediately feel at home at MEDISIP. Thank you for providing all the administrative support which allows us to focus on our research, but also for organising non-research related activities and for always being there to support everyone at bioMMeda and MEDISIP. Furthermore, thanks to Jurgen for all the technical aid and helping me repair my GPU by fixing new fans.

Of course, I also want to thank my current and former colleagues. Willeke and Thibault, as supervisors of my master thesis you taught me a lot of skills that were essential during my PhD. Thanks to you I got to know the MEDISIP research group and your encouragements made me jump into the PhD journey. Although you both successfully finished your PhD quite soon after I started, being surrounded (literally the two desks next to me) by my former supervisors made the start much easier. You also immediately set the bar high with your outstanding PhD defences. Next is Stijn who quickly brought me up to speed about everything related to primary brain tumours. The work in this book on computer-aided brain tumour diagnosis builds upon, and was therefore never possible without, your work. I will also always remember your very dry sense of humour and never saying no to an additional piece of cake. Kim, you always brightened up our office even though I was sometimes startled by your occasional cursing at your computer. I am vey much looking forward to having you as my colleague again soon. Jens Mincke, you were not often at the office (always taking care of your plants) but when you randomly showed up you always made sure to have a talk with everyone and set up some practical jokes thereby cheering up the entire office. Tim, I remember you as the cool rock dude. Thank you for being my office neighbour for two years. Paulo, I still admire your ever-lasting optimism and positivity. You were always up for a laugh and your presence was never unnoticed. I wish you a lot

of succes on your postdoc position in Brasil. Maya, as you live in Qatar I did not see you very often but I really liked to meet you at MIC in Manchester. I admire you for doing PhD research while being a mother and a physics teacher at the Texas A&M university in Qatar. Emma, my office mate during my entire PhD. One can always count on you to lend a helping hand. Your precision, work ethic and incredible research make you an example PhD student that everybody looks up to. I would also like to thank Prakash for co-organising our starters party together with Mariele and joining the trip to Chomutov for Marek's wedding. Which brings me to Marek, thank you (and Darja) for the invitation to your wedding in Czech Republic. I really enjoyed getting to know your home town, family, wedding traditions and the 'Tatratea mountains'. Jens Maebe, I am happy that someone will continue the work on using deep learning to enhance PET. I am sure you will do a great job. Ashkan, it was nice to also have someone in the group that does not do purely academic research but really tries to bring a new product idea into reality. With your presentation skills and passion in developing your product, Exoligamentz will certainly be a great success. Charlotte, not only our floor, but probably the entire UZ campus could hear when you arrived at the office. You always bring a lot of enthusiasm and energy. I will always remember our trip to MIC in Manchester together. On that trip I got to know you much better (including our common guilty pleasure ;)) and I really had a lot of fun. Too bad it was our last offline conference because of corona. I am glad that we will continue to be colleagues at Molecubes. Jolan, my neighbour in the office for the past two years. While Emma was a good influence to concentrate on my work, you were probably the opposite. Always ready for a (not so short) coffee break or a walk. I did really enjoy having you as my colleague though and I admire you for taking on big challenges that would seem impossible for others such as simultaneously doing a PhD and studying medicine. I am sure you will somehow manage to do both with great success. Gert and Emma Depuyt, you are both mostly at other departments and due to corona I did not get the change to get to know you very well. I am confident, however, that you will form a great team together with Jolan and set the state-of-the-art in EEG analysis with machine learning for many applications. And last but not least I want to thank Mariele. We started and submitted our PhD together and often went through the same phases. It was nice to always have someone to vent to and share frustrations and succes with. We also worked closely together on the project to improve PET imaging

with neural networks. I learned a lot from you and you taught me all about PET detectors which I knew nothing about before I started my PhD. After many data iterations, discussions, hand gestures ($\|$, $\backslash/$) and video calls during corona, I believe we accomplished great work. We not only worked together but also did a lot of fun things such as our starters party, the trip to Chomutov and Prague, training for and running a half marathon, going for drinks etc. Thank you for going through this PhD journey together and I wish you all the luck at your new home in Aalst. Furthermore, I want to thank the colleagues of bioMMeda for the often hilarious but also sometimes twisted and awkward (yes, I am thinking about you Gerlinde and Mathias ;)) lunch time discussions.

Mijn familie, en voornamelijk mijn ouders en zus verdienen natuurlijk ook een belangrijke plaats in dit dankwoord. Het is maar al te gemakkelijk dit als vanzelfsprekend te beschouwen, maar door jullie onvoorwaardelijke steun doorheen mijn studies en doctoraat ben ik zo ver gekomen. Bedankt om er te zijn op momenten wanneer ik het even niet meer zag zitten, voor de vele jaren dat ik in Gent op kot mocht, natuurlijk ook de financiële steun, de vele ritjes van en naar het station, de potjes met lekker eten.... kortom om er altijd te zijn voor mij.

Ten slotte blijft een heel belangrijk persoon in mijn leven nog over: Brecht - Lief. Bedankt om mij altijd te steunen en in mij te geloven, ook op momenten waarop ik het moeilijk had en gestresseerd rondliep waarbij ik niet altijd aanspreekbaar was. Samen gaan wonen bij de start van de lockdown en de laatste periode van mijn doctoraat was meteen een grote test, maar een die we met glans hebben doorstaan. Bedankt voor de vele mooie momenten en ik weet zeker dat we er nog veel tegemoet gaan.

Milan Decuyper
Oktober 2021, Gent

Summary

This PhD dissertation involves the application of artificial intelligence to advance medical imaging. Driven by the ever increasing amount of computational power and generated digital data, AI is employed to develop voice assistants, computer vision, self-driving cars, recommendation systems etc. These systems already achieve incredible performances that match or even outperform humans and are finding their way into our daily lives. Also in healthcare the need and potential of AI arises. Electronic health records contain a wealth of information that can be used towards personalised and precision medicine. Due to the immense quantity and complexity of this data, especially in medical imaging, and the limited available workforce, it is not possible to fully exploit all this information. For this reason, AI algorithms are being developed to improve the efficiency of the radiological workflow.

In this work we focus on two use cases of AI in medical imaging. The first use case is situated at the acquisition stage where we use neural networks to improve the spatial resolution of positron emission tomography (PET) detectors and consequently PET scanners. The second application is located at the very end of the imaging pipeline, on the analysis of pre-therapy magnetic resonance images (MRI) for computer-aided brain tumour segmentation and diagnosis.

PET detector calibration

The purpose of a detector in positron emission tomography is to stop incoming gamma rays and determine their energy, interaction position and arrival time inside the detector. Current clinical PET scanners use pixelated detector designs consisting of long and thin scintillation crystals. A downside of this design is a spatial resolution that is limited to

the pixel size. When decreasing pixel size to improve spatial resolution, this results in increased dead space and reduced sensitivity. A monolithic design, consisting of a large monolithic crystal connected to a pixelated photodetector array, can prove to be beneficial in terms of sensitivity and spatial, timing and energy resolution. Moreover, monolithic detectors allow intrinsic depth-of-interaction encoding resulting in accurate 3D positioning of the interaction inside the crystal. Monolithic PET detectors do, however, require lengthy calibration procedures and powerful positioning algorithms to determine the exact position of interaction. To train these algorithms, calibration data needs to be acquired by irradiating the detector with a source beam at pre-defined discrete positions. We investigated the use of artificial neural networks to estimate the 3D gamma interaction position from the light response measured by the photodetector array in a monolithic PET detector.

In a first study, simulation data was used to assess the optimal network complexity, amount of training data and training procedure. The ultimate achievable spatial resolution was investigated and compared with an established positioning algorithm called mean nearest neighbour. Optimal performance was achieved with a network containing three hidden layers of 256 neurons trained on 1000 events per training grid position. Results showed that a very high spatial resolution was obtained of around 0.50 mm FWHM across the entire detector. Comparison with mean nearest neighbour positioning demonstrated superior performance both in spatial resolution as in computational efficiency. When training neural networks, however, potential overfitting on the discrete training positions should be carefully evaluated through the use of validation data acquired at intermediate positions. Network complexity should be tuned to the calibration setup and we showed that by stopping the training process when performance on intermediate data stops improving, strong overfitting and thus non-uniform positioning can be prevented.

Furthermore, we investigated the potential degrading effect of intra-crystal Compton scatter and calibration source beam width on the achievable spatial resolution with neural networks.
Around 60% of the arriving gamma rays first undergo one or multiple Compton interactions inside the LYSO crystal before final photoelectric absorption. Estimation of the required first interaction position from the measured electronic signal is difficult as often only a small amount of energy is released when Compton scattering. Evaluation of spatial resolution with and without Compton scattered events revealed that

Compton scatter has a significant degrading effect on the overall positioning accuracy (mean 3D positioning error of 2.29 mm versus 0.49 mm). However, the positioning error depends on the scatter distance and only a small fraction of events scatters very far (10% more than 8 mm). A network specifically trained to position Compton scattered events did not result in an improvement in performance. We therefore investigated whether networks can identify far scattered events and could help to improve positioning accuracy. To this end, a network was trained to predict 3D scatter distance. This network could be used to filter out far scattered events in order to improve spatial resolution with a tradeoff in sensitivity which can be justified in certain applications. Considering the limited practicality of training a scatter prediction network in an experimental setup (no available labels), a different approach was investigated using a Bayesian neural network. This method allows to train one network to predict both the position as the positioning uncertainty related to Compton scatter without requiring additional information on Compton scattering. When filtering out 10% most uncertain events, the mean positioning error could be reduced from 1.54 mm to 1.23 mm.
A calibration source with a certain beam width can introduce differences between the ground truth position label and the actual first interaction position. These errors in the ground truth data could influence the training process of neural networks. Comparison between a network trained on data acquired with a perfectly narrow beam versus a calibration source with a realistic beam width of 0.6 mm showed no significant difference in achieved intrinsic spatial resolution. The beam diameter does, however, influence the measured spatial resolution (0.74 mm versus 0.52 mm FWHM) which should be taken into account when evaluating and comparing spatial resolution of different PET detectors.

The developed neural networks and training procedure were also evaluated on an experimental setup. Similar to the results on simulation data, high spatial resolutions (around 1 mm FWHM in detector centre) could be achieved with neural networks, superior to the mean nearest neighbour positioning algorithm (1.14 mm FWHM in centre region). Neural networks are trained on individual events, directly learn to infer the interaction position from the measured light distribution and produce continuous coordinate outputs, not restricted to a discrete calibration grid. This leads to an improved positioning accuracy of Compton scattered events and less degradation near the detector edges. Improved spatial resolution of PET detectors with neural networks can help reach

the physical limits of PET and a better detection of small tumours. Moreover, when achieving better spatial resolutions than required, there is room to trade resolution for other parameters, e.g.: less readout channels, inexpensive materials with less light output, larger detector thickness etc. Lastly, positioning events with the network is fast and parallelisable, especially when using powerful hardware like GPUs.

Computer-aided primary brain tumour diagnosis

The second part of this work focuses on the application of AI in medical image analysis, specifically for primary brain tumour segmentation and diagnosis. Primary brain tumours are a complex type of neoplasms that originate in the brain and are relatively rare with an incidence rate of 10.8 people per 100,000 per year. They are, however, a significant cause of cancer morbidity and mortality, especially in children and young adults where they are the leading cause of cancer deaths. The most common types of PBTs are glioma and meningioma. In this work, we focus on the characterisation of glioma. In the most recent WHO classification guidelines, increased emphasis is placed on the integration of molecular markers next to histopathological analysis. Integration of genotypic parameters for tumour classification intends to add objectivity and yield more narrowly defined diagnostic entities with respect to prognosis and optimal therapy. Three markers play a central role in the classification of glioma: histological grade, isocitrate dehydrogenase (IDH) 1 and/or 2 mutation and co-deletion of chromosome arms 1p and 19q. Determination of these markers requires tumour samples obtained through biopsy or resection. These invasive procedures involve risks and are not always possible to perform depending on tumour location and accessibility, the patient's clinical condition or when the patient refuses a surgical procedure. Therefore, non-invasive assessment of clinically relevant markers based on medical images can aid in characterising glioma and guide therapy and surgery planning, especially when extraction of tumour tissue is not possible or genetic testing not available. It has been shown that MR tumour phenotype is correlated with genetic markers and malignancy. However, visual interpretation and prediction of tumour properties remains very challenging and inaccurate. To increase the efficiency and accuracy of non-invasive glioma characterisation, AI algorithms are developed. Many existing approaches use manually obtained tumour segmentations which introduces subjectivity and variability in

performance. Moreover, they are often trained and evaluated on data from a small dataset acquired at one institution, limiting their generalisability to data from different centres. The goal of this work is to develop an accurate, robust and fully automatic pipeline to segment and characterise glioma using deep convolutional neural networks.

In a first study, we investigated the task of non-invasively distinguishing high-grade glioblastoma (GBM) from lower-grade glioma (LGG). The BraTS 2017 dataset consisting of 210 GBM and 75 LGG cases was used for this study. For every patient, four MRI sequences (T1, T1ce, T2 and FLAIR) were provided with manual tumour segmentation labels. Predictive performance of hand-engineered radiomics features that describe tumour shape, texture and intensity was compared with features extracted using a pre-trained CNN. Moreover, we compared the performance of pre-trained CNN features extracted from different input scales: one or multiple slices and with or without cropping to the tumour ROI. Classification of the features was done using a Random Forest classifier. Best performance was achieved with the radiomics features extracted from manually segmented tumour volumes (AUC of 96%). Features from a pre-trained CNN, on the other hand, had a high predictive value as well and allowed to design a fast and automatic binary grading system reaching an AUC score of 91%. Best performance was achieved when cropping the MRI to the tumour ROI (AUC of 94%).

Since manual tumour segmentation is time- and labour-intensive and prone to inter- and intra-observer variability, we developed an automatic tumour delineation network based on the U-Net architecture. The network was trained using the BraTS 2019 training dataset (335 patients) and evaluated on the BraTS 2019 validation set (125 patients). Accurate delineation of different tumour regions was achieved with average Dice scores of 90%, 83% and 76% for the total abnormal, tumour core and enhancing tumour regions respectively. In clinical practice, not all four input MRI (T1, T1ce, T2 and FLAIR) are always available. We therefore applied input channel dropout, i.e. randomly excluding input MRI during training, which we demonstrate to significantly increase robustness to missing input modalities. These scores match state-of-the-art results reported in the most recent BraTS challenges and we believe that the obtained performance is sufficiently high to be useful in a clinical setting. It can be debated whether further improving the Dice scores with a few percentages is clinically relevant as they are evaluated on manual delineations which are also not 100% accurate.

Objectivity and robustness could be more important when analysing brain tumour volumes and progression over time. Qualitative evaluation on independent data acquired at the Ghent University Hospital showed good generalisation performance.

To train a brain tumour classification network, a large dataset of 628 patients was collected from multiple public databases available on The Cancer Imaging Archive. The automatic segmentation network was applied to extract the 3D tumour region of interest from every MRI sequence. Subsequently, a classification 3D CNN was trained to predict tumour grade, IDH and 1p/19q co-deletion status. Multi-task learning was employed to simultaneously predict these three markers and to deal with missing ground truth labels in the dataset while reducing the risk of overfitting. On a test dataset of 100 patients, not used during training, the network achieved AUC scores of 93% for WHO grade, 94% for IDH mutation and 82% for 1p/19q co-deletion prediction. We additionally evaluated the classification performance on an entirely independent dataset of 110 patients retrospectively acquired at the Ghent University Hospital. On this dataset, AUC scores were reported of 94%, 86% and 87% for the three tasks respectively.

The above two-stage approach can have some downsides as the classification network only operates on the tumour region of interest which excludes potentially relevant information on location and surrounding tissue. Moreover, possible errors in the prior segmentation step could also influence the subsequent classification performance. As an alternative, we developed a network that performs simultaneous segmentation and classification based on the full brain MRI. The segmentation U-Net was extended with a classification branch and called Y-Net. Through the use of multi-task learning, techniques to reduce GPU memory consumption and appropriate patch extraction, one network could be trained on the large multi-institutional and heterogeneous database containing many cases with missing labels. A similar segmentation performance was achieved with average Dice scores of 89%, 84% and 75% for the whole tumour, tumour core and enhancing tumour regions respectively. In terms of classification performance, WHO grade could be predicted with 98%, IDH mutation with 96% and 1p/19q co-deletion with 87% AUC on the TCIA test dataset. On the independent GUH test data, the AUC scores were 96%, 83% and 90%. Classification performance is slightly higher than with the two-stage approach. This is possibly because the entire input MRI is now processed and the addition of the segmentation

task could provide additional regularisation to the training process. Finally, insights into the network's visual knowledge and extracted imaging features were obtained using several visualisation techniques. The feature embeddings of the network were plotted for every brain tumour case in the dataset after t-SNE feature reduction. This revealed different clusters of brain tumour cases with similar imaging characteristics such as ring enhancement, lesion size, frontal lobe location and T2-FLAIR mismatch. These are indeed known imaging features that are correlated with these tumour markers. Saliency maps, that visualise where the network places the most attention in the input MRI to make a certain prediction, showed that the network indeed looks at the relevant tumour regions. This allows an additional check to gain confidence in the network's predictions. Lastly, a synthetic input was generated that maximises the output scores for a glioblastoma, IDH wildtype tumour. Starting from random noise, a ring-enhancing tumour pattern appeared with a hypo-intense core in the T1ce channel and surrounding hyper-intense tissue on the T2 channels. This indicates that the network learned to attribute these features to this tumour type.

In conclusion, we have demonstrated that neural networks can improve the image acquisition process on detector level which eventually results in better image quality and affects the entire remaining imaging pipeline. Furthermore, an image analysis application was researched on primary brain tumour characterisation resulting in non-invasive and accurate brain tumour segmentation and diagnosis tools. Although challenges remain regarding standardised datasets and understanding of AI, both by experts and the general public, we can conclude that AI will have a profound impact on radiology. It will become and important tool supporting radiologists to increase efficiency, perform routine tasks and enable personalised and precision medicine. It will, however, not replace radiologists as many vital elements to the radiological profession can never be automated.

Samenvatting

Deze doctoraatsthesis behandelt de toepassing van artificiële intelligentie (AI) ter bevordering van medische beeldvorming. Gedreven door de steeds grotere hoeveelheid beschikbare rekenkracht en gegenereerde digitale data, wordt AI gebruikt om slimme assistenten, computervisie, zelfrijdende auto's, aanbevelingssytemen enzovoort te ontwikkelen. Deze systemen leveren al ongelooflijke prestaties die de mens evenaren of zelfs overtreffen en vinden hun weg in ons dagelijks leven. Ook in de medische zorg ontstaat de nood en het potentieel van AI. Elektronische medische dossiers bevatten een schat aan informatie die kan worden gebruikt voor gepersonaliseerde en precisie geneeskunde. Door de immense hoeveelheid en complexiteit van deze data, zeker binnen medische beeldvorming, en het beperkte aantal experten die deze data kunnen verwerken, is het niet mogelijk om al deze informatie volledig te benutten. Om deze reden worden AI algoritmes ontwikkeld die de efficiëntie van de radiologische workflow kunnen verbeteren.

In dit werk richten we ons op twee toepassingen binnen medische beeldvorming. De eerste bevindt zich in de acquisitiefase waar we neurale netwerken gebruiken om de spatiële resolutie van positronemissietomografie (PET) detectoren en dus ook PET scanners te verbeteren. De tweede toepassing bevindt zich helemaal aan het einde van het beeldvormingsproces, de analyse van pre-therapie magnetische resonantie (MR) scans voor computer-ondersteunde segmentatie en diagnose van primaire hersentumoren.

PET detector kalibratie

Het doel van een PET detector is om inkomende gamma stralen te stoppen en hun energie, interactiepositie en aankomsttijd in de detector

te bepalen. Huidige klinische PET scanners maken gebruik van geseg-
menteerde detectorontwerpen die bestaan uit lange en dunne kristallen.
Een nadeel van dit ontwerp is dat de spatiële resolutie beperkt is tot de
kristal grootte. Bij het gebruik van dunnere kristallen om de resolutie
te verhogen, resulteert dit in meer dode ruimte tussen de kristallen en
dus een lagere sensitiviteit. Een monolithisch ontwerp, bestaande uit
een groot monolithisch kristal verbonden met een gesegmenteerde foto-
detector matrix, kan gunstiger zijn op gebied van sensitiviteit en spatiële,
tijds- en energieresolutie. Bovendien maken monolithische detectoren in-
trinsieke interactiediepte-codering mogelijk, wat 3D-positionering van de
interactie in het kristal toelaat. Monolithische PET detectoren vereisen
echter langdurige kalibratieprocedures en krachtige positioneringsalgorit-
men om de exacte positie van interactie te bepalen. Om deze algoritmen
te trainen, moet kalibratie data worden verkregen door de detector te
bestralen op vooraf gedefinieerde discrete posities. We onderzochten het
gebruik van neurale netwerken om de 3D interactiepositie te schatten
op basis van de lichtrespons gemeten door de fotodetector matrix in een
monolithische PET detector.

In een eerste studie werd simulatiedata gebruikt om de optimale
netwerkcomplexiteit, hoeveelheid training data en trainingsprocedure
te bepalen. De ultiem haalbare spatiële resolutie werd onderzocht en
vergeleken met een gevestigd positioneringsalgoritme genaamd gemid-
delde dichtstbijzijnde buur. De beste performantie werd bereikt met een
netwerk bestaande uit drie verborgen lagen van 256 neuronen, getraind
op 1000 datapunten per kalibratiepositie. De resultaten toonden een
zeer hoge resolutie aan met een halfwaardebreedte van 0.50 mm over de
hele detector. Vergelijking met het gemiddelde dichtste buur algoritme
demonstreerde superieure prestaties met neurale netwerken, zowel in
spatiële resolutie als in rekenefficiëntie. Bij het trainen van neurale
netwerken moet echter de mogelijke overfitting op de afzonderlijke trai-
ningsposities zorgvuldig worden geëvalueerd. Dit kan door het gebruik
van validatiedata die is verkregen op tussenliggende posities. De net-
werkcomplexiteit moet worden afgestemd op de kalibratie-opstelling en
we hebben aangetoond dat door het trainingsproces te stoppen wanneer
de performantie op tussenliggende posities niet meer verbetert, sterke
overfitting en dus niet-uniforme positionering kan worden voorkomen.

Verder onderzochten we het potentieel degraderend effect van intra-
kristallijne Compton-verstrooiing en breedte van de kalibratie straal op
de positioneringsnauwkeurigheid van neurale netwerken.

Ongeveer 60% van de gammastralen ondergaat eerst een of meerdere Compton interacties in het LYSO kristal voordat ze definitief worden geabsorbeerd door middel van foto-elektrische absorptie. Het schatten van de vereiste eerste interactiepositie uit het gemeten signaal is moeilijk omdat er vaak slechts een kleine hoeveelheid energie vrijkomt bij Compton-verstrooiing. Evaluatie van spatiële resolutie met en zonder Compton-verstrooide data onthulde dat Compton-verstrooiing een significant degraderend effect heeft op de positioneringsnauwkeurigheid (gemiddelde 3D euclidische afstand tot de correcte positie van 2.29 mm versus 0.49 mm). De afwijking hangt echter sterk af van de verstrooiingsafstand en slechts een klein percentage van de data verstrooit zeer ver (10% meer dan 8 mm). Een netwerk dat specifiek is getraind om Compton-verstrooide gebeurtenissen te positioneren leidde niet tot een verbetering van de resolutie. We hebben daarom onderzocht of netwerken ver verstrooide gebeurtenissen kunnen identificeren om zo de performantie te verbeteren. Hiertoe werd een netwerk getraind om de 3D verstrooiingsafstand te voorspellen. Dit netwerk kon worden gebruikt om ver verstrooide gebeurtenissen uit te filteren en zo de resolutie te verbeteren met een klein verlies in sensitiviteit wat gewenst kan zijn in specifieke toepassingen. Gezien de beperkte bruikbaarheid van deze methode in een experimentele opstelling (geen beschikbare labels over verstrooiingsafstand) werd een andere aanpak onderzocht met behulp van een Bayesiaans neuraal netwerk. Deze methode maakt het mogelijk om één netwerk te trainen om zowel de positie als de positioneringsonzekerheid gerelateerd aan Compton-verstrooiing te voorspellen zonder dat aanvullende informatie over Compton-verstrooiing nodig is. Bij het uitfilteren van de 10% meest onzekere gebeurtenissen, kon de gemiddelde positioneringsfout worden verminderd van 1.54 mm tot 1.23 mm.

Een kalibratiebron met een bepaalde bundelbreedte kan verschillen introduceren tussen de positie labels (positie van de bron) en de daadwerkelijke eerste interactieposities. Deze fouten in de labels kunnen het trainingsproces van neurale netwerken beïnvloeden. Vergelijking tussen een netwerk getraind op data verkregen met een perfecte smalle bundel versus een kalibratiebron met een realistische bundelbreedte van 0.6 mm toonde geen significant verschil in bereikte intrinsieke spatiële resolutie. De bundeldiameter heeft echter wel invloed op de gemeten resolutie (0.74 mm versus 0.52 mm FWHM), waarmee rekening moet worden gehouden bij het evalueren en vergelijken van de resolutie tussen verschillende PET detectoren.

De ontwikkelde neurale netwerken en trainingsprocedure werden ook geëvalueerd op een experimentele opstelling. Vergelijkbaar met de resultaten op simulatiedata, werden hoge spatiële resoluties (ongeveer 1 mm FWHM in het centrum van de detector) bereikt met neurale netwerken, superieur aan het gemiddelde dichtste buur algoritme (1,14 mm FWHM in het centrum). Neurale netwerken worden getraind op individuele datapunten, leren om rechtstreeks de interactiepositie af te leiden uit de gemeten lichtverdeling en produceren continue waarden, niet beperkt tot een discreet kalibratieraster. Dit leidt tot een betere positioneringsnauwkeurigheid van Compton-verstrooide gebeurtenissen en minder degradatie nabij de detectorranden. Verbeterde spatiële resolutie van PET detectoren met neurale netwerken kan helpen de fysische limieten van PET te bereiken en kleine tumoren beter op te sporen. Bovendien is er bij het bereiken van betere resoluties dan vereist, ruimte om resolutie in te ruilen voor andere parameters zoals: minder uitleeskanalen, goedkopere materialen met minder lichtopbrengst, dikkere detectoren enz. Ten slotte is het positioneren van gamma interacties met het netwerk snel en parallelliseerbaar, vooral bij gebruik van krachtige hardware zoals GPU's.

Computer-geassisteerde primaire hersentumor diagnose

Het tweede deel van dit werk concentreert zich op de toepassing van AI in medische beeldanalyse, specifiek voor segmentatie en diagnose van primaire hersentumoren. Primaire hersentumoren vormen een complex type neoplasia die ontstaan in de hersenen en relatief zeldzaam zijn met een incidentie van 10.8 personen per 100 000 per jaar. Ze zijn echter een belangrijke oorzaak van morbiditeit en mortaliteit door kanker, vooral bij kinderen en jongvolwassenen waar ze de belangrijkste oorzaak zijn van sterfte door kanker. De meest voorkomende soorten primaire hersentumoren zijn gliomen en meningiomen. In dit werk richten we ons op de karakterisering van gliomen. In de meest recente classificatierichtlijnen van de WHO wordt naast histopathologische analyse meer nadruk gelegd op de integratie van moleculaire merkers. Integratie van genetische parameters voor tumorclassificatie is bedoeld om objectiever en nauwkeuriger gedefinieerde diagnostische entiteiten op te leveren met betrekking tot de prognose en optimale therapie. Drie merkers spelen een centrale rol bij de classificatie van gliomen: histologische graad, isocitraatdehydrogenase (IDH) 1 en/of 2 mutatie en co-deletie van chromosoomarmen 1p en 19q.

Bepaling van deze merkers vereist tumorweefsel die is verkregen door middel van biopsie of resectie. Deze invasieve procedures brengen risico's met zich mee en zijn niet altijd mogelijk, afhankelijk van de locatie en toegankelijkheid van de tumor, de klinische toestand van de patiënt of wanneer de patiënt een chirurgische ingreep weigert. Daarom kan niet-invasieve bepaling van klinisch relevante merkers op basis van medische beelden helpen bij het karakteriseren van gliomen en bij therapie en operatieplanning. Vooral wanneer extractie van tumorweefsel niet mogelijk is of genetische tests niet beschikbaar zijn. Het is aangetoond dat tumorfenotype op MR scans gecorreleerd is met genetische merkers en maligniteit. Visuele interpretatie en voorspelling van tumoreigenschappen blijft echter zeer uitdagend en onnauwkeurig. Om de efficiëntie en nauwkeurigheid van niet-invasieve glioomkarakterisering te verhogen, worden AI algoritmes ontwikkeld. Veel bestaande studies gebruiken manueel verkregen tumorsegmentaties die subjectiviteit en variabiliteit introduceren. Bovendien worden ze vaak getraind en geëvalueerd op data uit een kleine dataset die bij één instelling is verzameld waardoor de generaliseerbaarheid naar data van andere centra beperkt is. Het doel van dit werk is om nauwkeurige, robuuste en automatische algoritmes te ontwikkelen om gliomen te segmenteren en te karakteriseren met behulp van diepe, convolutionele neurale netwerken (CNN).

In een eerste studie onderzochten we het probleem van het niet-invasief onderscheiden van hooggradige glioblastomen (GBM) en lager-gradige gliomen (LGG). Voor dit onderzoek is gebruik gemaakt van de BraTS 2017 dataset bestaande uit 210 GBM en 75 LGG gevallen. In deze dataset zijn voor elke patiënt vier MR sequenties (T1, T1ce, T2 en FLAIR) en manuele tumorsegmentaties voorhanden. De voorspellende waarde van hand-gedefinieerde radiomics parameters of features die tumorvorm, textuur en intensiteit beschrijven, werd vergeleken met features die zijn geëxtraheerd met behulp van een CNN die vooraf werd getraind op niet-medische beelden. Bovendien vergeleken we de performantie van CNN features die werden geëxtraheerd uit verschillende MR input groottes: een of meerdere slices en met of zonder bijsnijden tot de tumor regio. Voor de classificatie is gebruik gemaakt van een Random Forest classificatie algoritme. De beste performantie werd bereikt met de radiomics features die werden geëxtraheerd op basis van handmatig gesegmenteerde tumor volumes (AUC van 96%). Features van een vooraf getrainde CNN hadden daarentegen ook een hoge voorspellende waarde en maakten het mogelijk om een snel en automatisch binair classificatie

systeem te ontwikkelen dat een AUC-score van 91% bereikte. De beste
score met CNN features werd behaald na het bijsnijden van de MR tot
de tumor regio (AUC van 94%).

Aangezien manuele tumorsegmentatie tijds- en arbeidsintensief is en
vatbaar voor variabiliteit tussen verschillende waarnemers, hebben we
een netwerk ontwikkeld voor automatische tumorsegmentatie op basis
van de U-Net architectuur. Het netwerk werd getraind met behulp van de
BraTS 2019 trainingsdataset (335 patiënten) en geëvalueerd op de BraTS
2019 validatieset (125 patiënten). Nauwkeurige segmentatie werd bereikt
met gemiddelde Dice scores van respectievelijk 90%, 83% en 76% voor
de totale abnormale, tumorkern- en contrast-capterende tumorregio's.
In de klinische praktijk zijn niet altijd alle vier de input MR sequenties
(T1, T1ce, T2 en FLAIR) beschikbaar. Daarom maakten we gebruik
van input kanaal dropout, d.w.z. het willekeurig op nul zetten van
een input MRI tijdens de training om zo het optreden van ontbrekende
sequenties te simuleren. We toonden aan dat dit de robuustheid tegen
ontbrekende sequenties aanzienlijk verhoogt. De behaalde Dice scores
komen overeen met state-of-the-art resultaten gerapporteerd in de meest
recente BraTS competities en we zijn van mening dat de verkregen per-
formantie voldoende hoog is om bruikbaar te zijn in een klinische setting.
Men kan zich afvragen of het verder verbeteren van de Dice scores met
enkele percentages klinisch relevant is, aangezien ze worden beoordeeld
op basis van manuele segmentaties die ook niet 100% nauwkeurig zijn.
Objectiviteit en robuustheid zijn mogelijk belangrijker bij het analyseren
van hersentumorvolumes en progressie in de tijd. Kwalitatieve evaluatie
op data verkregen in het Universitair Ziekenhuis Gent toonde een goede
generalisatie aan naar data van verschillende centra.

Om een hersentumor classificatienetwerk te trainen, werd een grote
dataset van 628 patiënten verzameld uit meerdere publieke databases,
beschikbaar op The Cancer Imaging Archive (TCIA). Het automatisch
segmentatienetwerk werd toegepast om de 3D tumorregio uit elke MR
sequentie te extraheren. Vervolgens werd een classificatie 3D CNN ge-
traind om de tumorgraad, IDH mutatie en 1p/19q co-deletie status te
voorspellen. Multitask leren werd gebruikt om deze drie merkers tegelij-
kertijd te kunnen voorspellen, om met ontbrekende labels in de dataset
om te gaan en terwijl het risico op overfitting te verkleinen. Op een test
dataset van 100 patiënten, niet gebruikt tijdens de training, behaalde het
netwerk AUC scores van 93% voor WHO graad, 94% voor IDH mutatie
en 82% voor 1p/19q co-deletie predictie. We evalueerden bovendien de

classificatie performantie op een volledig onafhankelijke dataset van 110
patiënten die retrospectief werden verkregen in het Gentse Universitair
Ziekenhuis. Op deze dataset werden AUC scores gerapporteerd van
respectievelijk 94%, 86% en 87% voor de drie merkers.

De bovenstaande twee-stappen methode, bestaande uit eerst seg-
mentatie en vervolgens classificatie, heeft mogelijks enkele nadelen. Het
classificatienetwerk wordt enkel toegepast op de tumorregio, wat poten-
tieel relevante informatie over de locatie en omliggend weefsel verwijdert.
Bovendien kunnen mogelijke fouten in de eerdere segmentatie stap ook
de daaropvolgende classificatie beïnvloeden. Als alternatief ontwikkelden
we een netwerk dat gelijktijdig hersentumoren kan segmenteren en clas-
sificeren op basis van de volledige hersenscans. Het segmentatienetwerk
werd uitgebreid met een classificatietak en kreeg de naam Y-Net. Door
het gebruik van multitask leren, technieken om het geheugengebruik
op de GPU te verminderen en een adequate patch-extractie methode,
kon één netwerk worden getraind op de grote multi-institutionele en
heterogene database met ontbrekende labels voor veel patiënten. Een
vergelijkbare segmentatie performantie werd bereikt als voorheen met
gemiddelde Dice scores van respectievelijk 89%, 84% en 75% voor de
hele tumor, tumorkern en contrast-capterende regio's. De WHO graad
kon worden voorspeld met 98%, IDH-mutatie met 96% en 1p/19q co-
deletie met 87% AUC op de TCIA test dataset. Op de onafhankelijke
testdata uit het universitair ziekenhuis waren de behaalde AUC scores
96%, 83% en 90%. De classificatie performantie is iets hoger dan bij de
twee-stappen methode. Dit komt mogelijk omdat de gehele MR scan nu
wordt ingevoerd en de toevoeging van de segmentatietaak zou kunnen
zorgen voor extra regularisatie van het trainingsproces.
Ten slotte werden inzichten in de visuele kennis van het netwerk en
geëxtraheerde features verkregen met behulp van verschillende netwerk
visualisatietechnieken. De features van het netwerk werden geplot voor
elke case in de dataset na t-SNE feature reductie. Dit onthulde ver-
schillende clusters van hersentumor cases met vergelijkbare kenmerken
zoals ringvormig contrast-capterend weefsel, laesiegrootte, locatie van de
tumor in de frontale kwab en T2-FLAIR mismatch. Dit zijn inderdaad
gekende kenmerken die gecorreleerd zijn met deze tumormerkers. Sali-
ency maps, die visualiseren waar het netwerk het meeste focust in de
MR om een bepaalde voorspelling te doen, lieten zien dat het netwerk
inderdaad kijkt naar de relevante tumorregio's. Dit laat een extra con-
trole toe om vertrouwen te krijgen in de voorspellingen van het netwerk.

Ten slotte werd een synthetische input gegenereerd die de outputscores voor een glioblastoom, IDH wildtype tumor maximaliseert. Startende van een input bestaande uit willekeurige ruis verscheen na verschillende iteraties een ringvormig contrast-capterend tumorpatroon met een kern bestaande uit lagere intensiteitswaarden in het T1ce kanaal. Op de T2 kanalen ontstond omringend hyper-intens weefsel. Hieruit blijkt dat het netwerk heeft geleerd deze kenmerken toe te schrijven aan dit tumortype wat overeenkomt met bestaande kennis rond correlaties tussen tumor fenotype en tumormerkers.

Samenvattend hebben we aangetoond dat neurale netwerken het beeldvormingsproces op detectorniveau kunnen verbeteren, wat uiteindelijk resulteert in een betere beeldkwaliteit en invloed heeft op de gehele resterende radiologische workflow. Verder werd een toepassing voor beeldanalyse onderzocht, meer specifiek de karakterisering van primaire hersentumoren. Dit resulteerde in niet-invasieve en nauwkeurige algoritmes voor het segmenteren en diagnosticeren van hersentumoren. Hoewel er nog uitdagingen zijn met betrekking tot gestandaardiseerde datasets en het begrijpen van AI, zowel bij deskundigen als bij het grote publiek, kunnen we concluderen dat AI een grote impact zal hebben op radiologie. Het zal een belangrijk hulpmiddel worden voor radiologen om de efficiëntie te verhogen, routinetaken uit te voeren en gepersonaliseerde en precisie geneeskunde mogelijk te maken. Artificiële intelligentie zal echter radiologen niet vervangen, aangezien veel elementen die essentieel zijn aan het radiologisch beroep nooit kunnen worden geautomatiseerd.

List of Abbreviations

Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the receiver operating characteristic Curve
BNN	Bayesian Neural Network
BraTS	Brain Tumour Segmentation
CADe	Computer-Aided Detection
CADx	Computer-Aided Diagnosis
CE	Cross-Entropy
CNN	Convolutional Neural Network
CNS	Central Nervous System
CSF	Cerebro Spinal Fluid
CT	Computed Tomography
CTA	CT Angiography
CVD	Cardiovascular Disease
DECT	Dual-Energy CT
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DOI	Depth Of Interaction
DWI	Diffusion Weighted Imaging
EANO	European Association of Neuro-Oncology
EM	Electron Microscopy

FISH	Fluorescence In-Situ Hybridization
FLAIR	Fluid-Attenuated Inversion Recovery
FN	False Negative
FNN	Feedforward Neural Network
FOV	Field Of View
FP	False Positive
FWHM	Full Width at Half Maximum
GAN	Generative Adversarial Network
GBM	Glioblastoma Multiforme
GD	Gradient Descent
GTB	Gradient Tree Boosting
GUH	Ghent University Hospital
IDH	Isocitrate Dehydrogenase
IHC	Immunohistochemistry
LGG	Lower-grade Glioma
LOR	Line of Response
MAE	Mean Absolute Error
MC	Monte Carlo
MCC	Matthews Correlation Coefficient
MGMT	O^6-methylguanine-DNA methyltransferase
ML	Machine Learning
MNIST	Modified National Institute of Standards and Technology
MRA	MR Angiography
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NIfTI	Neuroimaging Informatics Technology Initiative
NLL	Negative Log-Likelihood
NMR	Nuclear Magnetic Resonance
NPV	Negative Predictive Value
PBT	Primary Brain Tumour

PDE	Photon Detection Efficiency
PET	Positron Emission Tomography
PMT	Photomultiplier Tube
PPV	Positive Predictive Value
PSF	Point Spread Function
PWI	Perfusion Weighted Imaging
ReLU	Rectified Linear Unit
RF	Random Forest
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
SGD	Stochastic Gradient Descent
SiPM	Silicon Photomultiplier
SNR	Signal to Noise Ratio
SPECT	Single Photon Emission Computed Tomography
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Archive
TN	True Negative
TOF	Time-OF-Flight
TP	True Positive
WHO	World Health Organisation

Contents

1 | Introduction

1.1 Context

This dissertation involves the application of artificial intelligence (AI) in medical imaging. Although the idea of intelligent machines and artificial neural networks has been around for decades, AI has only recently known an unprecedented growth. The required computational power and large amounts of data are only recently available to bring AI into practice. Artificial intelligence is being applied in numerous industries and entering our daily lives in the form of voice assistants, face recognition, recommendation systems in advertising and entertainment (e.g. social media, Spotify and Netflix), intelligent driver assistance, household robots etc. These systems achieve incredible performances that match or even outperform humans.

Driven by the increasing digitisation of healthcare, AI is also finding its way into the healthcare industry. Wearables measure more and more health related data and the amount of lab tests, DNA analyses, treatment results and medical imaging is rapidly expanding. This data contains a wealth of information that presents opportunities to improve and personalise healthcare, but it is becoming increasingly difficult to efficiently and fully exploit all this intelligence. In medical imaging, the radiology workforce is struggling to meet the rising demand for imaging examinations which can lead to delayed diagnoses and potentially affect accuracy [1]. This rising demand also puts pressure on the throughput of medical image acquisition with imaging technology that continuously advances in resolution and quality and becomes more complex, multimodal, 3D and dynamic. Hence there is also a need for AI systems in healthcare to improve efficiency and enable personalised and precision medicine.

Even though there is a lot of interest and enthusiasm on the development and application of AI, it is also feared. Concerns range between AI taking over jobs to robots taking over the world. In healthcare there are concerns on AI ethics, trustworthiness and interpretability, responsibility and legal aspects, privacy and security etc. News reports appear that announce the extinction of radiologists and numerous other professions. It is clear the a good understanding of this technology and debate on these questions is of utmost importance as it is not a matter of *if* but *how* AI will influence our society.

Current AI technology is still very narrow meaning that an AI algorithm is able to perform only one specific task. The algorithm can reach high performances for this task, but for every task, different AI systems need to be developed. General AI that exhibits human like intelligence is still far from being a reality. In medical imaging, many therefore believe that AI will help to perform routine tasks and increase efficiency of the radiological workflow. But it will remain the doctor who ultimately decides as many aspects to the radiological profession can never be performed by a machine regarding expertise, human attitude, empathy, mutual understanding, family situation and support etc.

This PhD dissertation focuses on the use of AI in medical imaging. AI can be applied throughout the entire medical imaging pipeline: from acquisition, image enhancement and post-processing to image analysis including detection, segmentation and diagnosis of diseases. Under the influence of recent advances in computer vision, a lot of research is situated on the use of AI for analysing medical images [2]. Algorithms based on deep learning reach performances that exceed humans in numerous tasks such as pneumonia detection on chest X-ray, breast cancer mammography screening, analysing skin lesions etc. The use of AI in medical imaging remains, however, challenging as the amount and size of curated datasets is still limited, certainly when compared to natural imaging datasets employed in computer vision. Data is scattered across clinical centres with high heterogeneity is imaging protocols, quality, recorded modalities and annotations. Curation of medical imaging datasets is time consuming and requires expert knowledge. Medical images are also very complex with large 3D volumes, many different modalities with differing characteristics and complex, highly variable healthy and pathological structures. Moreover, as these systems are used in critical settings that can have a direct influence on diagnosis and treatment planning, they need to be trustworthy and interpretable.

Two different applications of AI in medical imaging are explored in this work. One is situated in the acquisition stage, more specifically on improving the spatial resolution of positron emission tomography (PET) detectors. The second application is located at the end of the medical imaging pipeline, on the analysis of pre-therapy brain MRI for automatic segmentation and diagnosis of primary brain tumours.

The purpose of a PET detector is to detect gamma rays emitted by a radioactive tracer that is injected into the body that is imaged. It is important to determine the exact position of arrival of the gamma ray within the PET detector as this has a direct influence on the spatial resolution of the final image. In this dissertation, we investigate the use of artificial neural networks to accurately determine the gamma ray arrival position from the electronic signal measured by the detector. Positioning accuracy is evaluated in a simulation and experimental setup and compared with an established positioning algorithm called mean nearest neighbour.

Primary brain tumours (PBTs) are a complex type of neoplasms. They are often difficult to treat and associated with low survival rates, depending on tumour type, as the brain is a complex and a vital organ itself. Recent guidelines of the World Health Organisation (WHO) have put increased emphasis on the integration of molecular markers to differentiate primary brain tumours [3]. These markers have prognostic value and enable better and more personalised therapy planning. In order to determine these molecular markers, invasive procedures are required such as biopsy or resection that involve risks. These surgical procedures are not always possible depending on the tumour location, the patient's clinical condition or when the patient refuses surgery. Being able to determine tumour characteristics non-invasively based on pre-therapy MRI can therefore be beneficial for initial prognosis and therapy planning. Many studies have already reported correlations between tumour phenotype (appearance on MRI) and genetic markers [4–8]. AI algorithms can learn to fully exploit these patterns in medical images and perform an accurate diagnosis. Existing AI systems are, however, often not fully automatic (require manual delineation of the tumour) and trained and evaluated on a single small dataset. Hence their robustness to data from different centres remains to be validated.

In this work, deep learning models are developed for non-invasive and fully automatic segmentation and classification of primary brain tumours. Moreover, their generalisation capacity to data from different

institutions is evaluated.

1.2 Outline

In what follows, an overview is provided of the structure of this dissertation.

As artificial intelligence is an essential topic in this work, **chapter 2** provides a thorough explanation of important AI, machine learning and deep learning concepts that are necessary to understand the remaining part of this book. The chapter focuses on artificial neural networks and especially convolutional neural networks as this is one of the most common AI algorithms in medical imaging.
Chapter 3, covers the role of AI in medical imaging. The need, potential and challenges of AI in healthcare are discussed followed with a brief overview of the most common medical imaging modalities. Positron emission tomography is explained in more detail as this is relevant for the first main topic of this work. Finally, an overview is provided of state-of-the-art applications of AI throughout the entire medical imaging pipeline.

The rest of this book is divided into two parts: one on PET detector calibration (three chapters) and a second on computer-aided brain tumour diagnosis (five chapters).

In **chapter 4** the use of neural networks to determine the gamma ray arrival position in a PET detector is investigated using simulation data. The optimal architecture, amount of training data and training procedure is evaluated and potential pitfalls related to the training and evaluation of neural networks are identified and addressed. Furthermore, the positioning performance is directly compared to mean nearest neighbour positioning which is an established algorithm for gamma ray positioning in PET detectors.
There are several factors that can potentially degrade the gamma ray positioning accuracy of neural networks. **Chapter 5** will explore two different factors being Compton scattering of the gamma ray inside the crystal and precision of the calibration beam used to acquire training data. Their influence on the positioning accuracy is evaluated and, if necessary, different techniques are explored that could help mitigate these effects.

The methodology of training neural networks for positioning of gamma rays developed in chapters 4 and 5 will be validated on experimental data in **chapter 6**. Two different detector setups will be evaluated and performance will again be compared with mean nearest neighbour positioning.

The second part of this book on computer-aided brain tumour diagnosis starts with an introductory chapter, **chapter 7**, providing some background information on neuroanatomy, nomenclature on different brain tumour types and markers as defined by the WHO and their relation between prognosis and optimal therapy. Furthermore, a literature review is included on recent work on brain tumour segmentation and classification with AI.

In **chapter 8**, the important task of determining brain tumour grade is investigated as tumour malignancy is highly predictive for prognosis and optimal therapy planning. The predictive performance of more traditionally used hand-engineered imaging features is compared with deep learning features extracted with a convolutional neural network, pre-trained on natural images. Additionally the effect of the provided region to the network (only tumour region of interest or full MRI) on the classification accuracy is examined.

As brain tumour segmentation is important in the diagnosis and management of primary brain tumours and is often a required pre-processing step before classification, an automatic segmentation algorithm is developed in **chapter 9**. Automatic and accurate delineation of brain tumour tissues is necessary as manual segmentation is time-consuming and suffers from inter- and intra-reader variability. Additionally, the segmentation algorithm is made robust to missing input MRI modalities as often not all MRI sequences are available in clinical practice and in the dataset that is collected and used in the subsequent chapters.

In **chapter 10**, a convolutional network will be trained from scratch that not only predicts tumour grade but also important molecular markers according to the most recent WHO guidelines. To this end, a large dataset is collected from multiple public databases. The segmentation algorithm from chapter 9 is applied to these clinical scans to extract the tumour region of interest which is fed into the classification network. An additional independent dataset acquired at the Ghent University Hospital is used to evaluate the generalisation performance to data from different centres.

The two-stage approach proposed in chapter 10 (segmentation followed

by classification) can have some downsides as the classification network only operates on the tumour region of interest which excludes potentially relevant information on location and surrounding tissue. Moreover, possible errors in the prior segmentation step can also influence the subsequent classification performance. As an alternative, a network that performs simultaneous segmentation and classification based on the full brain MRI is explored in **chapter 11**. To train such a network different techniques are used to deal with limited GPU memory, data heterogeneity and missing labels. The segmentation and classification performance of this approach is compared with the two-stage pipeline in chapter 10 on the same data. Furthermore, several visualisation techniques are implemented to gain insights into the imaging features that are automatically learned by the network.

Finally, **chapter 12** concludes this dissertation and discusses some future work and research directions.

2 | Artificial intelligence

The purpose of this chapter is to provide an introduction to artificial intelligence and its different terminology such as Machine Learning (ML), Deep Learning (DL), narrow versus general AI etc. Necessary machine learning and deep learning concepts are explained in more detail to support subsequent chapters and describe current state-of-the-art AI technology applied in medical imaging. I refer the reader to the following works for an in-depth review on artificial intelligence [9], machine learning [10, 11] and deep learning [12].

2.1 Introduction

So what is Artificial Intelligence? The term Artificial Intelligence was first coined in 1955 by John McCarthy in a conference proposal for "a study to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" [13]. Modern dictionaries such as the Oxford english dictionary state: "The capacity of computers or other machines to exhibit or simulate intelligent behaviour." However, what is intelligent behaviour or intelligence in general? On this day there is still no clear definition of AI as defining intelligence itself is already very difficult and subject to a lot of philosophical discussion.

An operational definition of intelligence was proposed by Alan Turing in 1950 [14]. The imitation game, more famously known as the Turing Test, tries to evaluate whether a machine can think. A computer passes as intelligent if a human interrogator cannot tell whether the responses come from a person or not. Although the value of the Turing Test

has been debated ever since, it is still relevant today and remains an important milestone to reach for many years to come.

This shows that the idea of intelligent machines has been around for decades. Yet why is AI only recently growing so rapidly and finding its way into society? Throughout the years AI research has known a lot of successes but also many disappointments. Overestimation of AI's progress and failure to meet excessive promises from the early days led to a rapid decline in funding for AI research and a so called 'AI winter'. AI continued to show some progress under different names such as 'machine learning' or 'pattern recognition', however, commercial AI industry advances were very limited.
The recent revival of AI research in the 21st century is driven by two key elements. The first and most important factor is the exponential increase in computational power and data storage, also known as Moore's law. Computers are getting faster, smaller and more affordable and thereby accessible to everyone. This in turn leads to an increasing digitisation of our world and an explosion in data which is the second requirement to bring AI technology invented many years ago, such as neural networks, finally into practice. Complex tasks like image analysis or natural language processing require complex algorithms not only resulting in a need for fast processors to execute these huge number of calculations but also for a lot of data to discover patterns and learn to perform these tasks.

These two elements proved to be key to lead AI into a new summer, rapidly progressing both in academia as in industry. It is increasingly playing a role into our daily lives with applications in speech recognition (voice assistants such as Siri and Alexa), image recognition (face recognition, self-driving cars...), recommendation systems in retail and entertainment and even analysis of healthcare data for faster and more accurate diagnoses. AI will advance numerous industries and is anticipated to be the key driver of the fourth industrial revolution but is also seen as a potential threat to job security and human society incited by futuristic science-fiction movies where robots take over the world. It is clear that a good understanding of this technology is important to facilitate debate on ethical questions not defining if but how it will influence our society. This is especially important in healthcare where AI can have a direct influence on people's lives and possibly survival. In 2016, several industry leaders including Google, Facebook, Apple and Microsoft joined together in a partnership to formulate best practices on AI technology and advance the public's understanding of AI [15].

One often makes the distinction between narrow and general AI. Narrow AI is trained to perform one specific task and is what we currently find in our devices. For every task (recognising objects in images, speech recognition, playing chess...) a different algorithm is trained. General AI on the other hand is able to perform a lot of different tasks and exhibits human like intelligence. While a lot of narrow AI applications with impressive performance are developed today, general AI is still far from being a reality.

Understanding AI terminology can be difficult as it is often interchangeably used with Machine Learning and Deep Learning although they do not refer to the same thing. Where AI is the most general term and encompasses any technique to bring intelligence to a machine, machine learning and deep learning subsequently cover more specific types of AI as illustrated in figure 2.1. Machine learning algorithms allow a computer to learn how to perform a task without explicitly being programmed [16]. In other words ML is an approach to develop AI systems. One type of ML is a network of simple connected processing units or neurons often organised in layers. When these neural networks contain enough layers (typically more than three) and neurons one talks about deep learning.
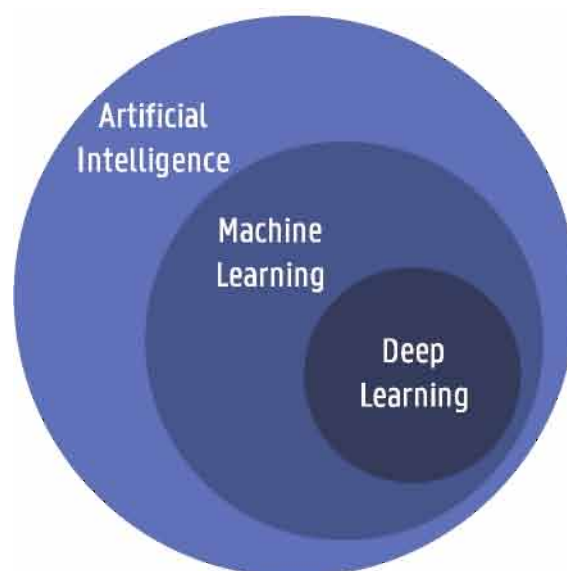
Figure 2.1: Diagram illustrating relation between different AI terminology.

2.2   Machine learning

Machine learning is programming computers to perform a certain task using example data or past experience. This is useful (or even necessary) in cases where humans are unable to precisely explain their expertise (for example face recognition) or when algorithms constantly need to be adapted (changes in time or users). Figure 2.2 shows a schematic overview of different ML components illustrated with a brain tumour detection example. A model, defined up to some parameters, receives brain MRI as **input** and needs to provide as **output** whether the brain scan shows a tumour or not. Based on **example data**, i.e. labelled brain MRI, a **learning algorithm** optimises the model parameters to improve a certain **performance measure**. When training is finished and the model achieves sufficient performance, it can be used to detect tumours in new MRI.
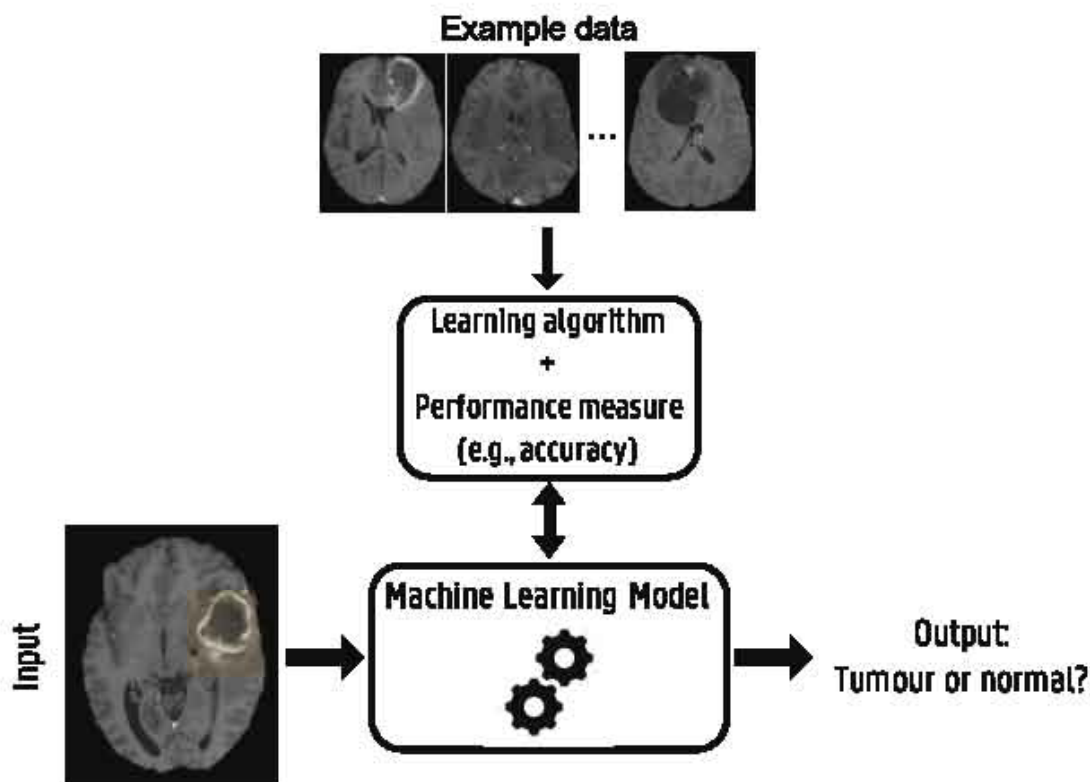
Figure 2.2: Schematic overview of different machine learning components and their interaction.

In the following sections, we will discuss the different types of machine learning and explain the elemental ML algorithms: linear and logistic

regression which form the foundation to understand neural networks described in section 2.3. Furthermore sections 2.2.4 and 2.2.5 will discuss performance evaluation and the main challenge of machine learning: generalisation. Finally, the random forest classification algorithm is explained in section 2.2.6 as this classifier will be used in chapter 8.

2.2.1 Types of machine learning

Based on the type of example data and available information, different types of machine learning can be defined.

Supervised learning

In supervised learning, the most common type of machine learning, example data consists of known input-output pairs. Labelled data is available and the model is trained such that its output is as close as possible to the desired label for every input. After training, the model can be applied to new unlabelled input data. Supervised learning can further be categorised into classification and regression. In case of classification, the output consists of a discrete set of output values or classes as in the brain tumour detection example in figure 2.2. In case of regression, the output is a continuous variable. For example, house price prediction or survival time estimation.

Unsupervised learning

The second type of ML is unsupervised learning, where no output labels are available. The aim is to find hidden structure in the example input data. One method is clustering into different groups of similar inputs. For example, tissue segmentation in medical images where pixels with similar intensities are grouped together.

Reinforcement learning

The final type of learning is often used in game playing or robot control and is called reinforcement learning. The goal now is for an artificial agent to learn a policy on which actions to take in an environment in order to reach a certain goal or maximise a cumulative reward. Hence

there is not one sequence of best actions or in any intermediate state there is not one best action but an action is good if it is part of a good policy that in the end leads to a maximal reward. The agent explores the environment and possible actions using trial and error. Based on past good action sequences, the machine learning algorithm should be able to learn a good policy.

2.2.2 Linear regression

An example of a supervised regression ML algorithm is linear regression. We will start with explaining the simplest case, univariate linear regression, using with a well known house price prediction example. The goal is to train a model, in this case a univariate linear function, that allows us to predict the price of a house based on its size (in squared metres). A univariate linear function (i.e. a straight line) has the following form:

$$\hat{y} = w_1 x + w_0$$

where x is the input of the model (the size of the house), $\hat{y}$ is de output (the house price) and w_1 (the slope) and w_0 (the intercept) are called the weights of the model.

To train the house prediction model we use a training dataset containing input-output pairs or training samples (x,y). In figure 2.3, an example training dataset is plotted. The goal is to find the line that best fits the training samples, i.e. the optimal weights w_1 and w_0 that minimise the distance between each training sample and the line. In order to evaluate how well a line fits to the data, a loss or cost function needs to be defined. Often the Mean Squared Error (MSE) loss is used defined as

$$\text{MSE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

with N the number of data samples, y the ground truth label and $\hat{y}$ the predicted output of the model.
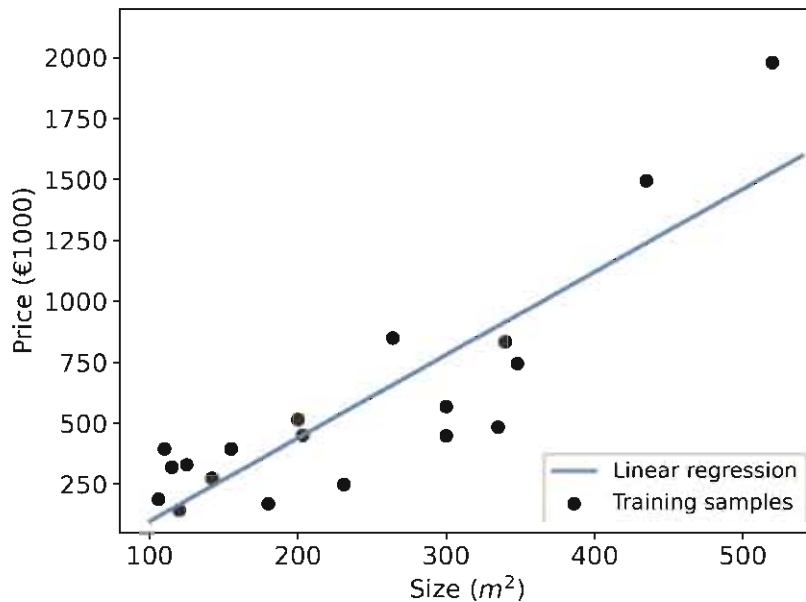
Figure 2.3: Linear regression illustrated with a house price prediction example. Training samples are plotted of size versus house price along with the linear function that corresponds with a minimal squared error loss.

Gradient descent

The optimal weights can be found using an iterative optimisation algorithm called gradient descent (GD). This algorithm initialises the model with randomly chosen weights and iteratively updates them in a direction that minimises the loss based on the dataset. This direction is given by the negative partial derivative or gradient of the loss with respect to the weights. For linear regression, the gradient of the MSE loss with respect to the weight w_i is calculated as follows:

$$\frac{\partial \text{MSE}}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{N} \sum_{i=1}^{N} ((w_1 x_i + w_0) - y_i)^2$$

$$= \frac{2}{N} \sum_{i=1}^{N} ((w_1 x_i + w_0) - y_i) \frac{\partial}{\partial w_i} ((w_1 x_i + w_0) - y_i)$$

For w_1 and w_0 this results in

$$\frac{\partial \text{MSE}}{\partial w_1} = \frac{2}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i) x_i$$

$$\frac{\partial \text{MSE}}{\partial w_0} = \frac{2}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)$$

and the weights are updated as

$$w_1 = w_1 - \alpha \frac{2}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i) x_i$$

$$w_0 = w_0 - \alpha \frac{2}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)$$

where α is defined as the learning rate or step size by which the weights are updated. The learning rate is an important hyperparameter of gradient descent. Setting the learning rate too high can prevent proper convergence by overshooting the minimum. A learning rate that is too low results in very slow convergence. Each iteration, the weights can be updated based on the entire training dataset (batch gradient descent), only on a small random subset or mini-batch (mini-batch gradient descent) or only on a single sample (stochastic gradient descent, SGD). Mostly mini-batch gradient descent is used as calculating the gradient on the entire dataset is computationally inefficient. One iteration over the entire training set (either in one step or in multiple steps) is often called an epoch.

Multivariate regression

The univariate linear regression example can easily be extended to a multivariate case where we have more than one input feature. For example, next to the size of the house, the energy performance, distance to shops, number of bedrooms, etc., can be important to accurately determine the price. We then have

$$\hat{y} = w_0 + \sum_{i=1}^{M} w_i x_i$$

with M the number of input features. Defining the input value x_0 as always equal to 1 we can write this as the product of the weight and input vectors $\mathbf{w}$ and $\mathbf{x}$:

$$\hat{y} = \mathbf{w}^T\mathbf{x}$$

Different features can be expressed in various units and thereby have large differences in range. The size of a house, for example, ranges between 100 to 500 m^2 in figure 2.3 while the number of bedroom would typically vary between 1 to 5. It is important to address these differences in range to improve the convergence speed of gradient descent and reduce oscillation when features are highly uneven. The features can be scaled to a fixed range, typically between 0 and 1. An other approach is standardisation to zero mean and unit variance. These statistics are calculated across the samples in the training dataset.

Higher order regression

Linear regression can also be extended to higher order polynomial regression. In the univariate case this results in

$$\hat{y} = w_0 + \sum_{p=1}^{P} w_p x^p$$

with p the order of the polynomial. The multivariate case additionally includes product terms of different features. For example, with two input features x_1 and x_2, we can define

$$z_1 = x_1, z_2 = x_2, z_3 = x_1^2, z_4 = x_2^2, z_5 = x_1 x_2$$

and apply linear regression on the five dimensional input $\mathbf{z} = [z_1, z_2, z_3, z_4, z_5]$. Hence, instead of defining a nonlinear function in the original space, a nonlinear transformation to a new space can be applied where we define a linear function. The input space can be extended to any order to better fit the training data. High order polynomials, however, might better fit the individual training samples, but not capture the inherent relation between input and output. The model complexity should carefully be tuned to match the complexity of the function underlying the data. This will be discussed in more detail in section 2.2.5 on generalisation.

2.2.3 Logistic regression

Unlike the name suggest, logistic regression is used for classification problems where the output value is categorical. In case of binary classification, the output is limited to values 0 or 1 instead of any continuous value as with linear regression. This is solved by introducing the sigmoid function to the linear regression model. The sigmoid function defined as

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

transforms the output to a value between 0 and 1 which can be interpreted as a probability (see figure 2.4). Hence the probability of an input sample $\mathbf{x}$ belonging to class 1 can be written as:

$$P(C_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$$

A threshold, usually 0.5, is then defined to determine whether the sample belongs to a certain class.
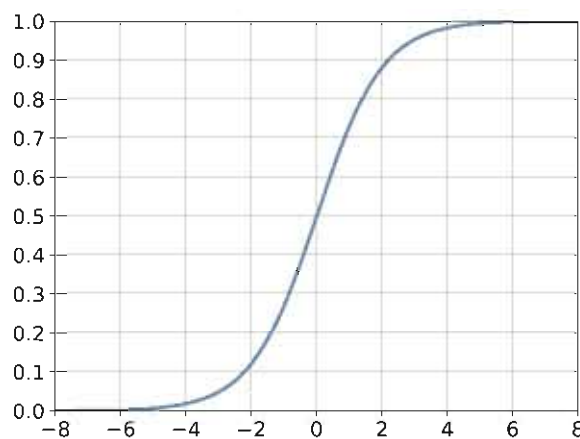
Figure 2.4: The sigmoid function.

Similarly to linear regression, gradient descent can be used to find the optimal weights of the model. Now a different loss function is used, namely the negative log-likelihood (NLL) or cross-entropy loss (CE) defined as:

$$CE(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

The binary classification example can be extended to multiple classes by using the softmax function instead of sigmoid:

$$P(C_j|\mathbf{x}) = \text{softmax}(\mathbf{w}_j^T \mathbf{x}) = \frac{e^{\mathbf{w}_j^T \mathbf{x}}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^T \mathbf{x}}}, \quad j = 1, ..., K$$

This way, the sum of the probabilities of all classes equals to one.

2.2.4 Performance evaluation

To evaluate how well a machine learning model performs, various metrics are used that compare the model's predictions to the ground truth.

For regression tasks, distance measure are used such as mean squared error or mean absolute error (MAE):

$$MAE(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|$$

In case of classification, most metrics are derived from the confusion matrix shown in table 2.1 for binary classification. A prediction can be identified as a true positive TP or a true negative TN when it correctly indicates that the sample belong to the 'positive' or 'negative' class respectively. When the prediction is wrong, it is identified as a false positive FP or false negative FN.

Table 2.1: Confusion matrix of a binary classifier.

		Ground truth	
		Positive	Negative
Predicted	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

The most well known classification metric is accuracy:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Accuracy is not always a reliable performance measure, especially for unbalanced classes where one class occurs much more often than the other. For instance, the case of classifying tumours as benign or malignant where 90% of the tumour samples in the data are benign and only 10% malignant. When the model always predicts benign, it achieves a high accuracy of 90%. One would think that the model performs very well while it actually misses all malignant tumours. In case of class imbalance, it is important to determine both the sensitivity (percentage of correctly identified positive samples) and specificity (percentage of correctly identified negative samples) of the model:

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{FP + TN}$$

Similarly one can define the positive predictive value (PPV) or precision and the negative predictive value (NPV) as the fraction of true positive or negative samples over all samples that are predicted as positive or negative.

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{FN + TN}$$

The harmonic average between sensitivity and precision is defined as the f1 score, also known as Sørensen-Dice coefficient or Dice score.

$$f1 = \frac{2 \times precision \times sensitivity}{precision + sensitivity} = \frac{2TP}{2TP + FP + FN}$$

The Dice score can also be interpreted as a measure of overlap between two sets and is often used to evaluate segmentation algorithms. A balanced measure to evaluate a binary classifier, even if the classes are of very different sizes, is Matthews correlation coefficient (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The metric ranges between -1 and 1, where -1 indicates total disagreement (reversed prediction), 0 indicates no better performance than random guessing and +1 indicates perfect prediction.
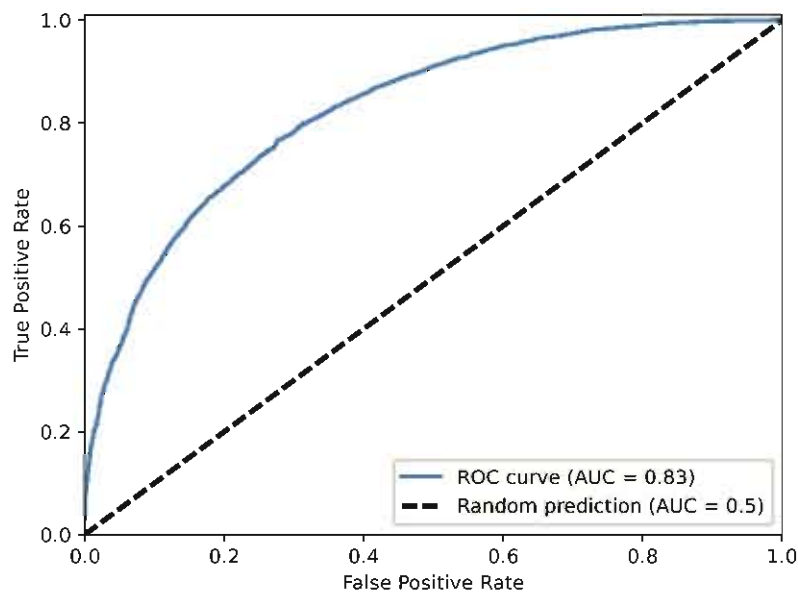
Figure 2.5: Example ROC curve.

The above measures use the predicted class labels after thresholding the output probabilities and are therefore dependent on the chosen threshold. An aggregate measure that evaluates performance across the entire range of possible thresholds is the area under the receiver operating characteristic (ROC) curve or simply AUC. The ROC curve plots the true positive rate versus the false positive rate for different classification thresholds and is illustrated in figure 2.5. Predictions that are no better than chance result in a diagonal ROC curve corresponding with an AUC of 0.5. An AUC of 1 indicates 100% correct predictions and an AUC of 0 indicates 100% wrong predictions. Hence classifiers that perform well have an ROC curve closer to the upper left corner with an AUC closer to 1.

2.2.5   Generalisation

The main challenge of machine learning is to train a model that performs
well on new, unseen data. This is called generalisation. To assess the
generalisation performance of a model, the available dataset is typically
split into a train, validation and test set. The training set is used to
optimise the model weights. During training we minimise a certain error
measure calculated on the training set, called the training error. The
validation set is used to evaluate the generalisation performance of the
model during training. Hence no weights of the model are optimised
using validation data but the number of training iterations, model com-
plexity and hyperparameters are tuned to minimise the validation error.
After training the model is finally evaluated on the test dataset to assess
the predictive power on unseen samples.
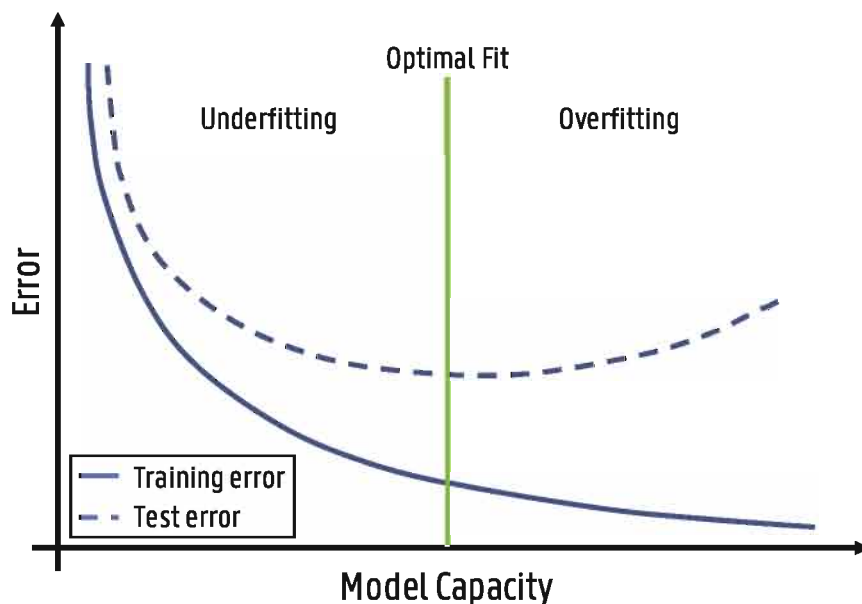
Figure 2.6: Illustration of typical relationship between model capacity and
training and test error. At low capacity both training and
test error are high, the model is underfitting. When capacity
increases, the error decreases but the gap between training and
test error broadens. Eventually this gap becomes too large and
test error starts to increase while the train error keeps decreasing
i.e. the model is overfitting.

In order for a machine learning model to perform well, it needs to
have a low training error and a low test error i.e. a small gap between
training and test error. This corresponds with the two key challenges in

machine learning: underfitting and overfitting. When a model is not able to sufficiently reduce the training error, it is underfitting. Overfitting occurs when there is a large difference between the training and test error. The balance between under- and overfitting can be controlled by adjusting the model capacity as illustrated in figure 2.6. A model that does not have enough capacity results in a high training and test error. When the capacity increases, both errors decrease but the gap between training and test error broadens. Eventually this gap will become too large and the test error will start to increase while the train error keeps decreasing i.e. the model is overfitting.

We illustrate this with the house price prediction example of section 2.2.2. The linear function fitted in figure 2.3 might be too simple to properly fit the underlying trend where the price appears to rise faster for larger houses. We can increase the model's capacity by fitting a polynomial instead of a linear function. Figure 2.7a shows a polynomial of order 3 fitted to the training samples. This regression line is a better fit to the samples than the linear function of figure 2.3 which was too simple and underfitting. If we further increase the model capacity to a polynomial of order 9, see figure 2.7b, we clearly observe that the model is overfitting. The polynomial better fits the individual training samples but is very erratic and does not follow the expected underlying relation between house size and price, especially the peak around a size of 500 m^2.
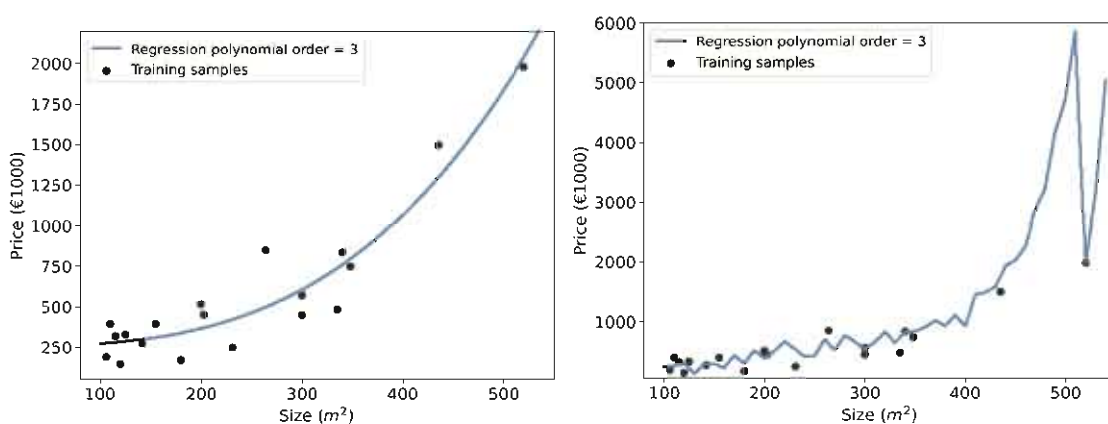
(a) Polynomial regression with order 3.   (b) Polynomial regression with order 9.

Figure 2.7: Polynomial regression illustrated with a house price prediction example. Training samples are plotted of size versus house price along with the fitted polynomial curve.

Regularisation

We have seen in previous section that the amount of under- and overfitting can be controlled by altering the capacity or number of parameters of the model. However, instead of changing the variety of functions that the model can represent, we can also incorporate a preference towards certain functions to limit the amount of overfitting. This is called regularisation or *"Regularisation is any modification we make to a learning algorithm that is intended to reduce its generalisation error but not its training error"* [12].

For example, we can modify the loss function J of linear regression to express a preference towards smaller weights by adding an extra term that penalises weights with high squared L2 norm:

$$J = \text{MSE} + \lambda \mathbf{w}^T \mathbf{w}$$

where λ is a hyperparameter controlling the strength of regularisation. This type of regularisation, called weight decay, allows to obtain solutions that put weight on fewer number of features or that have a smaller slope.

We illustrate this again with the house price prediction example. The polynomial regression with order 9 of figure 2.7b is repeated with weight decay in figure 2.8. The regularisation strength was set to $\lambda = 1.0$. The benefit of regularisation is clearly visible as the fit is now much smoother and more similar to the polynomial in figure 2.7a.

Weight decay is only one example of the many regularisation techniques and ways to control overfitting. Throughout section 2.3.1 several other regularisation techniques will be discussed that are applicable to deep learning.

2.2.6   Random forest

In this section, we will explain one of the most popular classifiers: random forest (RF). This classifier will be used in chapter 8 to classify brain tumours as high grade or lower grade.
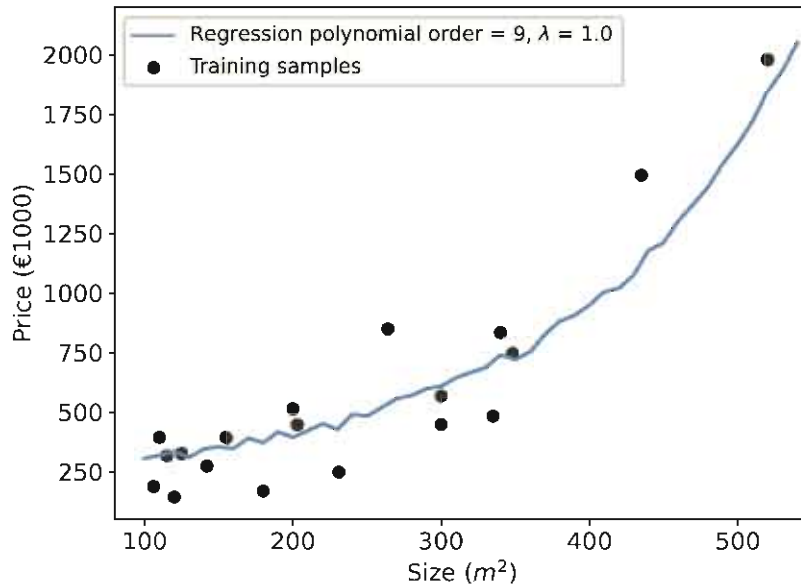
Figure 2.8: Polynomial regression illustrated with a house price prediction example. Training samples are plotted of size versus house price along with the fitted polynomial curve of order 9. Weight decay with strength $\lambda = 1.0$ was added to the loss function to prevent overfitting.

Decision trees

The basic building block of a random forest classifier is a decision tree. Decision trees are hierarchical models consisting of internal decision nodes, branches and finally leaf nodes. At each node, starting from the root node, a certain test is applied (e.g. is the size of the tumour larger or smaller than a certain threshold) and one of the branches is taken depending on the outcome. This process is recursively repeated until arriving at a leaf node giving the final output. Decision trees have the advantage that they can be visualised and are interpretable as they can be written as a set of *if-then* rules. For this reason they are applied to a broad range of tasks including medical applications such as diagnosis.

The goal is thus to build a decision tree that optimally classifies example data with a minimal amount of nodes. This is typically done using a greedy top-down procedure. Starting from the root node, the training data is iteratively split into smaller and smaller subsets. At each node, the best split needs to be found. In other words, the attribute has to be identified that makes the data in the child nodes as 'pure'

as possible i.e. containing data that belongs to just one class. One
measure that is often used to calculate impurity is entropy. The entropy
or impurity at node m is calculated as:

$$i(m) = -\sum_{k=1}^{K} p_m^k \, \log_2 \, p_m^k$$

where p_m^k is the fraction of samples arriving at node m that belong
to class k. Hence at each node, the split is selected that maximally
decreases the entropy. This process is repeated until the nodes are pure
or if certain stopping criteria are met such as maximum depth, minimal
number of samples per leaf etc. Learning a tree that classifies the training
data perfectly may lead to overfitting due to noise in the data and poor
decisions towards the leaves as they are based on a small number of
samples. For this reason, growing is often stopped when reaching a
sufficient purity or when there is no longer sufficient data to reliably
split the nodes. This is called prepruning. Another technique, called
postpruning, removes subtrees that do not have sufficient evidence after
the tree is fully grown. The output of a leaf node can be the class label of
the majority class or a probability calculated as the fraction of samples
in the node that belong to the majority class.

Random forest

Random forests make use of a bagging (bootstrap aggregating) technique
to create an ensemble of a lot of decision trees [17]. The idea is that the
error probability of a combination of many weak learners, in this case de-
cision trees, is lower than the error probability of the individual learners
and this way overfitting can be reduced. The principle of a random forest
classifier using bagging is illustrated in figure 2.9. Every tree is grown
on a subset of the training data randomly sampled with replacement to
introduce randomness between the different trees. Additionally, at every
node, the best split is chosen based on a random subset of the input
features. The outputs from all trees are aggregated through majority
voting to determine the final output class.
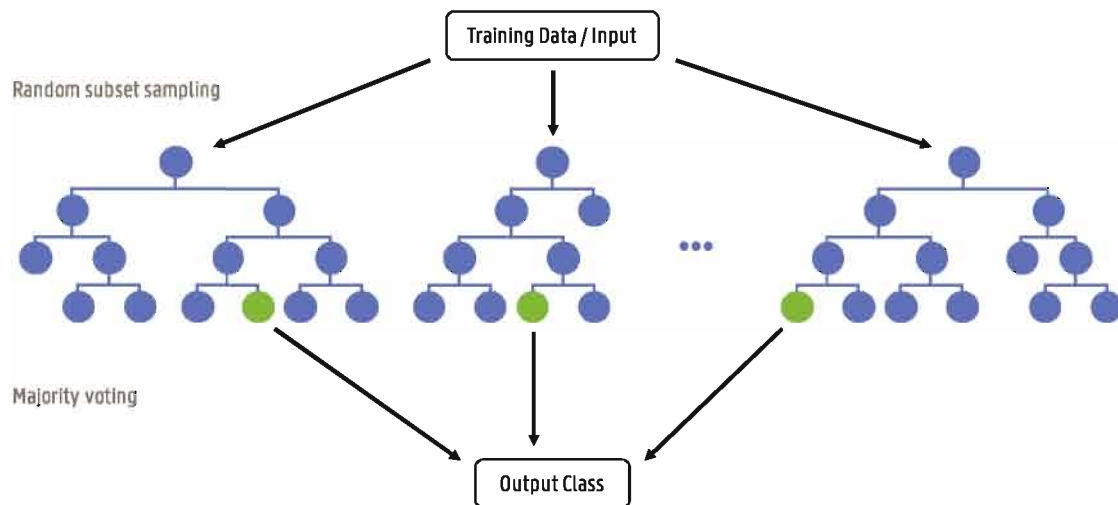
Figure 2.9: Illustration of a random forest classifier. Every tree is grown based on a random subset of the training data. When evaluating an input, it is propagated through every tree in the forest and the final output class is decided using majority voting.

2.3 Deep learning

In this section we will explain a type of machine learning algorithm inspired by the biological functioning of the brain: artificial neural networks (ANN). Starting from the basic building block, the artificial neuron, the principle behind feedforward neural networks (FNN) will be described. Driven by the rapid increase in computational power and amount of data, these neural networks became increasingly complex which brings us to the domain of deep learning. We will describe how these complex networks can be trained, including several regularisation techniques to tackle the challenge of overfitting. Afterwards, we will explain a type of neural networks specialised to process imaging data. Their design and theoretical background will be discussed together with several example architectures that play a key role in computer vision and in this work.

2.3.1 Artificial neural networks

Artificial neuron

In 1943, Warren McCulloch and Walter Pitts proposed a model of artificial neurons based on physiological function of neurons in the brain

where each neuron has two states "on" or "off" [18]. They showed that all logical operators (and, or, not etc.) can be implemented with a simple network of connected neurons and that any computable function can be represented by a network of connected neurons [9].
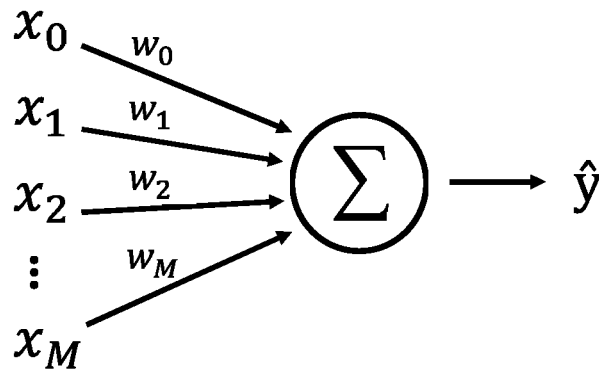
Figure 2.10: Artificial neuron or perceptron where the output $\hat{y}$ is calculated as a weighted sum of input $\mathbf{x} = [x_0, x_1, ..., x_M]$ and weights $\mathbf{w} = [w_0, w_1, ..., w_M]$.

An artificial neuron or perceptron with input $\mathbf{x} = [x_0, x_1, ..., x_M]$. and output $\hat{y}$ is illustrated in figure 2.10 [19]. The output is calculated as a weighted sum of the neuron's inputs where mostly $x_0 = 1$, called the bias term. This brings us back to the linear regression equation $\hat{y} = \mathbf{w}^T \mathbf{x}$ of section 2.2.2. Hence a neuron is just a different graphical representation of linear regression and al the theories and concepts explained in sections 2.2.2 and 2.2.3 are applicable to neurons as well. Similar to section 2.2.3 on logistic regression, the sigmoid or softmax functions can be added to the output for classification tasks. These functions are typically called activation functions.

Feedforward network

Multiple neurons can be connected to form a neural network. When the network only has connections in one direction, it is called a feedforward neural network or multilayer perceptron. When loops are added to the network, i.e. neuron outputs are fed back into its inputs, we are dealing with recurrent neural networks. They are especially suited to process sequential data, for example in speech recognition. In this work, however, only feed forward neural networks are discussed.
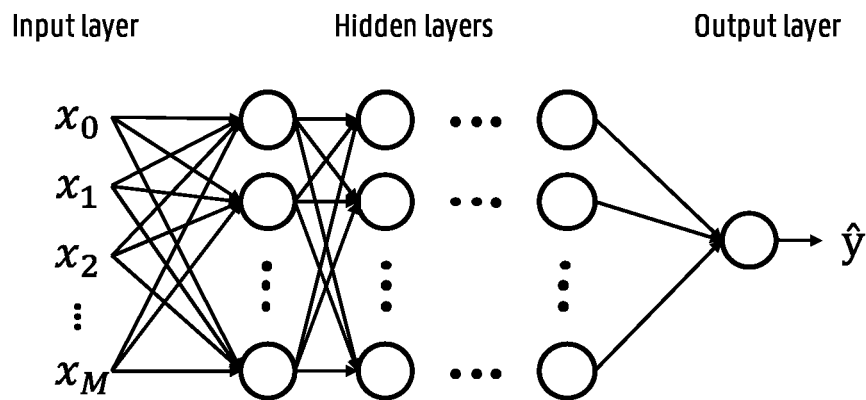
Figure 2.11: Feedforward neural network arranged in input layer, hidden layers and output layer.

Feedforward networks are typically arranged in layers starting with the input layer followed by one or several hidden layers and finally the output layer as illustrated in figure 2.11. The number of hidden layers determines the depth of the network. Hence the term deep learning for networks with many hidden layers. Figure 2.11 shows a network with one output value, however multiple output values can the estimated by adding additional output neurons.

The power of feedforward neural networks is shown by the universal approximation theorem. This theorem states that feedforward networks with at least one hidden layer (with any nonlinear activation, see next section) and a linear output layer can approximate any continuous function [20, 21]. In other words, a feedforward neural network with just one hidden layer is sufficient to represent any function. It is, however, not guaranteed that the optimisation algorithm will be able to learn that function as the hidden layer may be too large and fail to generalise. It is often beneficial to use deeper models with less neurons per layer to achieve a better generalisation error.

Activation functions

To model nonlinear functions, nonlinear activation functions need to be added after every hidden layer. Otherwise, the network would perform a linear combination of linear combinations which is just another linear combination. Many kinds of activation functions exist, but we will only discuss the most common types. The sigmoid function seen in section 2.2.3 is one example of an activation function. An other,

closely related, activation function is the hyperbolic tangent illustrated
in figure 2.12:

$$\tanh(z) = 2\sigma(2z) - 1 = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$
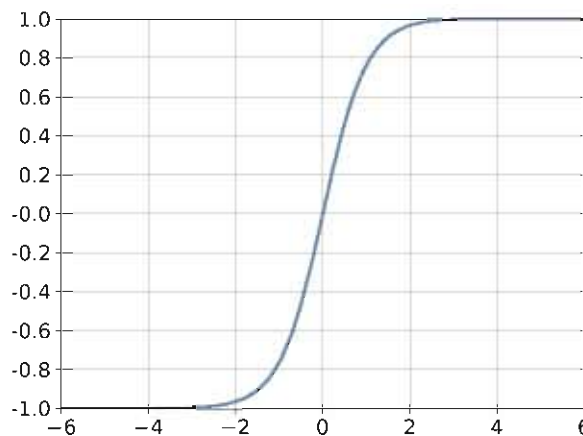
Figure 2.12: The hyperbolic tangent function.

Both the sigmoid and hyperbolic tangent function saturate for high
and low values and are only sensitive to their input close to zero. This
can make gradient based learning difficult. Moreover, the derivatives
of the sigmoid and tanh functions lie within 0 and 1. This means
that every time the gradient is back-propagated to earlier layers (see
section 2.3.2), the signal gets smaller and smaller. This is also known
as the vanishing gradient problem. For these reasons, the use of sigmoid
and tanh activation in feedforward networks is discouraged.

An alternative activation function that does not suffer from the van-
ishing gradient problem is the Rectified Linear Unit (ReLU) shown in
figure 2.13a [22]. The derivative of the ReLU function is either 0 or 1
so the gradient will not vanish. It is not differentiable at zero, but this
is typically solved by returning either 0 or 1. One drawback to ReLU is
that the activation is zero for negative inputs which can result in dead
neurons that are never updated as the gradient is always zero. To resolve
this, Leaky ReLU was proposed which has a small slope for $x < 0$ (see
figure 2.13b) [23].

(a) The ReLU function.

(b) The Leaky ReLU function with a negative slope of 0.1.
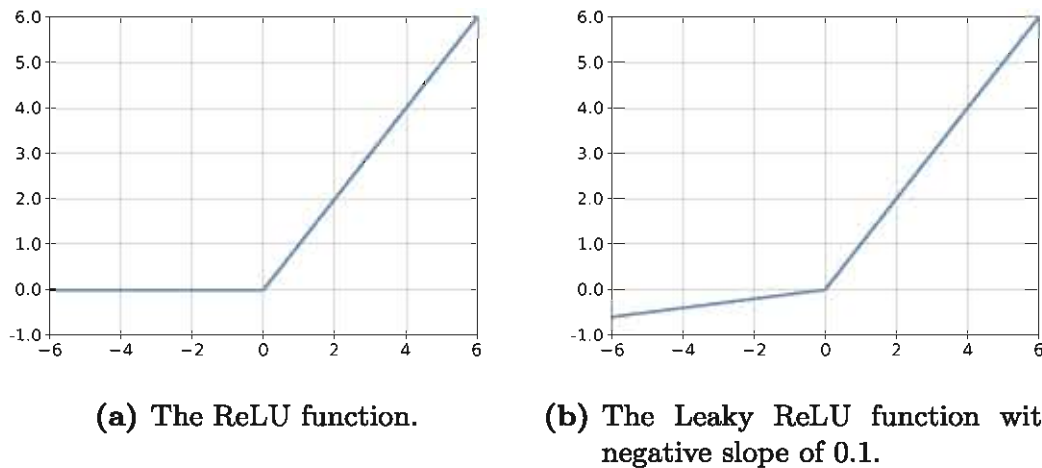
Figure 2.13: The ReLU and Leaky ReLU activation functions.

2.3.2 Training neural networks

Neural networks can be trained using gradient descent similar to linear regression in section 2.2.2. The weights are optimised in three steps: forward propagation, backward propagation and weight update. During forward propagation, input samples are propagated from the input, through the hidden layers to the output layer of the network. A loss is calculated between the output predictions and the ground truth labels. This loss is then backpropagated from the output layer to the input where, at every layer, the gradient of the loss with respect to the weights is computed using the chain rule [24]. The weights are then updated using the negative gradient with a certain step size or learning rate α.

Gradient descent with momentum

As mentioned in section 2.2.2, gradient descent updates are often based on a small randomly sampled subset of the training set, called a mini-batch. Larger batches provide a more accurate estimate of the gradients, but are computationally expensive and the batch size is often limited by the memory of the training hardware. Therefore, small batches are often beneficial and can additionally offer a regularisation effect due to the noise they add during the training process. However, they can also result in noisy gradients as they are calculated based on a few samples. Large oscillations in weight updates can cause slow convergence. To overcome this problem, gradient descent with momentum is introduced. The idea

is to use an exponentially decaying moving average of the past gradients
to update the weights. The weight update then becomes:

$$\Delta w_i = -\alpha \nabla_{w_i} \mathrm{L}(\mathbf{w}) + \eta \Delta w_i$$

with hyperparameter η determining how quickly the contribution of
previous gradients decays.

Adam

The adaptive moment estimation (Adam) algorithm is a modified gra-
dient descent algorithm. It automatically adapts the individual learning
rates of the parameters using running averages of the gradients and
second moments of the gradients [25]. Through the combination of
momentum and adaptive learning rates, Adam is considered faster that
standard gradient descent.

Batch Normalisation

As discussed in section 2.2.2, it is important to scale different input
features to a similar range to improve the learning process. The idea
behind batch normalisation is similar and is used to improve the learning
of deep networks. Very deep neural networks consist of many layers and
with every iteration, the weights of every layer are updated based on the
assumption that the other layers do not change. Changes to the early
layers, however, will affect the deeper layers. To minimise this effect,
batch normalisation is introduced ensuring that the input of each layer is
re-normalised to zero mean and unit variance. Hence, after every hidden
layer, a batch normalisation layer normalises the batch again using the
mean and standard deviation of the current mini-batch [26]. At test
time, running averages of the mean and standard deviation calculated
during training can be used to allow evaluation of the model on a single
sample.

2.3.3   Regularisation

In section 2.2.5, we have explained that the central problem of machine
learning is to build a model that not only performs well on the training

data but also on new unseen inputs. One strategy of regularisation, weight decay, is already described. In this section we will discuss several additional regularisation techniques applicable to deep neural networks.

Data augmentation

The best strategy to reduce overfitting is to train the model on more data. Of course, in practice, the amount of available training data is limited and it is not always possible to collect new additional data. Especially in a medical context for example where data annotation is labour intensive and requires expert knowledge. Data augmentation allows to artificially create new data samples based on the existing training set. Most data augmentation techniques are based on transformations or alterations that the model should be invariant to. For instance, after horizontally flipping an image of a cat, the image still contains a cat. Other methods suited for imaging data are translation, cropping, rotation, adding noise, blurring, changing the image intensity, elastic deformation etc. One should always be careful, however, that the applied transformations do not alter the correct label. For example, in digit recognition, 180° transformations are not appropriate with respect to the difference between '6' and '9'.

Early stopping

When training neural networks, we typically observe a behaviour where the training error steadily keeps decreasing while the validation error starts to increase again after some time. This is similar to the behaviour of the error as a function of model capacity illustrated in figure 2.6. Therefore, instead of training a neural network for a fixed number of iterations, it can be beneficial to monitor the validation error during training and terminate the training process when no further improvement of the validation loss is observed for a predefined number of iterations. The optimal network state is then chosen at the point in time where validation error was lowest. This strategy is known as early stopping.

Dropout

Another regularisation technique, effective in a lot of application domains, is dropout [27, 28]. Here neurons of the network are randomly

dropped during training with a certain probability p as illustrated in figure 2.14. Hence, for every sample in the mini-batch, different units are set to zero and a different subnetwork is created. Therefore, dropout can be thought of as a way to create and train an ensemble of many subnetworks and thereby improve the generalisation performance (similar to random forest as an ensemble of many decision trees in section 2.2.6). Another view on why dropout has a regularisation effect is that it prevents co-adaptation of different neurons. By removing different neurons at every iteration, neurons that are included should perform well regardless of which other neurons that are included in the network. Hence it forces the neurons to be relevant in many contexts.

(a) Original neural network.   (b) Neural network after applying dropout.
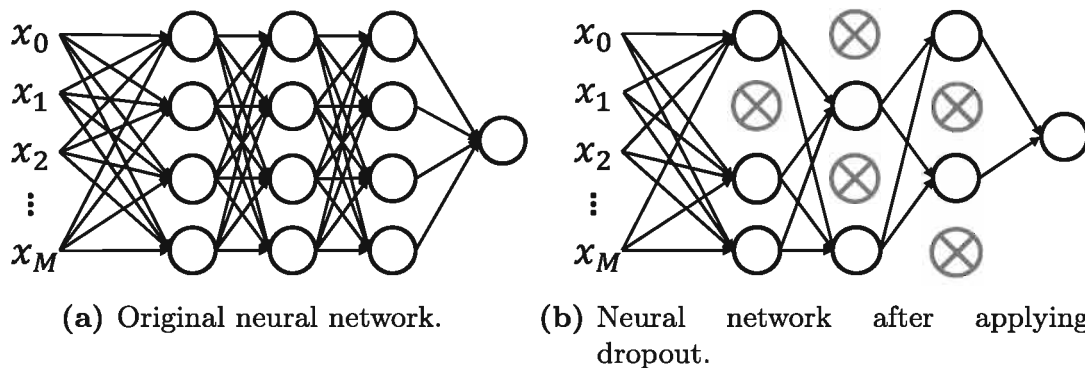
Figure 2.14: Illustration of dropout applied on a feedforward neural network.

Transfer learning

Transfer learning refers to techniques where knowledge learned from one task is transferred to another task instead of training a network from scratch. It is expected that features learned to identity for example cats and dogs in images can be applied to other image recognition tasks as well. This is especially useful in case only a small amount of data is available for the new target task. Through the use of a good starting point, i.e. a network pre-trained on a different related task for which a lot of data is available, high performances can be achieved with only a limited amount of data. One can identify two approaches to transfer learning.

A first method is to use a pre-trained network as feature extractor. One or several of the final layers of the network are removed. Data from the new task is then propagated through the network and the output, i.e. deep features, of the last remaining hidden layer is then used to train a

new less complex machine learning model.

A second approach is to replace the final output layer of the network with a new layer and train or fine-tune the weights of the final layers on data from the new target task. The weights of the early (feature extraction) layers are frozen and only the weights of the final layers are fine-tuned.

Multitask learning

Training a network to simultaneously perform various tasks is called multitask learning. Multitask learning helps the network to learn features that are relevant for multiple tasks which reduces the risk of overfitting [29]. Typically the initial layers of the network are shared between the different tasks and at the end the network is split into different parts that are specific for each task. The different tasks must, of course, be sufficiently related such that neurons trained for one task are applicable for the other tasks as well.

2.3.4   Bayesian deep learning

When analysing data to make predictions, it is often desired to have information on how certain a model is about its output. Uncertainty and probabilistic modelling is of fundamental interest in Bayesian machine learning [30]. Deep learning models, however, often produce point estimates of parameters and predictions with little information on model uncertainty. This can be problematic when, for example, providing out of distribution input data to a model with respect to the data distribution it was trained on. Imagine the example where a model trained on brain scans to detect brain tumours is given a scan of an entirely different structure. The desired behaviour would be for the model to make a prediction but with additional information that it is highly uncertain and the input lies outside the data distribution. Other situations that can lead to uncertain predictions are: noisy data, model parameter uncertainty and uncertainty on the optimal model structure [31]. Uncertainty is important in relation to AI safety. Especially in systems that can directly or indirectly affect human lives like in medical systems or autonomous vehicles.

Different types of uncertainty can be defined: aleatoric and epistemic uncertainty [32].

Aleatoric uncertainty captures noise inherent in the data. This can be measurement noise due to sensor noise or motion. Aleatoric stems from the Latin word Aleator which means 'dice player'. One can further distinguish two types of aleatoric uncertainty: homoscedastic and heteroscedastic. Homoscedastic concerns constant observation noise for every input. Heteroscedastic uncertainty depends on the input, where some inputs can have higher noise outputs than others.

Epistemic uncertainty refers to uncertainty in the model parameters i.e. which model optimally fits our data. This type of uncertainty can be explained away with more data.

Most regression or classification machine learning models output a single prediction value and do not capture uncertainty. The softmax probability scores in classification models are often misinterpreted as confidence levels [31]. Out of distribution samples can result in a high softmax output. So, even with a high softmax score, the model can be uncertain on its prediction.

Bayesian neural networks (BNNs) offer a way to model uncertainty by inferring distributions (e.g. Gaussian) over the model weights instead of point estimates. It was first proposed by MacKay [33] and Neal [34] and further developed with variational techniques by Graves [35], Kingma and Welling [36], and Bach and Blei [37]. Inferring the posterior in a Bayesian neural network is difficult and often approximations are used such as variational inference. Here, the posterior is modelled as a simple variational distribution like a Gaussian. The distribution's parameters are fitted to be close to the true posterior though minimisation of the Kullback-Leiber divergence [31]. These techniques often require many more parameters, cannot scale to complex models and large amounts of data or need specific models which all limit practicality.

In this work we include an intuitive explanation on two practical techniques to model epistemic and heteroscedastic aleatoric uncertainty that scale well to complex models, large data and are applicable to existing models that are widely used. For a more complete and mathematical discussion on Bayesian deep learning we refer the reader to the work by Yarin Gal [31].

MC Dropout

Gal and Ghahramani [38] show that training with dropout (see section 2.3.3) can be cast as approximate Bernoulli variational inference in

Bayesian deep learning. Bernoulli variables require no additional parameters and allow an efficient implementation of Bayesian NN. Through application of dropout at test time and by averaging the outputs of multiple stochastic forward passes, the predictive posterior can be approximated. This is referred to as Monte Carlo (MC) dropout. In practice, an input sample is passed multiple times through the network, each time with different neurons randomly set to zero. The predictive mean over all samples is then used as the final prediction and the variance can be used as an estimate of model uncertainty. As dropout applies a Bernoulli distribution on the model weights, MC dropout is a way to estimate epistemic uncertainty. A more detailed explanation and proof can be found in [31, 38, 39].

Predicted variance

Next to estimating model uncertainty, it can be interesting to model heteroscedastic aleatoric uncertainty as well. Hence we are interested in learning the variance as a function of the input. Heteroscedastic uncertainty can be learned by adding an additional output such that the model predicts both the desired output value $\hat{y}$ as the variance $\hat{\sigma}^2$ associated with the input [31, 32]. The loss function J is then adapted to (for one input sample):

$$ J_{BNN} = \frac{1}{\hat{\sigma}^2} L(\hat{y}, y) + log \ \hat{\sigma}^2 $$

where $L(\hat{y}, y)$ is the original regression or classification loss. One can observe that to minimise the loss, the network needs to learn to associate a higher variance to wrongly predicted samples as this effectively reduces the first term. The second term discourages the model to predict high uncertainties for all samples (and thus ignoring the data). More details on heteroscedastic uncertainty modelling (optionally combined with MC dropout) can be found in [31, 32].

2.3.5   Convolutional neural networks

Convolutional neural networks (CNN) are a type of neural networks specialised to process structured input data [40]. This can be image data thought of as a 2D grid of pixels or even a 3D grid which is often the case

with medical images. An other example is time-series data considered as a 1D grid where samples are taken at regular time steps. Convolutional neural networks have been successfully applied to numerous computer vision applications reaching state-of-the-art performance.

In this section we will describe the fundamental layers used in CNNs: convolutional, pooling and fully connected layers and explain the motivation behind using convolutions in neural networks. Afterwards we will discuss a selection of key CNN architectures throughout the history of computer vision.

Convolutional layer

A convolutional layer consists of several kernels, containing the trainable weights or parameters of the layer, that are convolved with the input. They have the same number of dimensions as the input and an equal depth but are usually much smaller in the other dimensions. The kernel size determines their receptive field. For a 2D convolutional layer the kernel size is defined as: *width* × *height* × *depth*.

Figure 2.15 illustrates a 2D convolutional operation (with a depth of 1). The kernel size is set to a width and height equal to 3 resulting in a receptive field of 3 × 3. The kernel slides over the entire input with a predefined step size or stride, and at every position, a dot product is performed between de kernel and the current input patch. This way, a feature map is created containing the output responses of the kernel at every spatial position. The size of the output feature maps depends on the kernel size $W_K \times H_K$, the stride or step size S and the optional amount of zero padding P around the border of the input according to the following formula:

$$W_{out} = \frac{W_{in} + 2P - W_K}{S} + 1$$

$$H_{out} = \frac{H_{in} + 2P - H_K}{S} + 1$$

Hence for figure 2.15 with $W_{in} = 6$, $H_{in} = 6$, $K = 3$, $S = 1$ and no zero padding, the resulting output size is 4×4. Every convolutional layer consists of several kernels and produces an equal amount of feature maps. These features maps are concatenated resulting in an output with size $W_{out} \times H_{out} \times D_{out}$, where D_{out} is the depth of the output and equal to

the amount of kernels. Consequently, the input depth of a convolutional layer depends on the amount of channels of the network input image (one for a grayscale and three for an RGB image) or the number of feature maps produced by the previous convolutional layer.

We can see that convolutional layers have a lot of hyperparameters that need to be defined: number of kernels, kernel size, stride and padding.
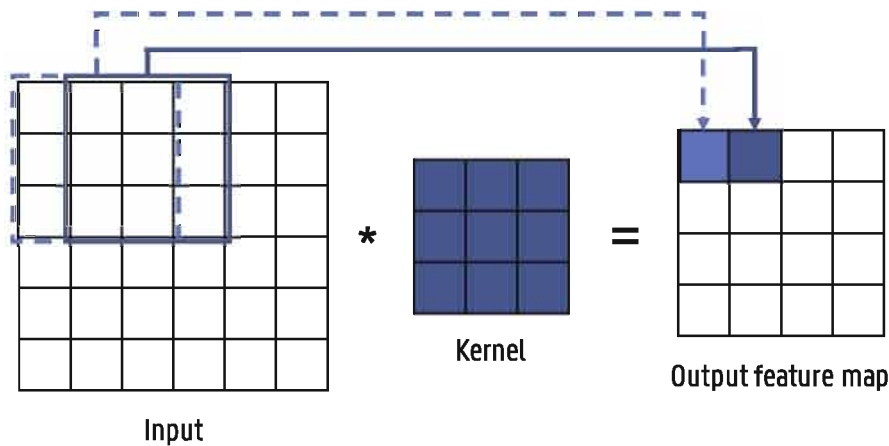
Figure 2.15: Illustration of a convolution operation between a 2D input and a kernel with size = 3 and stride = 1.

The motivation behind using convolutional layers is twofold: sparse connectivity and parameter sharing.

Sparse connectivity means that, in contrast to traditional feedforward networks, the output neurons are not connected to all input units. Input images can contain millions of pixels. Instead of connecting a neuron with every input pixel, relevant features such as edges can be detected using kernels that are much smaller than the input. Although the receptive field of each kernel is small, deeper layers that interact with multiple outputs of earlier layers have an increasingly large receptive field with respect to the input. This allows the network to model complex interactions between simple building blocks across the input.

Parameter sharing denotes that the same kernel is used multiple times across the entire input while in a fully connected network, each weight is only used once. Consequently, a feature only needs to be learned once instead of multiple times for every location. Parameter sharing also causes a convolutional layer to be translational equivariant. This means that, if the input is translated, the output translates in the same way. This is especially useful when features, that detect edges for example,

are relevant across the entire input. Moreover, because of parameter
sharing, the input size does not have to be fixed which allows to process
inputs with varying sizes. Sparse connectivity and parameter sharing
results in a large reduction in number of parameters which improves
statistical efficiency and reduces memory requirements and amount of
computations [12].

Pooling layer

Pooling or subsampling layers reduce the size of the input by calculating
summary statistics over a predefined neighbourhood. As the number of
parameters in the next layers depend on the input size, pooling allows to
improve the computational efficiency and reduce memory requirements.
Different statistics can be computed such as max- and average pooling
as depicted in figure 2.16. The neighbourhood size is usually set to 2×2,
effectively reducing the input size by half. The pooling operation can
also be learned using convolutional layers with a stride larger than one.

(a) 2×2 max pooling   (b) 2×2 average pooling
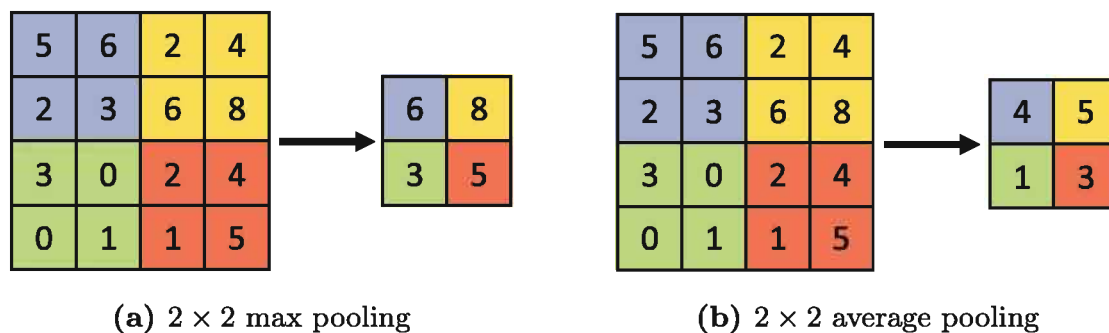
Figure 2.16: Two types of pooling with neighbourhood size 2×2.

Fully connected layer

At the end of a convolutional neural network, often one or several fully
connected or dense layers are applied. This a standard feedforward
neural network layer as seen in section 2.3.1 where every neuron is
connected to every input. These final fully connected layers use the
features extracted by the convolutional layers to determine the final
output class. Hence the convolutional layers are generally seen as the
feature extractors of the CNN and the fully connected layers as the
classifier.

2.3.6   LeNet

Although the idea behind convolutional layers already dates back to 1980 [41], the first modern convolutional neural network was proposed by LeCun et al. [42] in 1998. They trained a CNN to recognise hand-written digits on bank checks. Their dataset, known as the Modified National Institute of Standards and Technology (MNIST) database, is standardised and still used to benchmark different deep learning architectures.
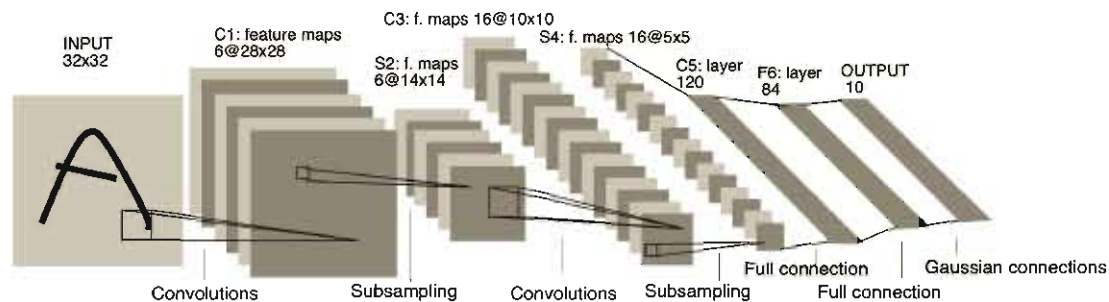
Figure 2.17: Architecture of LeNet-5 proposed by LeCun et al. [42] used for hand-written digit recognition. Image from LeCun et al. [42]. © 1998 IEEE.

The proposed architecture, called LeNet-5 and shown in figure 2.17, consists of two convolutional and subsampling layers followed by two fully connected layers with tanh activation. The final output layer is composed of 10, one for each class, Euclidean radial basis function units which compute the Euclidean distance between the input vector and their parameter vector. A 32×32 pixel greyscale image is provided at the input of the network. The first convolutional layer contains six 5×5 kernels and uses a stride of one resulting in six 28×28 feature maps. Both pooling layers halve the input size and the second convolutional layer has sixteen 5×5 kernels. The network was trained using stochastic gradient descent and MSE loss and data augmentations such as translations, squeezing and shearing were applied to reduce overfitting. A test error was achieved of 0.8%.

2.3.7   AlexNet

The next milestone in deep learning was achieved when a CNN won the ImageNet Large-scale Visual Recognition Challenge in 2012. ImageNet

is a large publicly available dataset containing millions of images labelled with 1000 object classes [43]. AlexNet, proposed by Krizhevsky et al. [44], won this challenge with a top-5 test error rate of 15.3% which was significantly better than the 26.2% error rate achieved by the second-best entry.
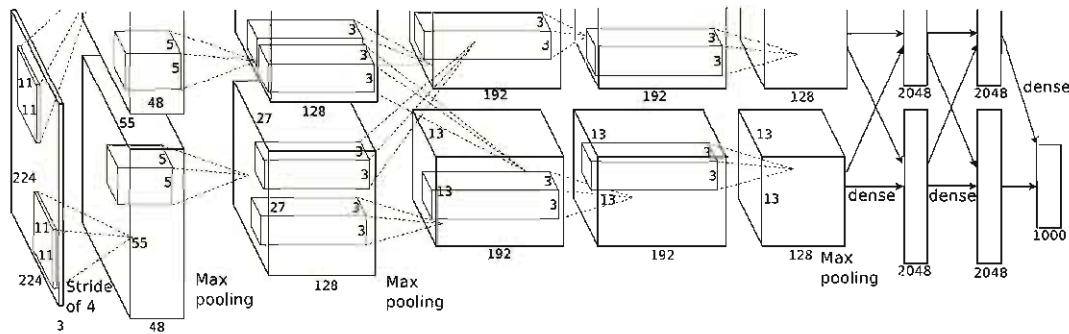
Figure 2.18: Architecture of AlexNet, the winning algorithm of the ImagNet Large-scale Visual Recognition Challenge in 2012. Image from Krizhevsky et al. [44].

The AlexNet architecture shown in figure 2.18 consists of five convolutional layers with max pooling and three fully connected layers where the last dense layer has 1000 outputs with softmax activation. After every convolutional and dense layer, ReLU activation is applied to enable faster convergence. Due to GPU memory limitations, two GPUs were required to train the model which is why the network splits into two streams. The model was trained using gradient descent with mini-batch of 126, momentum and weight decay. Data augmentation and dropout in the fully connected layers was applied to reduce overfitting.

2.3.8   VGG

The effect of network depth on image recognition performance was investigated by Simonyan and Zisserman [45]. They experimented with networks consisting of 11, 13, 16 and 19 layers and reached an error rate of 7.3% on ImageNet in 2014. Every network ends with three fully connected layers similar to AlexNet so the number of convolutional layers varied between 8 and 16. The VGG16 architecture is illustrated in figure 2.19. Convolutional layers with 3×3 kernels are used with ReLU activation. The now common practice of doubling the number of kernels

after every max pooling operation was first presented in their work. The training procedure is similar to the one used by Krizhevsky et al. [44] with a mini-batch size of 256 samples.
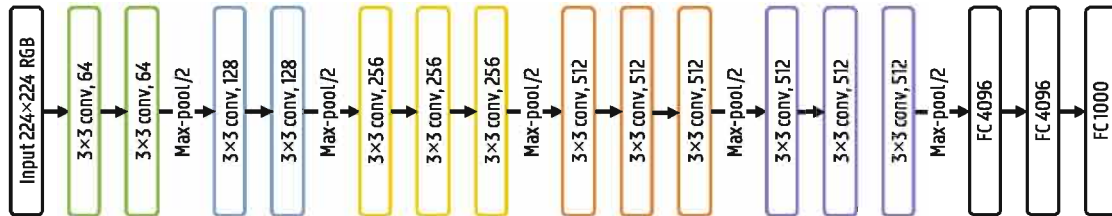
Figure 2.19: The VGG16 architecture with 13 convolutional layers and 3 fully connected layers.

2.3.9 ResNet

The winner of the 2015 ImageNet challenge is the ResNet architecture proposed by He et al. [46]. Previous results with the AlexNet and VGG architectures indicate that increasing network depth strongly improves the image recognition capacity. It was found, however, that when further adding additional convolutional layers the training accuracy saturated and even started to degrade. As this behaviour was observed on the training accuracy, it was not caused by overfitting. This shows that current optimisers find it hard to train increasingly deep networks. A deeper model that performs equally well as its shallower counterpart should exist as it can be constructed by adding layers performing an identity mapping to the shallow network. Based on this idea, He et al. [46] introduced the use of skip connections or residual blocks. The residual block is depicted in figure 2.20. Instead of directly learning the underlying mapping $G(x)$, the layers learn the residual $F(x) = G(x) - x$ due to the skip connection. Their results show that it is easier to optimise the residual function than the original mapping. Hence skip connections allow better optimisation of deeper networks.

A ResNet architecture with 33 convolutional layers and one fully connected layer (ResNet34) is illustrated in figure 2.21. Pooling was performed using convolutions with a stride of two (indicated by /2). Using an ensemble of different ResNet architectures containing up to 156 layers, an error rate was achieved of 3.6%

The use of skip connections was later further exploited by Huang et al. [47] through the introduction of dense blocks containing multiple
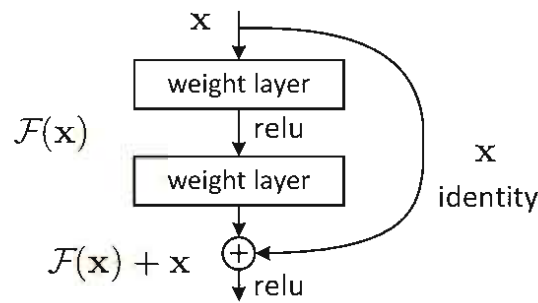
Figure 2.20: Illustration of a residual block. Image from He et al. [46]. © 2016 IEEE.
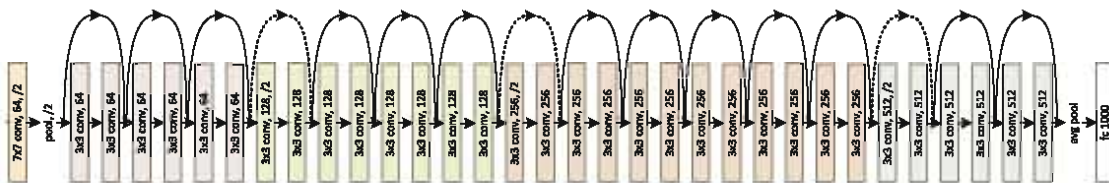
Figure 2.21: The ResNet34 architecture. Image from He et al. [46]. © 2016 IEEE.

convolutional layers with a lot of skip connections.

2.3.10 U-Net

In 2015, U-Net was proposed by Ronneberger et al. [48] as a biomedical image segmentation architecture. They employed the architecture in several segmentation challenges such segmenting neuronal structures in electron microscopy (EM) stacks or cell segmentation in light microscopy images and won with a large margin [48].

The typical use of CNNs was to classify an entire image into a single class label. In many computer vision tasks, however, localisation is required where every pixel is labelled with the class of the object it belongs to. These so-called semantic segmentation tasks were usually tackled using standard classification CNN architectures. Each pixel is separately classified by providing a local region (also called patch) around the pixel to the classification network. Using a sliding-window approach all pixels of an image are classified. This approach has the advantage that additional training data can be generated as a lot of patches can be extracted from one image. This is especially useful in biomedical tasks were the amount of training data is often limited. There are also

two drawback to this strategy. First of all, segmentation of an image is inefficient as many overlapping patches need to be propagated through the network. Secondly, finding the optimal patch size is difficult due to the trade-off between larger patches containing more context and smaller patches for better localisation.

To combine both context and good localisation accuracy, Long et al. [49] introduced the fully convolutional network. The idea is to add upsampling layers after the usual contracting classification network to increase the resolution of the output back to the input image resolution. No fully connected layers are used to preserve spatial information. To increase the output resolution, simple bilinear upsampling can be employed. An other approach is to use transposed convolutions, also called up- or deconvolutions, where the upsampling parameters are learned. The output size of the transposed convolution layer depends on the chosen kernel size and stride. A transposed convolution operation with a stride of two and kernel size 2×2 is illustrated in figure 2.22.
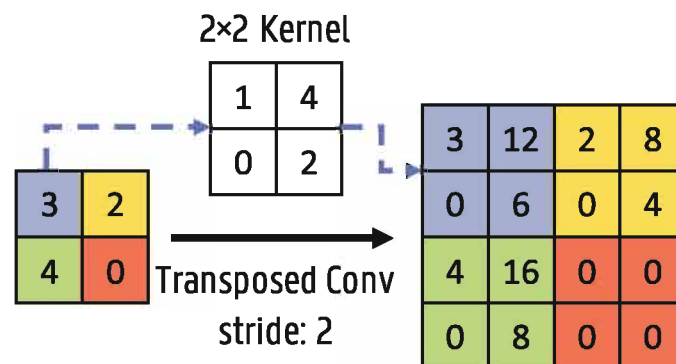
Figure 2.22: Transposed convolution operation with a 2×2 kernel and stride 2.

In the U-Net architecture this upsampling path is further extended with convolutional layers, allowing to propagate context information to the higher resolution layers [48]. This results in a more or less symmetric u-shaped architecture with a contracting and expansive path (see figure 2.23). This type of architecture is also called an encoder-decoder or auto-encoder network. To improve localisation, skip connections are added between the high resolution features of the encoder path and the upsampled feature maps in the decoder path. U-Nets efficiently use semantic and spatial information for accurate segmentation and are still the state-of-the-art for many segmentation tasks.
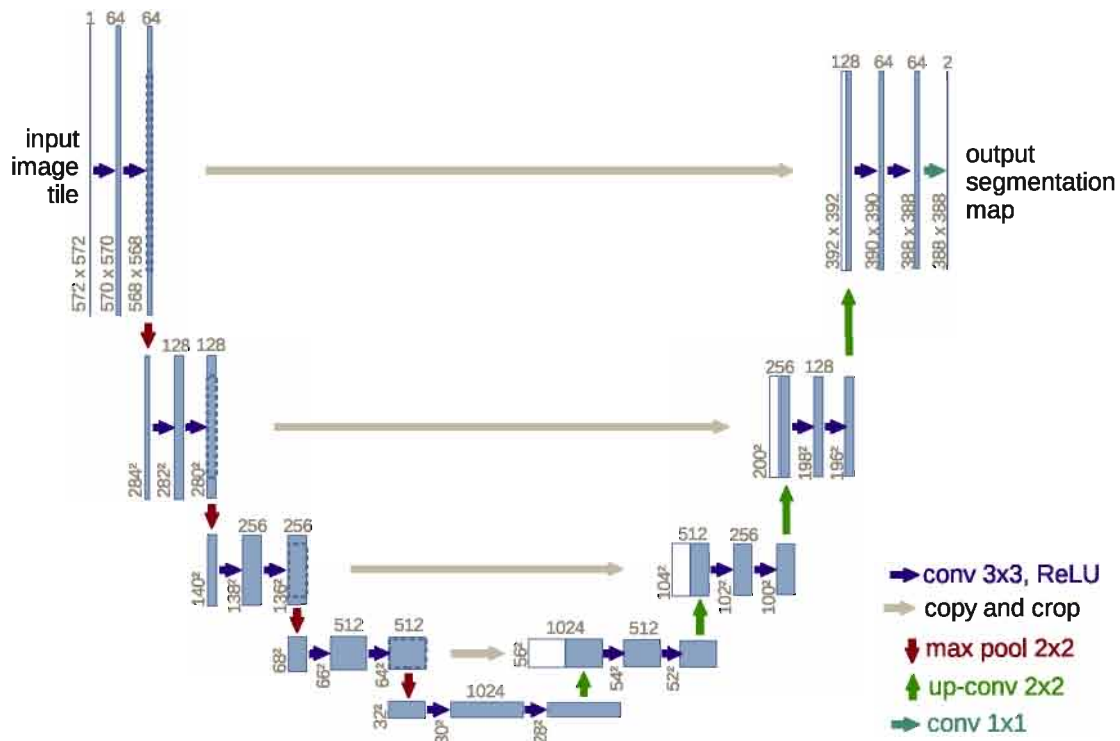
Figure 2.23: The U-Net architecture. Reprinted by permission from Copyright Clearance Center: Springer Nature, Ronneberger et al. [48], © 2015.

2.4 Conclusion

In this chapter, we gave an overview and intro into artificial intelligence, how it evolved throughout history and why AI development started to boom in the last two decades. The relation between AI and its subfields machine learning and deep learning as a means to build AI systems was clearly described. The main principles behind machine learning were introduced that also form the basis of neural networks and deep learning. In the last part, deep learning networks and training procedures were described with a focus on convolutional neural networks as this is the most common type of network used in medical imaging.

3 | Artificial intelligence in medical imaging

This chapter covers the role and state-of-the-art of AI in medical imaging. In the introduction, the need, potential and challenges of AI in healthcare are discussed. Afterwards, a brief outline of several most common medical imaging modalities is presented and positron emission tomography, relevant to part I of this work, is covered in more detail. Finally an overview is provided of state-of-the-art AI applications throughout the entire medical imaging chain.

3.1 Introduction

In previous chapter we have seen that the rapid progress of AI over the last decades has been possible due to the ever increasing amount of computational power and available data. This growing amount of data is witnessed across all industries, including healthcare. All kinds of patient data are recorded and stored into electronic health records such as lab results, reports, DNA analysis, activity and health data from wearables etc. A major volume of healthcare data comes from medical imaging. Due to advances in medical image acquisition, novel imaging procedures are introduced and the amount of diagnostic imaging is growing fast [50]. From 2D X-rays in the early days, medical imaging evolved to multimodal, dynamic and 3D CT, MRI and PET exams. This rising amount and complexity of imaging data increases the workload of radiologists. The Royal College of Radiologists, for example, has warned of shortages in the radiology workforce growing every year [1]. Radiologists struggle to meet the rising demand for imaging examinations resulting in delayed diagnoses and potentially affecting the accuracy of clinical decisions.

At the same time, the increasing amount of healthcare data contains a wealth of information that presents opportunities for personalised and precision medicine. As the huge amount of data is overwhelming for physicians, we need sophisticated AI algorithms to exploit all this information. We have seen in previous chapter that enough training data is a key requirement to develop these AI algorithms. Hence the rising amount of healthcare data is putting great pressure on the healthcare industry but is simultaneously providing the opportunity to revolutionise healthcare.

In case of medical imaging, artificial intelligence can be employed to improve the entire imaging pipeline. This is also reflected in the amount of publications about AI in radiology on PubMed as shown in figure 3.1. AI can be applied during image acquisition (see section 3.3.1) and reconstruction (see section 3.3.2) to advance image quality, acquisition speed and reduce cost. Moreover, it can be used for image denoising, registration and translation between different modalities (see section 3.3.3). Finally, a lot of AI applications are developed for medical image analysis including abnormality detection, segmentation and computer-aided diagnosis (see section 3.4).

There remain, however, several challenges to the adoption of AI in medical imaging. Although the amount of imaging data is rising fast, the number of curated datasets is still limited. Data is scattered across clinical centres with highly varying imaging protocols, recorded modalities, patient groups, included patient information, annotations etc. Data curation and annotation of medical images is time consuming, requires expert knowledge and is subject to inter- and intra-observer variability. It is difficult to gather enough data for rare pathologies and the distributions between different classes are often highly unbalanced. For these reasons, the availability of medical imaging data to train AI algorithms is still limited, certainly when compared with natural image datasets like ImageNet containing millions of images. An initiative that tries to solve this issue is The Cancer Imaging Archive (TCIA) which hosts a large archive of publicly available medical image datasets [51]. Medical image analysis is also more complex. The imaging data is often 3D which adds an additional dimension of complexity. They can have large variations in resolution, contain noise and artefacts and lack contrast which influences the performance of AI algorithms. Many applications also require information from multiple images combining different contrasts, functional and anatomical information or temporal behaviour.
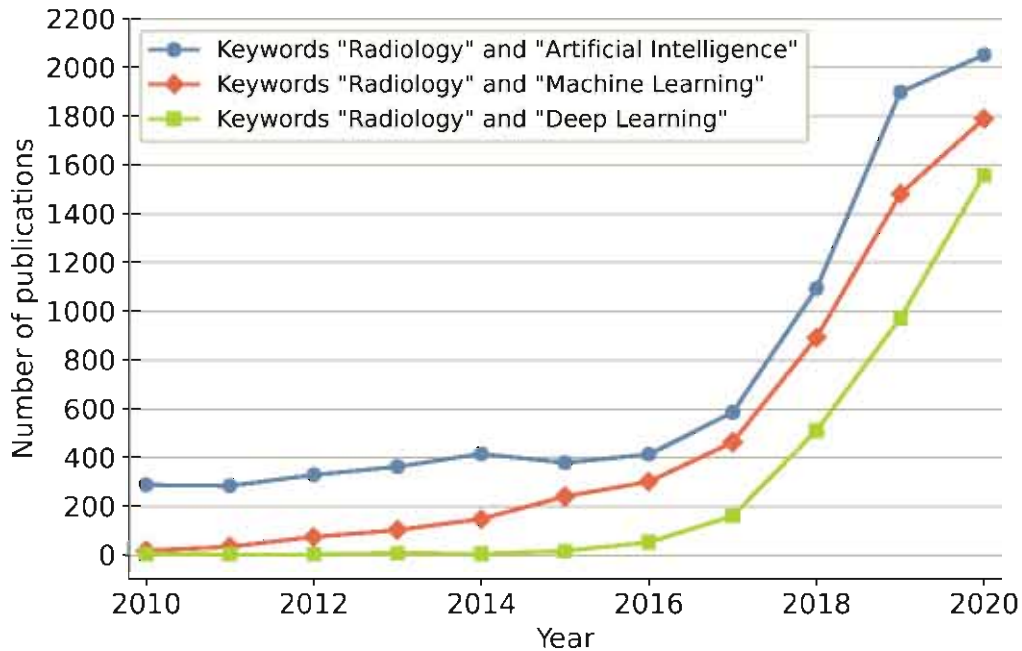
Figure 3.1: The growth of AI in radiology reflected in the number of publications on PubMed when searching on the terms "radiology" with "artificial intelligence", "machine learning" or "deep learning".

All these elements pose specific challenges to the design of medical image analysis tools. Moreover, detection, segmentation and interpretation of anatomical structures, both normal and pathological is inherently very complex. They have varying shapes, intensities and show large inter- and intra-subject variability. AI systems need to be robust to perform well under this wide variety of conditions.

Finally, as these AI tools can have a direct influence on diagnosis and treatment planning, more research is necessary towards explainable AI in order to understand and trust these algorithms. Deep learning algorithms are often seen as a black box and it is difficult to understand how and why the algorithm makes certain predictions and under what circumstances it might fail.

3.2   Medical imaging

Medical imaging encompasses techniques to image the structure and function of the human body for research, diagnostic and treatment pur-

poses. It allows to, often non-invasively, look at the interior of the body and plays an increasing role in healthcare management. In this section we will give a brief overview of the most common imaging modalities. The principle behind positron emission tomography will be covered in more detail as this is necessary to understand part I of this work.

3.2.1 Brief overview

In medical imaging one can distinguish structural and functional modalities. Where structural imaging refers to the visualisation of the anatomy, functional imaging measures the physiological activity of the human body such as metabolism and blood flow. Figure 3.2 shows an overview of different structural and functional imaging modalities that will be discussed below. For a more complete overview of the different medical imaging techniques, we refer the reader to the book *Fundamentals of Medical Imaging* by Paul Suetens [52].

X-ray

X-ray is short-wave electromagnetic radiation and was first discovered by Wilhelm Konrad Röntgen while experimenting with cathode tubes. He noticed that fluorescent screens started glowing when struck by light emitted from the tube, even when the tube was inside a box. Hence, the tube was not only emitting light but also a new kind of radiation. Moreover, Röntgen found that the radiation was attenuated differently by various materials and that the projection of an object could be captured on a photographic plate. Since different tissues inside the human body have varying X-ray absorption coefficients depending on their density and thickness, the medical potential soon became clear.

Today, X-ray or radiography is still widely used in clinical practice with its main applications in imaging of the skeleton (fractures), chest, lung, dental and breast (mammography). Drawbacks to radiography are the limited contrast between soft tissues with similar densities and exposure of the subject to ionising radiation.

Computed tomography

Computed tomography (CT) uses X-rays to produce a 3D image. Hence the same principle is used as in radiography but many radiographs are
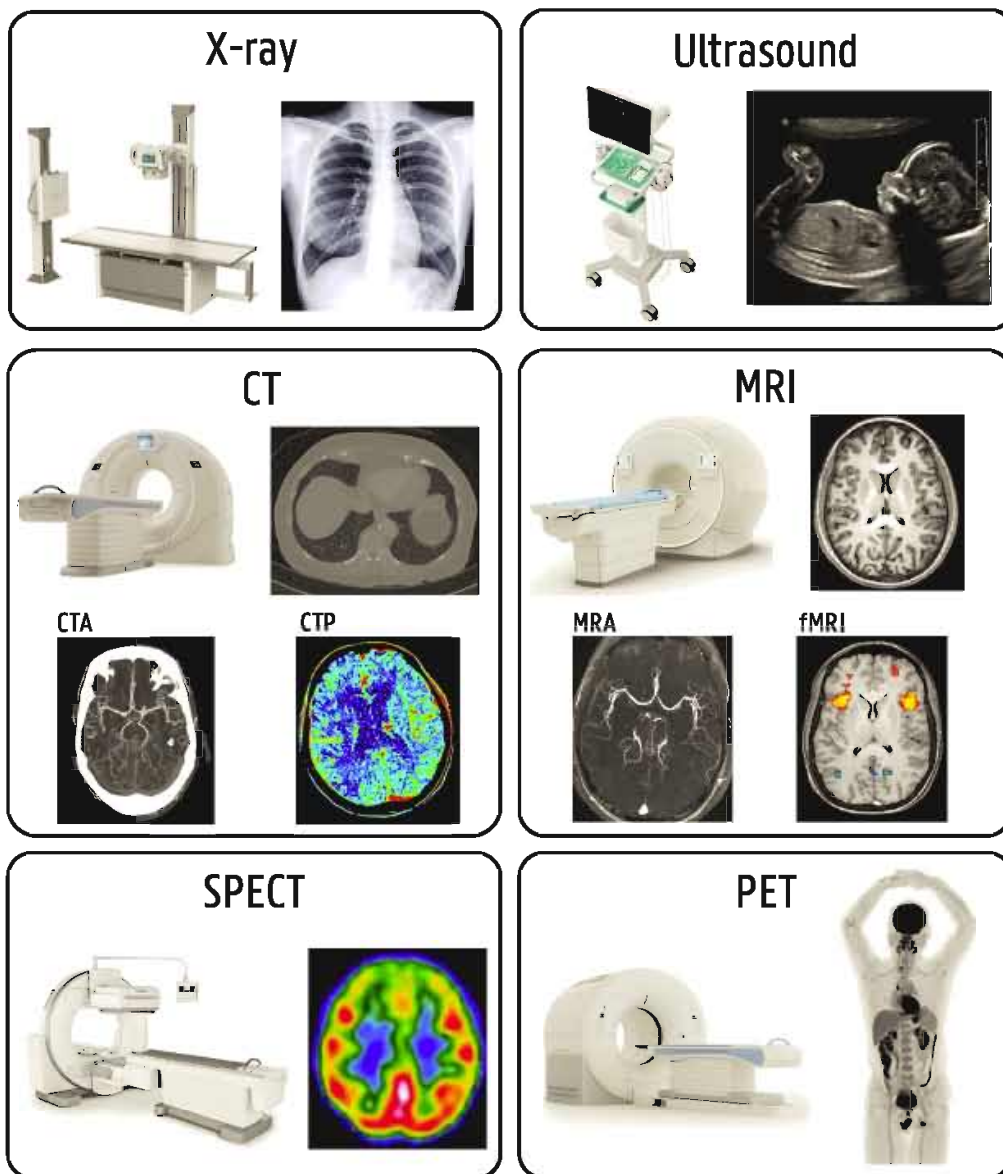
Figure 3.2: Overview of the most common medical imaging modalities.

acquired from different rotation angles to obtain one CT scan. One cross-section or slice is formed by rotating the X-ray source and detector at one transversal location. By moving the scanner table, the whole body can be scanned. The mathematical foundation of reconstructing a 3D image from many projections at different angles was already proven by Radon in 1917. The experimental development of the technique for CT was done by Annan Cormack in 1963 and the first CT scanner was built by G. Hounsfield in 1971. Since then CT has dramatically improved with advances such as helical and dual-energy CT which reduce the scanning time and patient dose and optimise the image quality. Current state-of-

the-art CT can scan the full body in less than one minute.

Pixels in a CT scan are a measure of X-ray beam attenuation or density and are scaled in Hounsfield Units (HU). They denote the ratio of the linear attenuation coefficient μ of a voxel to the linear attenuation coefficient of water μ_{water} and is given by:

$$CT(x,y) = 1000 \frac{\mu(x,y) - \mu_{water}}{\mu_{water}}$$

Lungs typically have very low intensities (HU below zero), soft tissues range between 20 and 80 HU and bones have a much higher value than water (300-2000 HU).

Compared to radiography, CT can produce 3D images with better tissue contrast. An important drawback remains the radiation dose delivered to the patient. The most common investigations include musculoskeleton system, abdomen, head and neck and thorax. To study arteries and veins, contrast agents with high attenuation coefficients can be used. This is called CT Angiography (CTA). CT can also be employed for functional imaging of blood flow for example which is called CT perfusion (CTP). Dual-energy CT (DECT) uses two separate x-ray photon energy spectra. This allows better visualisation of different tissues and their chemical compositions, e.g. iodine and calcium content, that have different attenuation properties at different energies.

Magnetic resonance imaging

Magnetic resonance imaging (MRI) is a relatively recent medical imaging technique with the first nuclear magnetic resonance (NMR) image taken in 1973 [53]. In contrast to CT, MRI does not use ionising radiation but is based on interactions between hydrogen atoms in the body and electromagnetic waves. Charged particles have a quantum mechanical property called spin which can be regarded as a rotating motion about its axis, creating a magnetic moment. Nuclei, consisting of an even or odd number of subatomic particles, have a zero or non-zero magnetic moment. Hydrogen nuclei have a non-zero magnetic moment as they consist of just one proton. As hydrogen is abundant in the human body, MRI focuses on visualising hydrogen-containing tissues like muscles, brain, kidney etc.

The main component of an MRI scanner is a large superconducting magnet producing a strong static uniform magnetic field. Typical magnetic field strengths are 1.5T (Tesla) or 3T and recently even 7T systems are being installed in academic centra. By placing the subject in this external field, a small fraction of the protons inside the body align with this field. Radio-frequency pulses are then applied tuned to the resonance or Larmor frequency with which the magnetic moment precesses around the magnetic field. These pulses disturb the equilibrium and result in a flip of the net magnetisation vector. The magnetisation vector now has a transversal component which can be measured with radio-frequency receiver coils of the MRI scanner. When the radio-frequency pulses are switched off, the magnetisation returns to its equilibrium, also called relaxation. The relaxation time depends on the local environment of the atoms and is therefore tissue dependent. As a result, different tissues will have different intensities in the MRI. By varying the acquisition protocol determining the time points of excitation and acquisition, tissues will show different values and the optimal contrast between tissues of interest can be chosen. Spatial information (slice selection and position encoding within the slice) can be encoded by superimposing small gradients to the magnetic field.

Similar to CT, MR Angiography (MRA) can be used to visualise blood vessels but without the need of contrast agents. However, contrast can still be used for visualisation of blood. Various MR protocols exist for functional imaging such as diffusion weighted imaging (DWI), perfusion weighted imaging (PWI) or functional MRI (fMRI) visualising brain activity based on changes in oxygenation level in the blood. MRI offers better soft-tissue contrast compared to CT and can be used to image all parts of the human body that contain hydrogen without using harmful ionising radiation. A downside is that MRI is slower than CT and MRI intensities do not have a direct physical meaning and are not standardised. This results in large differences in intensities depending on the MRI scanner and scanning protocol making it challenging to compare data across different centres and apply artificial intelligence models.

Ultrasound

Ultrasound imaging was initially developed for military war purposes to detect submarines (SONAR) during World War I. Only later, during World War II, ultrasound started to be used for clinical applications

and the first 2D grey scale image was produced during the 1950s. The basic principle behind ultrasound is quite simple. A sound wave (with a frequency higher than the upper limit of human hearing) excited by an ultrasonic probe, also called a transducer and based on piezoelectric elements, propagates through the imaged medium and reflects at the interface between different tissues. These reflections are measured as a function of time and using known velocities of the waves inside the medium, positional information is obtained. However, other phenomena like diffraction, attenuation, scattering etc. appear besides reflection which complicate ultrasound imaging and need to be taken into account.

Ultrasound imaging is portable, has relatively low cost, harmless and high temporal resolution making it a very popular medical imaging technique. It is best known from gynaecology but is also widely used for detection of liver tumours, liver cirrhosis, prostate and spleen cancer, echocardiography, cranial ultrasound and carotid imaging. Motion (e.g. blood flow) can be visualised as well by means of Doppler imaging.

Nuclear medicine

The goal of nuclear medicine imaging is to visualise the distribution of a molecule inside the body and to derive information on the function or metabolism of certain organs or to detect tumours with high activity uptake. To this end, nuclear medicine uses radioactive isotopes and the tracer principle. A radioactive isotope is labelled to a molecule that is involved in a certain metabolic process and is injected in the body. With the advent of the Anger Scintillation camera in 1957, γ-rays emitted by this radioactive isotope could be detected allowing to measure the concentration of the molecule inside the body as a function of position and time. Based on the work by David e. Kuhl and Roy Edwards published in 1963 [54], two different tomographic techniques were developed: single photon emission computed tomography (SPECT) and positron emission tomography. As the name suggest, SPECT uses single photon emitting isotopes whereas PET uses positron emitting isotopes followed by annihilation of the positron with an electron into two gamma rays travelling in opposite directions. The principle and components of PET scanners will be explained in more detail in section 3.2.2.

The strength of PET and SPECT is that the sensitivity is orders of magnitudes higher than other functional imaging techniques of CT and MRI. Weaknesses are the lower spatial (around 5 mm versus 1 mm) and

temporal (minutes versus seconds) resolution and exposure of the subject to radiation. PET and SPECT systems are nowadays combined with CT or even MR into hybrid SPECT/CT, PET/CT, PET/MR systems allowing to simultaneously obtain both structural and functional information. The most important clinical applications of nuclear medicine are in oncology, thyroid function, bone metabolism, functional cardiac imaging and lung embolism.

3.2.2 Positron Emission Tomography

In previous section we already briefly explained the principle behind nuclear medicine and positron emission tomography. Here we will discuss the different components of a PET scanner in more detail. Figure 3.3 illustrates the principle and different components of a PET system.
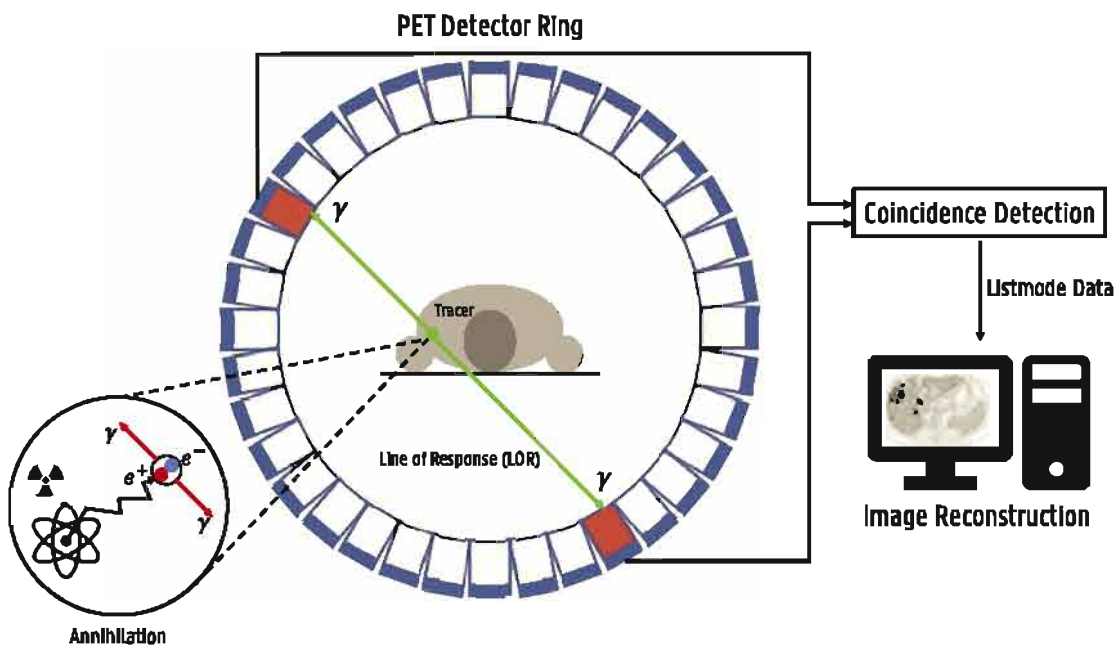
Figure 3.3: A graphical illustration of the principle and workflow of positron emission tomography.

The goal of PET is to image the distribution of a certain molecule of interest (e.g. glucose) inside the body [55]. This molecule is labelled with a radioactive isotope to form a tracer which is injected in the subject. This tracer is administered in very low amounts such that the studied biological process is not affected and the absorbed dose by the patient remains as low as possible. Common PET radioisotopes are ^{11}C, ^{18}F,

^{15}O, ^{13}N and ^{68}Ga. The most important tracer is 2-deoxy-2-^{18}F-fluoro-D-glucose (^{18}F-FDG) which is a glucose analog where one hydroxyl group is replaced with ^{18}F. It is used to visualise glucose metabolism with its main application in oncology as malignant tumours show an increased glucose metabolism.

The radioisotopes have a surplus of positive charge and decay by emitting a positron (e^+) and a neutrino (ν). For ^{18}F the following reaction occurs:

$$^{18}F \rightarrow {}^{18}O + e^+ + \nu$$
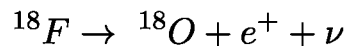
The released positron travels a short distance (within a few mm, depending on the positron energy), called positron range, through the surrounding tissue until it annihilates with a nearby electron generating two photons or gamma rays travelling in opposite direction. Each photon has an energy of 511 keV (rest energy of an electron). These photons are detected by PET detectors that are placed in a ring around the subject. PET scanners operate under the assumption that the path of the photons and the annihilation points are on the same line (collinearity) and that photons originating from the same decay arrive around the same time at the detector ring. Through coincidence detection, i.e. when two photons are detected within a certain time window (around 10 ns), the line of response (LOR) where the annihilation occurred can be recorded. By recording many LORs, the tracer distribution can be calculated using reconstruction algorithms like filtered back projection or iterative reconstruction methods such as maximum likelihood expectation maximisation [56, 57].

PET Detector

An optimal PET detector should have following properties: a high stopping power or sensitivity of 511 keV photons and a good energy, temporal and spatial resolution. A high sensitivity is required in order to visualise low radiation doses. A high temporal and energy resolution allows optimal coincidence detection and spatial resolution is important to accurately determine the LORs. The fundamental components are a scintillator and photodetectors. Figure 3.4 shows a typical pixelated clinical PET detector design consisting of discrete long and narrow scintillation crystals coupled to an array of photodetectors.
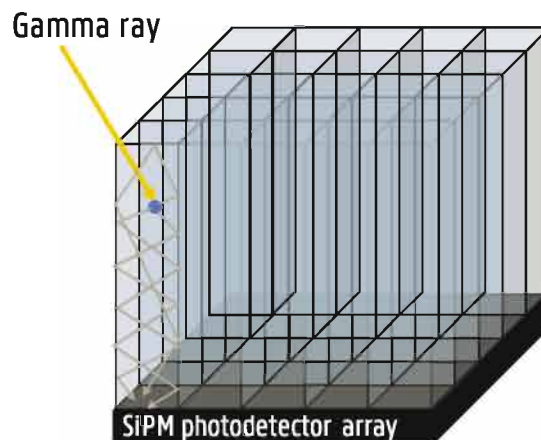
Figure 3.4: Illustration of PET Detector design with pixelated crystals.

Scintillation crystals are responsible for stopping the incoming gamma rays and to re-emit the absorbed energy in the form of light. During the 80s and 90s mostly pixelated BGO crystals were used. BGO crystals have a high stopping power but a limited light output. Today, L(Y)SO crystals have become the standard PET scintillators as they have a good stopping power, high light output and are quite fast [58]. To further raise the stopping power, the crystal thickness can be increased. This results, however, in increased cost and can degrade spatial resolution (see section on degrading factors below).

The photodetectors convert the emitted light to electrical signals. The most common devices are photomultiplier tubes (PMTs). They are, however, increasingly being replaced by Silicon photomultipliers (SiPMs) due to their fast response time, compactness and high gain with low voltage.

Based on the measured electronic signal, the pixel can be determined where the gamma interaction occurred. This pixel position is then used to estimate the LOR.

Image degrading factors

Besides the PET detectors properties, other factors influence the image quality of PET as well.

Positron range and acollinearity impose fundamental limits on the achievable spatial resolution of PET systems.

The positron travels a short distance through the subject to lose kinetic

energy before annihilating with an electron. Hence the location of annihilation is not the same as the location of positron emission. This positron range depends on the radioactive isotope and the surrounding tissue. Isotopes that emit low energy positrons will have a short positron range. Typical root mean square effective ranges are around 0.5 mm to 3 mm [55].

The photons resulting from annihilation will not be emitted exactly in opposite directions due to momentum of the positron and electron. This effect is called acollinearity or non-collinearity. The angular distribution follows a Gaussian distribution with a FWHM of 0.5°. The effect of acollinearity on spatial resolution, expressed in FWHM, is dependent on the detector ring diameter D according to $R_{180°} = 0.0022D$. For a PET scanner with a diameter of 80 cm this result in a FWHM of approximately 2 mm [55].

In case no depth-of-interaction (DOI) measurement in the PET detector crystal is present, the parallax effect further degrades the spatial resolution. This effect is illustrated in figure 3.5. Depending on the depth of interaction in the crystal, the actual line or response differs from the measured LOR.
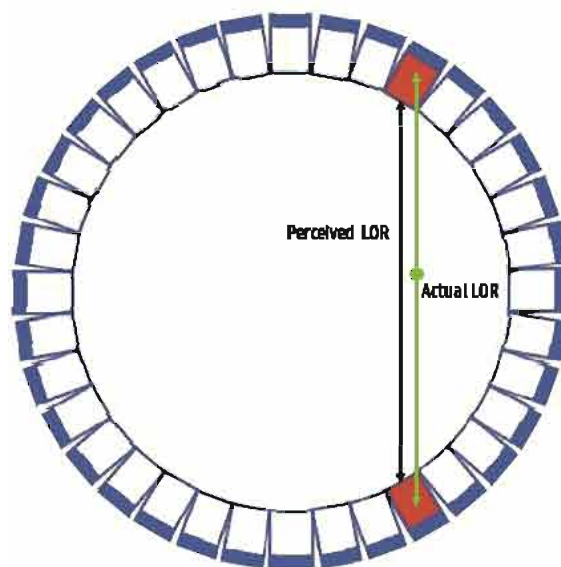
Figure 3.5: Illustration of the parallax effect. Depending on the depth of interaction in the crystal, the actual line or response (LOR) differs from the measured LOR.

The emitted 511 keV photons have to travel through the body before being detected. Hence there is a chance that one or both of the photons undergo Compton scatter (deflection) or absorption in the body. This

leads to a decreased number of coincidence events and thus measured LORs. For 511 keV photons, the probability of absorption is higher than the probability of Compton scattering. Because the probability of attenuation for photons originating from the centre of the body is much higher, it is important to implement attenuation correction during reconstruction using an attenuation map of the body. This attenuation map can be measured through a transmission scan or can be derived from a low dose CT in PET-CT scanners [55].

Measured coincidences are not alway true coincidences i.e. containing the annihilation point along its line of response. Coincidence events also include scattered and random coincidences depending on the detector energy and temporal resolution as illustrated in figure 3.6.
Scattered coincidences occur when at least one of the photons is Compton scattered resulting in a loss of energy and deviation from the correct LOR direction. It is possible to distinguish scattered from true events through energy filtering. In practice, the detector energy resolution is limited and coincidences within a certain energy window around 511 keV are accepted to maximise the detection of true coincidences while minimising the amount of scattered coincidences.
Two photons from different annihilation events arriving at the detectors within the coincidence time window is called a random coincidence. Similarly to the energy window, the coincidence time window should be large enough to avoid loss of true coincidences but not too large to limit the number of random coincidences.

Innovations in PET

Continuing progress is made to improve the coverage, sensitivity and spatial resolution of PET systems.
To increase the coverage and sensitivity, there is a trend towards PET scanners with larger axial Field of View (FOV) and even to total-body PET systems [55]. Typical whole body clinical PET scanners axially cover 20 to 25 cm of the body at a time. Through the addition of detector rings to extend the scanner along the length of the body, many more events can be detected thereby increasing the signal to noise ratio (SNR). The higher sensitivity can also be used to reduce the scan time or radiation dose while maintaining the same SNR. The main difficulty of these systems is the evident increase in cost.
Other innovations are mainly focused on improving the sensitivity and
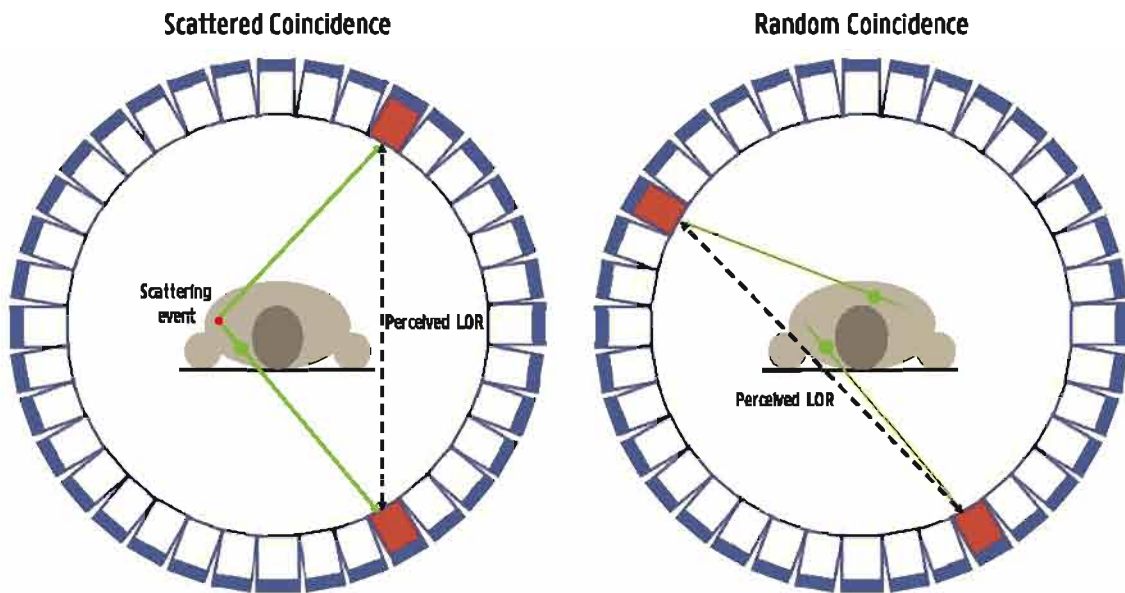
Figure 3.6: Illustration of scattered and random coincidence events.

spatial, temporal and energy resolution of the PET detectors through the use of better crystals, photodetectors, faster electronics, improved detector designs and computational methods to better estimate the location, time and energy of interactions. Improvements in timing resolution are also important to enable time-of-flight (TOF) PET [55]. By measuring the time difference between coincident photons, which depends on the distance between annihilation and the detectors, the position of annihilation along the LOR can be determined. Theoretically it is possible to directly infer the annihilation site using TOF information without the need for a reconstruction algorithm. However, the difference in arrival time changes only around 67 ps per cm difference in location and current PET detectors do not have the necessary timing resolution for direct reconstruction.

Monolithic PET detectors

Current clinical PET scanners employ pixelated scintillation detectors (see figure 3.4) where the scintillation light is restricted to the individual crystals thereby limiting the spatial resolution of the detector to the pixel size (typically 3-5 mm) [58]. Improving the spatial resolution by reducing pixel size negatively impacts other desirable parameters like sensitivity (more dead space between crystals), timing and energy resolution and increases cost.

As an alternative, the use of monolithic detectors is investigated to increase spatial resolution without the aforementioned trade-offs (see figure 3.7)[59–62]. Monolithic detectors can provide a better sensitivity (no dead space), temporal, energy and spatial resolution compared to their pixelated counterpart. Additionally, DOI information is intrinsically present in the measured scintillation light distribution when using monolithic detectors [63]. Interactions far from the photodetectors result in a broader distribution compared to interactions occurring close to the photodetectors. Estimation of the 3D interaction position, including DOI, is important to draw the most accurate line of response and reduce parallax errors during image reconstruction. Moreover, using DOI information, it is possible to improve timing resolution by correcting for the photon travel time between the interaction position and the photodetectors [64, 65].
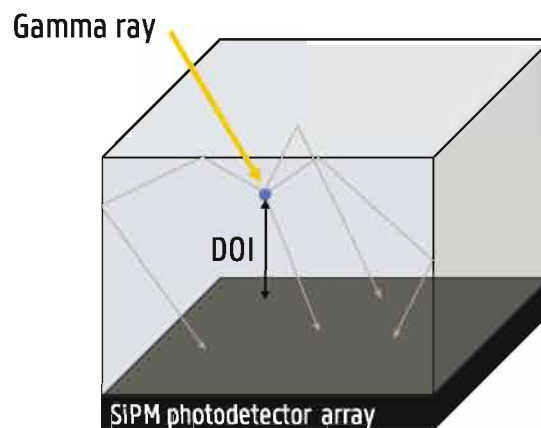
Figure 3.7: Illustration of a monolithic PET Detector design.

Already implemented in preclinical systems [66, 67], monolithic detectors are likely to also appear in clinical scanners [58]. The main challenges are (i) the lengthy calibration procedure and directly linked to that (ii) the gamma event positioning algorithm.
Acquisition of calibration data is required to develop the positioning algorithm. In a typical calibration setup, a collimated 511 keV pencil beam is used to irradiate the crystal in discrete steps over the entire crystal surface. This way, light distributions can be acquired at known beam positions. This process is time-consuming but is not seen as a limiting factor as techniques to expedite calibration are actively investigated in other research groups [58, 68, 69].
The second challenge is to develop an accurate and efficient algorithm

capable of correlating the measured light distribution with all possible interaction positions within the crystal. The positioning performance of monolithic detectors typically degrades near the edges [63]. Due to truncation of the light distribution, the positioning algorithms have less information to accurately estimate the interaction position. This edge effect can be limited using powerful positioning algorithms and high granularity readout [61]. A wide range of positioning algorithms have been proposed and will be discussed in the next section.

3.3 AI in medical image formation

Artificial intelligence techniques can be applied throughout the entire image formation process [70–74]. This includes AI to improve the raw measurement data, for better reconstruction of the image and post-processing to enhance the image quality or for translation between different modalities.

3.3.1 Acquisition

Improving the quality of the raw measurement data during image acquisition has a beneficial effect throughout the entire subsequent imaging pipeline, from reconstruction to analysis. The work by Sloun et al. [71] covers the use of deep learning on all aspects of ultrasound imaging with applications to improve raw signal acquisition using neural networks for adaptive beamforming. In this section we focus on the use of AI for positioning and timing of photons in monolithic PET detectors [72] as this is relevant for part I of this work. We have seen in previous section that spatial and timing resolution are critical parameters of PET detectors that largely influence the performance of the PET system.

Positioning

Most of the AI applications in PET detectors are on the estimation of the interaction position of the 511 keV photons in the detector. The scintillation light distributions captured by the photodetectors are used to determine the position of absorption.

In pixelated detectors, the determination of the interaction pixel is quite straightforward using centroid weighted methods such as Anger

Logic [72]. Hence there has been few investigations on the use of more complex machine learning methods to estimate the crystal of interaction in pixelated detectors. However, AI can play a role in obtaining DOI information which is normally not available in these detectors. Pizzichemi et al. [65] proposed a linear method to infer DOI in a pixelated detector with single side readout using light sharing through a light guide at the front of the detector. A DOI resolution of 4.1 mm FWHM was achieved with $1.53 \times 1.53 \times 15$ mm^3 crystals arranged in an 8×8 array. Zatcepin et al. [75] later improved upon this method through the use of neural networks for DOI estimation resulting in a better uniformity and DOI resolution of 3 mm FWHM.

In previous section we have seen that monolithic detectors can provide a better sensitivity, temporal, energy and spatial resolution and intrinsic DOI encoding. They require, however, powerful positioning algorithms able to correlate the measured light distributions with all possible interaction positions within the crystal. A wide range of positioning algorithms have been proposed such as maximum likelihood estimation [76–78], k-nearest neighbour [60, 79, 80], gradient tree boosting (GTB) [69, 81] and neural networks [82–84].

In Pierce et al. [78] a $50 \times 50 \times 10$ mm^3 LYSO crystal fixed to a 65-channel position-sensitive photomultiplier tube was evaluated using Gaussian maximum likelihood for interaction position estimation. Calibration data was acquired with a 0.9 mm beam source traversing the crystal in 1.52 mm steps. A new multiplexing scheme was proposed to reduce the 65-channel signal to 7 channels. They reported a FWHM of 1.2 mm in the detector centre and 1.9 mm near the edge of the crystal. Using GTB, a spatial resolution of 1.4 mm FWHM has been reported for a LYSO crystal of $32 \times 32 \times 12$ mm^3 [69]. Calibration data with a pitch of 0.75 mm was acquired with a parallel hole or fan beam collimator. Besides the photon counts for all pixels, additional features were used as input such as centre of gravity, main pixel, row and column sums, etc. Separate GTB models were trained for x- and y-position. A similar approach was proposed to add DOI estimation (FWHM of 2.12 mm) by acquiring calibration data through side irradiation. The GTB models can be adapted based on required performance versus memory restrictions [81].
A spatial resolution of 1.7 mm FWHM and an average DOI resolution of 3.7 mm FWHM could be achieved with a $32 \times 32 \times 22$ mm^3 LYSO crystal and the k-nearest neighbour algorithm [60]. Through the use of

an algorithm that preselects only the most useful reference events, the k-nearest neighbour classification algorithm could be accelerated.

In Stockhoff et al. [80], optical simulations were used to investigate the lower limit of intrinsic spatial resolution for a $50 \times 50 \times 16$ mm^3 LYSO crystal using a mean nearest neighbour approach. Reference light distributions were calculated for a grid of 49×49 positions and interpolated to a step size of 0.25 mm. A 2D spatial resolution of 0.56 mm FWHM was reported with a DOI error of 1.6 mm. Nearest neighbour algorithms achieve state-of-the-art spatial resolution but are computationally expensive as for every new event a distance metric needs to be calculated with a potentially large set of reference light distributions of known incident positions.

The use of neural networks for 3D positioning has been proposed in Wang et al. [83]. Events are processed in two steps: first a global network roughly estimates the x- and y-coordinates to select a sub-area of the detector. In a second step, separate x, y and DOI networks for every sub-area further refine the position. For a LYSO crystal of $25.5 \times 25.5 \times 10$ mm^3 a plane and DOI resolution of around 2 mm FWHM is achieved. Iborra et al. [84] propose an ensemble of neural networks, trained on simulation data, to estimate the 3D photoelectric interaction position. Separate ensembles are trained for each coordinate and evaluation on measured test data shows an average resolution of 2-2.4 mm FWHM. They report that training and testing the ensemble to predict the first (Compton or photoelectric) interaction gave poor results. However, positioning of the first interaction is required to draw the most accurate line of response.

Timing

PET detectors with good timing resolution are important for random scatter rejection and to allow time-of-flight estimation. In most PET detectors, the time of interaction is estimated using linear methods that measure the time when the photodetector signal crosses a certain threshold [85]. This is likely not an optimal use of the information contained in the detector waveforms. Machine learning algorithms that use the digitised rising edge of the photodetector signals could more accurately determine the time of interaction. Ground truth TOF data can easily be acquired by moving a point source over a small distance range between pairs of detectors as the TOF difference is exactly determined by the

known distance and speed of light.

Berg and Cherry [86] proposed a convolutional neural network to predict the TOF difference from the pair of digitised detector waveforms of a coincident event. The detectors consisted of an LYSO crystal coupled to a photomultiplier tube. An improvement in timing resolution of 20% was achieved compared to leading edge discrimination (185 ps versus 231 ps).

3.3.2 Reconstruction

Most medical imaging modalities do not directly generate data in image space but require an image reconstruction step to form the image from the acquired raw signals. A first type of reconstruction algorithms are analytical methods such as filtered back-projection for tomography data or inverse fast Fourier transform for spatial frequency data in MRI. These methods are popular due to their computational simplicity but often suffer from poor resolution-noise tradeoffs [87]. Instead, iterative reconstruction methods like expectation-maximisation can be used where the image is recursively updated to better match the measured data according to a forward model describing the system's physics and sensor and noise statistics [87]. They result in an improved image quality by reducing noise and artefacts but are computationally expensive and may contain errors in the forward model.

A recent development in the field of image reconstruction is the introduction of deep learning approaches [87–89]. Neural networks can be trained to directly reconstruct the image from the raw projection or k-space data using known input and output pairs as illustrated in figure 3.8. This approach is entirely data-driven as the inverse mapping and noise characteristics are learned from the data without underlying assumptions on the imaging process. Examples of direct reconstruction methods with deep learning are automated transform by manifold approximation (AUTOMAP) [90], DeepPET [91] and direct PET image reconstruction network (DPIR-Net) [92].

The AUTOMAP architecture consists of three fully connected layers followed by a convolutional encoder-decoder. They show flexibility in learning the inverse mapping for various MRI acquisition strategies (undersampled, misaligned Fourier, Radon projection and spiral non-
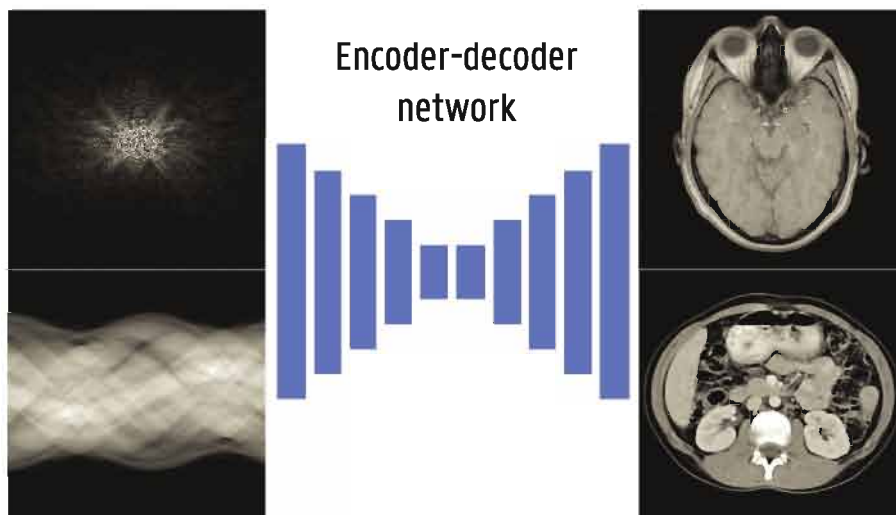
Figure 3.8: Direct MRI or CT reconstruction from raw k-space or sinogram
data respectively using an encoder-decoder deep neural network.

cartesian Fourier). Results show diminished noise and artefacts with
AUTOMAP compared to conventional reconstruction methods. More-
over, they show that the inverse mapping could also be learned from
natural images in ImageNet and emphasise that their approach can be
generalised to other reconstruction problems across a broad range of
different modalities.

DeepPET similarly uses an encoder-decoder convolutional neural net-
work to directly reconstruct high quality PET images from the sinogram
data. The network was trained on simulated data derived from a whole-
body digital phantom. Results demonstrate higher quality images com-
pared to conventional techniques in only a fraction of the time.

DeepPET was later used by Hu et al. [92] as a generator in a Wasserstein
generative adversarial network (GAN) for direct PET image reconstruc-
tion, called DPIR-Net, reaching further improvement in image quality.
GANs have recently gained a lot of attention in the computer vision and
medical imaging community, especially in image synthesis, enhancement
and translation tasks which will be explained in next section [93].

Remaining drawbacks to direct reconstruction are the large amounts of
data required to learn the complex mapping, limited interpretability and
current approaches operate on 2D slices instead of full 3D reconstruction
due to memory constraints.

For this reason there is an increasing interest in model-based networks
that incorporate existing domain knowledge prior to training. Existing
reconstruction techniques are translated into a neural network where

the different mathematical operations are mapped to different networks layers.

Wurfl et al. [94] showed that the filtered back-projection algorithm for CT can be translated into a neural network. The weights are initialised with known values from the analytical approach. This way, prior to training, the performed operation is identical to the FBP algorithm. By further training the network on known input-output pairs, noise characteristics can be incorporated thereby improving the reconstruction accuracy.

Iterative reconstruction methods can similarly be translated into neural networks through algorithm unrolling [95]. The number of iterations is fixed and each update or iteration is mapped to one or several network layers. These layers are then stacked to form an end-to-end mapping between the raw data and the final reconstructed image which can then further be optimised using regular network training. Applications of unrolled algorithms in MRI [96–101], CT [102, 103] and PET [89, 104] show that they can improve image quality and reconstruction speed compared to traditional iterative methods [105].

3.3.3 Enhancement and translation

In previous section, we have seen that deep learning reconstruction techniques can learn to correct noise and artefacts resulting in improved image quality. Other than using deep learning during reconstruction, it can also be applied as a post-processing tool in image domain to enhance image quality. Possible applications include restoring high dose from low dose scans [106–113], fully sampled from undersampled scans [114, 115], super-resolution [116–119], denoising [120–124] etc. Next to image enhancement, deep learning is also used for registration [125–130] and translation [131–137] of different modalities. Registration refers to the process of aligning different modalities or scans at different time points such that anatomical structures spatially coincide. Cross-modality synthesis or translation can be useful to generate other modalities without additional acquisition time and cost. For example, pseudo-CT generation from MRI allows MRI-guided radiation therapy requiring CT equivalent images for positioning and dose calculation. Additionally, in PET/MR systems, a CT image needs to be synthesised to calculate the attenuation map necessary for attenuation and scatter correction.

State-of-the-art deep learning solutions for the above image-to-image enhancement and translation tasks mostly use U-Net type networks com-
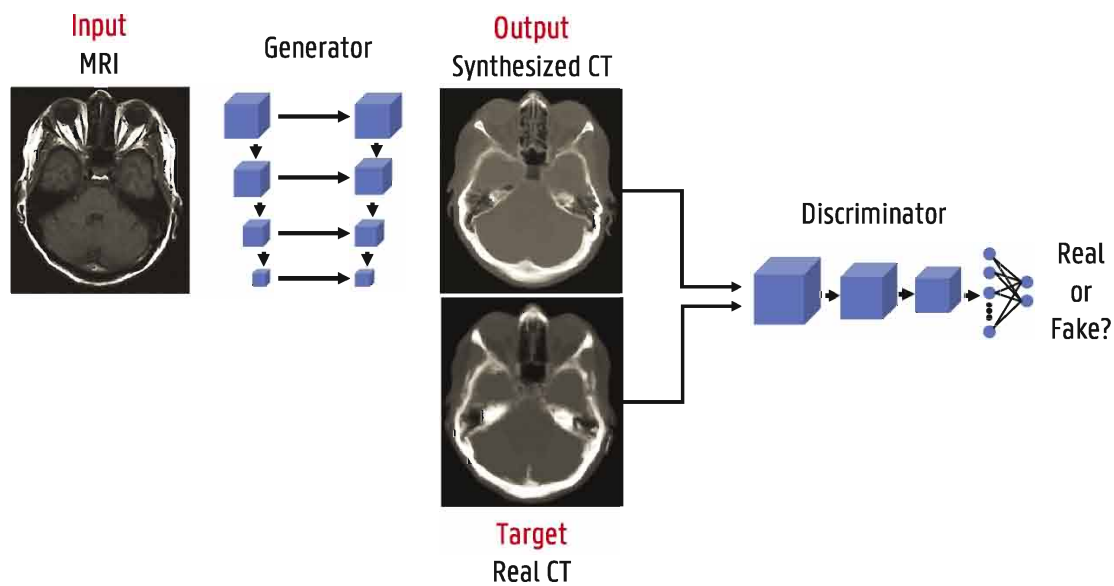
Figure 3.9: Generative adversarial network (GAN) framework illustrated with a pseudo-CT from MRI generation examples.

bined with a discriminator network to form generative adversarial networks [93]. Generative adversarial training is a framework where two networks, a generator and a discriminator, are simultaneously trained and compete against each other [138, 139]. This is illustrated with a pseudo-CT from MRI generation example in figure 3.9. The generator focuses on image synthesis and tries to fool the discriminator which is trained to identify real versus synthesised images. While training, the gradients are back propagated from the discriminator to the generator so the parameters of the generator are adapted to produce realistic images according to the discriminator. Next to this adversarial loss other loss functions such as L1 Loss are incorporated as well to retain image details. GANs and variants thereof, e.g. cycleGAN [140], are widely used in image reconstruction and enhancement.

3.4   AI in medical image analysis

A lot of AI algorithms applied in medical imaging are to improve the efficiency and accuracy of medical image analysis and even to extract information that is not (yet) perceived by human experts. Different applications can be identified being segmentation, treatment monitoring, prognosis, computer-aided detection (CADe), computer-aided diagnosis

(CADx) etc.

Given that a vast number of medical image analysis applications of AI have been reported, it is infeasible to cover all literature in this work. We therefore selected several important works across different commonly found anatomical application areas. This illustrates the potential and current progress of AI in medical image analysis. For more exhaustive literature surveys, we refer the reader to Litjens et al. [2], Zhou et al. [74], Mazurowski et al. [141], Ranschaert et al. [142], Rueckert and Schnabel [143], and Ibrahim et al. [144].

3.4.1 Approaches

There are two main approaches to medical image analysis, being the more traditional radiomics pipeline and, more recently, the end-to-end deep learning algorithms. Radiomics is mostly used in limited data settings which was mostly the case in the early days of medical image analysis with AI. In recent years, the availability of larger medical imaging datasets has increasingly resulted in a transition towards deep learning approaches.

Radiomics

Radiomics refers to the extraction and analysis of large amounts of quantitative imaging features [145]. The aim is to convert medical images into quantitative mineable data and to make current radiological practice, which is often more qualitative, quantitative and standardised. In other words, many quantitative features are extracted from the 2D or 3D medical images which can then be analysed by machine learning algorithms to find correlations with certain disease characteristics such as prognosis and disease type. When the relation between image features and genomic patterns are investigated one often refers to radiogenomics. The typical radiomics workflow consist of a segmentation, feature extraction and analysis step as illustrated in figure 3.10.

To extract radiomics features, the structures of interest need to be segmented. This is often done manually by an experienced radiologist or with (semi-)automatic segmentation algorithms. From these delineated structures, many features can be extracted describing its shape, volume, texture, intensities etc. The last step is then to analyse the extracted
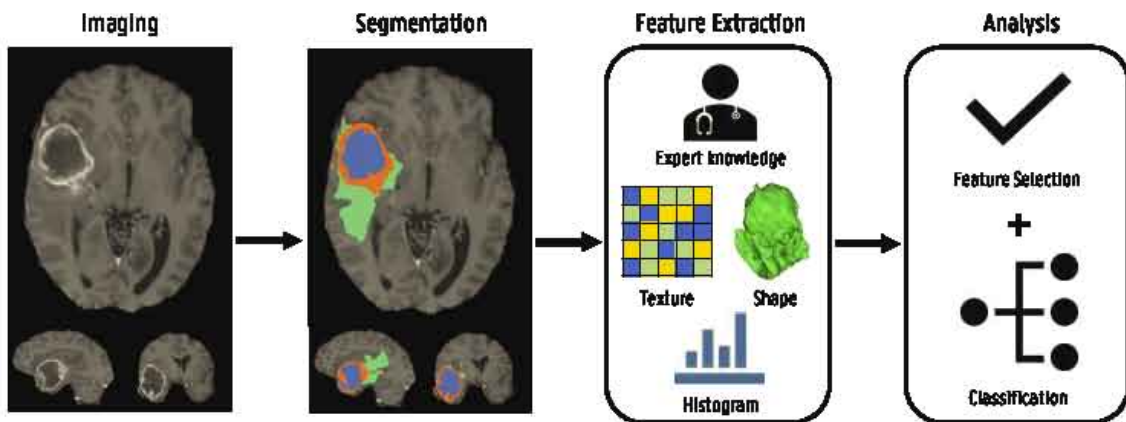
Figure 3.10: Illustration of the radiomics workflow.

features. This often starts by removing redundant and irrelevant features to select a minimal subset of highly predictive features with respect to the considered task. One can use specific features selection algorithms or find the features that result in the best performance of the subsequent machine learning model. For final prediction, usually more traditional machine learning algorithms are used like random forests and support vector machines.

There are several challenges to the radiomics approach regarding imaging, segmentation, feature extraction and efficiency. First of all, there is a large variety in scanners and imaging protocols between different institutions resulting in strongly differing image characteristics such as resolution, contrast, noise, slice thickness, intensity values etc. These differences have a strong impact on the extracted radiomics features reducing robustness and generalisability of the trained models across different centres. Therefore, standardised imaging protocols are preferred and data from different sources should be normalised both in space and intensity.

Secondly, since features such as shape are based on the segmentation masks, accurate and reproducible delineation is of crucial importance. Manual segmentation suffers from interreader variability and is labour intensive making it unfeasible for large databases. (Semi-)automatic segmentation algorithms are therefore increasingly developed. Training and evaluation of these algorithms is often done using manual delineations making the assessment of their true accuracy difficult. For this reason consistency and reproducibility might be more important properties for radiomics analysis. To this end, manual-interference should be min-

imised.

Thirdly, a vast amount of features can be defined and extracted. Consequently many of the extracted features can be redundant or irrelevant for the task at hand. As seen in chapter 2, too many features can result in overfitting and proper feature selection is therefore very important. At the same time, the features are hand-engineered and defining the optimal features for a certain task is not straightforward. This way, import information in the medical images might be missed.

Finally, the entire pipeline of (manual) segmentation, features extraction and analysis can be time-intensive which is often not desired in clinical applications.

Deep Learning

To address the above challenges associated with radiomics, there is a transition towards the use of end-to-end deep learning approaches. They directly receive the medical images as input and provide at the output the desired outcome prediction. Often the workflow is still split into a segmentation and classification part to allow the prediction algorithm to focus on the relevant regions of interest. However, no manual feature extraction is necessary as the deep learning networks automatically learn the most optimal features. Both in the segmentation and classification stages, deep networks can pave the way for state-of-the-art, unbiased, fast and automatic medical image analysis.

The challenge with deep learning on the other hand is the requirement of even more data to train the complex (3D) networks. Large datasets are not always available and strongly application dependent. Moreover, deep learning often lacks interpretability. In radiomics, the features used by the model to make a certain prediction can be identified and interpreted whereas deep learning is seen as a black box. Hence, although there is an increasing use of deep learning approaches to achieve state-of-the-art performances, radiomics is still often employed when limited data is available and insight in the decision process is necessary.

In chapter 8 we perform a comparison between radiomics and deep features from a pre-trained CNN for brain tumour grading. The remaining part of this dissertation focusses on deep learning.

3.4.2   Segmentation

As discussed in previous section, segmentation of structures of interest is an important task in medical image analysis. It is not only an important pre-processing step to improve further classification and diagnosis, it is also relevant for therapy planning and assessing therapy response. Automatic segmentation has many advantages compared to labour intensive manual segmentation suffering from interreader bias and low reproducibility and is therefore widely investigated [146–148].

Where the early segmentation systems used region-growing, clustering and traditional machine learning approaches based on hand-crafted features, deep learning approaches now dominate the state-of-the-art in medical image segmentation. The most well-known CNN architecture for medical image segmentation is U-Net originally proposed by Ronneberger et al. [48] for segmenting neuronal structures in electron microscopy stacks and cell segmentation in light microscopy images. U-Nets and its modifications are the state-of-the-art architectures in many segmentation tasks.

Milletari et al. [149] proposed a 3D variant of the U-Net architecture, called V-Net, with residual blocks in the encoding and decoding paths for prostate segmentation in MRI. They used a novel cost function to train de model based on the Dice score (see section 2.2.4). This allows a more balanced evaluation of segmentation performance in case the structure of interest is much smaller compared to the entire image. Since then, Dice loss is one of the most used cost functions for segmentation tasks. They trained and evaluated their model on the PROMISE12[1] dataset of the MICCAI Prostate MR Image Segmentation challenge organised in 2012 and reached an average Dice score of 87%.

A self-configuring deep learning method for medical image segmentation, called nnU-Net was proposed by Isensee et al. [150]. It automatically adapts pre-processing steps, network architecture (2D, 3D or cascaded U-Net), training and post-processing depending on the task and dataset properties. nnU-Net achieves state-of-the art results in many biomedical segmentation challenges and won first place in the Medical Segmentation Decathlon[2] organised in 2018 [151]. The aim was to evaluate the generalisability of a segmentation algorithm across

---

[1] https://promise12.grand-challenge.org/Home/
[2] http://medicaldecathlon.com

many different tasks instead of designing specialised solutions for one specific task. The challenge includes segmentation of 10 structures: liver, colon, pancreas and lung tumours in CT, brain tumours and prostate in multi-modal MRI, hippocampus and cardiac in mono-modal MRI and hepatic vessels and spleen in CT. Several segmentation examples from the Medical Segmentation Decathlon are included in figure 3.11.

In chapters 7 and 9, brain tumour segmentation will be covered in more detail.
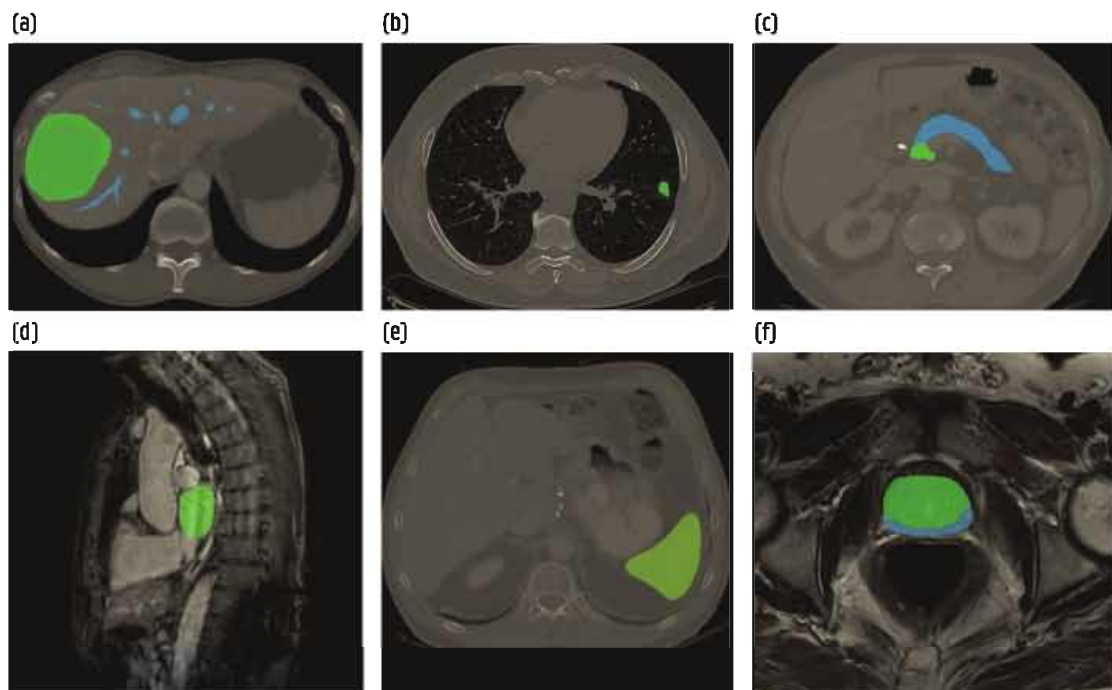
Figure 3.11: Segmentation examples from the Medical Segmentation Decathlon [151]. **(a)** Hepatic vessel (blue) and tumour (green) in CT. **(b)** Lung tumour (green) in CT. **(c)** Pancreas (blue) and tumour (green) in CT. **(d)** Left ventricle (green) in MRI. **(e)** Spleen (green) in CT. **(f)** Prostate peripheral (blue) and transitional (green) zones in MRI.

3.4.3 Detection and diagnosis

Computer-aided detection consists of localising organs or abnormalities such as lesions. It can be seen as a pre-processing step followed by further diagnosis of the found region of interest (ROI). Note that some of the discussed studies may overlap with the subject of segmentation covered in previous section.

Chest pathology

One of the most widely studied topics is lung nodule detection in low-dose CT scans which is an important step in identifying early stage lung cancer [152]. Early detection reduces lung cancer mortality and screening programs are increasingly implemented. As interpretation of lung CT scans to find small lung nodules is tedious, error-prone and time consuming this puts a lot of pressure on radiologists. Different algorithms for automatic lung nodule detection were compared in the LUng Nodule Analysis 2016 (LUNA16) challenge [153]. This challenge made use of the publicly available LIDC-IDRI dataset containing 888 chest CT scans with lung nodule annotations performed by four radiologists [154, 155]. Most of the proposed methods consist of two stages: a candidate detection stage and a false positive reduction stage. The candidate detection stage typically makes use of a 2D (slice-level) or 3D U-Net architecture and often has a high sensitivity at the cost of many false positives. Therefore, the false positive reduction stage additionally classifies the found ROIs as a true nodule or not using standard classification CNN architectures. Through the combination of different solutions, a sensitivity of over 95% was achieved at less than 1 false positive per scan.

To analyse screening CT scans for lung cancer, the found nodules with nodule detection algorithms need to be classified according to malignancy [152]. Many different types of algorithms have been proposed for benign-malignant pulmonary nodule classification including more traditional radiomics approaches as well as 2D or 3D convolutional neural networks. Diagnosis of lung cancer based on low-dose CT was the topic of the 2017 kaggle Data Science Bowl [3]. The top ten submissions all used deep learning algorithms often with a similar approach as for lung nodule detection. Figure 3.12 shows an illustration of a typical lung cancer screening pipeline with 3D CNNs. The winning algorithm consisted of two modules: a 3D region proposal (nodule detection) network and a second module evaluating the cancer probabilities for the five detected nodules with highest detection confidence [156]. Both modules made use of a modified U-Net architecture. Few years later, google researchers published an end-to-end lung cancer screening algorithm using [157]. They employ a 3D inflated inception architecture [158] which builds upon the inception network for 2D image classification pre-trained on ImageNet but inflates the filters into 3D. Their model achieves state-of-

---

[3]https://www.kaggle.com/c/data-science-bowl-2017

the-art performance (AUC of 94%) on the National Lung Cancer Screening (NLST) dataset [159] containing 6716 cases and on an independent clinical validation set of 1139 cases. This performance was on par or even outperforming six radiologists.

Other applications of AI in chest pathology include diagnosis of pulmonary embolism, tuberculosis, airway diseases, interstitial lung disease and others [160].
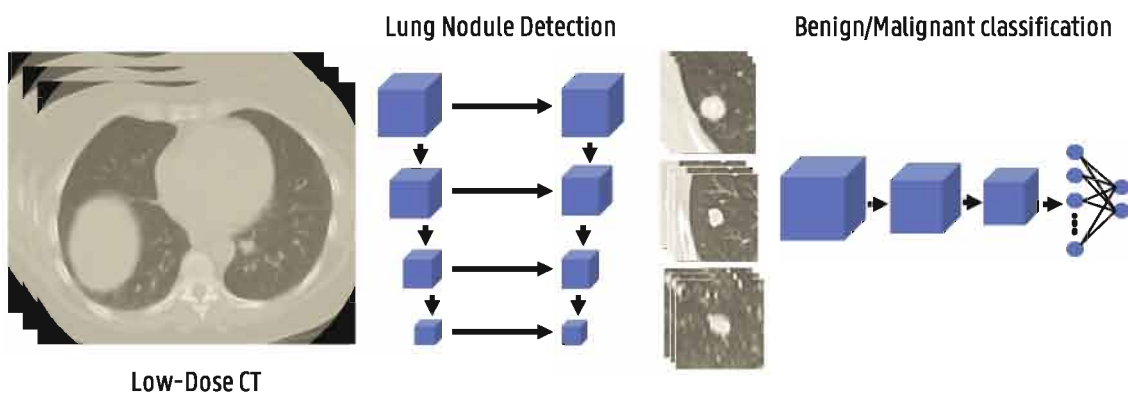
Figure 3.12: Illustration of a typical lung cancer screening pipeline consisting of a lung nodule detection and a malignancy classification stage.

Recently, medical imaging such as X-ray and CT have played an important role in diagnosis and management of COVID-19. Many artificial intelligence tools have been developed and contributed to improve the safety, efficiency and accuracy of the imaging workflow to fight COVID-19 [161–168]. Murphy et al. [161] have proposed an AI system to detect COVID-19 pneumonia in chest X-rays. After pre-processing consisting of image normalisation and lung segmentation using U-Net, a CNN was used for patch- and image-level classification. The network was pre-trained to detect tuberculosis and subsequently fine-tuned to detect pneumonia in general and COVID-19 pneumonia. Evaluation on a test dataset of 454 chest radiographs from an independent Dutch hospital shows an AUC score of 0.81 which was comparable to the performance of six chest radiologists. Prokop et al. [169] aimed to introduce a standardised reporting system for CT of COVID-19. They assess the suspicion of COVID-19 infection of a scale from 1 (very low) to 5 (very high). An AI tool to automatically asses CO-RADS score and extend of infection was proposed by Lessmann et al. [165]. The system consisted of three successively applied deep learning algorithms performing lobe segmentation, lesion segmentation and CO-RADS scoring respectively.

Pulmonary lobe segmentation was performed using a two-stage U-Net [170]. For segmentation of ground-glass opacities and consolidation in the lungs, a 3D U-Net was used built with the nnU-Net framework [150]. It was trained on 108 scans with corresponding manual delineations. By computing the percentage of affected parenchymal tissue, the severity score could be assessed. To determine the CO-RADS score, again the 3D inflated inception architecture [158] was employed.

Breast cancer

Another well researched use case of AI in radiology is breast cancer screening [171, 172]. Randomised trials show reduced mortality from breast cancer after mass screening with mammography leading to a widespread implementation of screening programs. This results in an increased workload for radiologists but also a lot of data. Mammography reading, i.e. finding masses and/or calcifications and identifying them as benign or malignant, is complex and knows large inter and intra-observer variations leading to missed lesions but also to many false positives. False positive testing leads to additional healthcare costs and emotional stress for patients and family. To reduce the error rate, blinded double-reading by two independent readers was introduced in many European countries, increasing the workload even further.

A large publicly available dataset for computer-aided breast cancer screening is the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) on TCIA [51, 173, 174]. It contains mammography data from 1566 participants with corresponding ROI segmentations and verified pathology information. In 2017, the digital mammography DREAM challenge was organised aiming to develop algorithms that can improve early breast cancer detection [175]. Similarly to lung nodule analysis, most state-of-the-art CAD systems for breast cancer screening rely on deep learning algorithms and consist of a candidate detection stage and a classification stage.

Kooi et al. [176] compared a state-of-the-art CAD system relying on manually designed radiomics features with a convolutional neural network (see figure 3.13). Both systems were trained on a large dataset of 45000 mammograms and used the same candidate detection approach. To obtain lesion candidates, a random forest classifier was trained on pixel based first and second order gaussian kernel features. An AUC score of 91% and 93% was achieved with the radiomics approach and

with the CNN respectively. Through combination of the CNN with the manual features, the performance could be improved to an AUC of 94%. Comparison with certified radiologists showed no significant difference in performance.
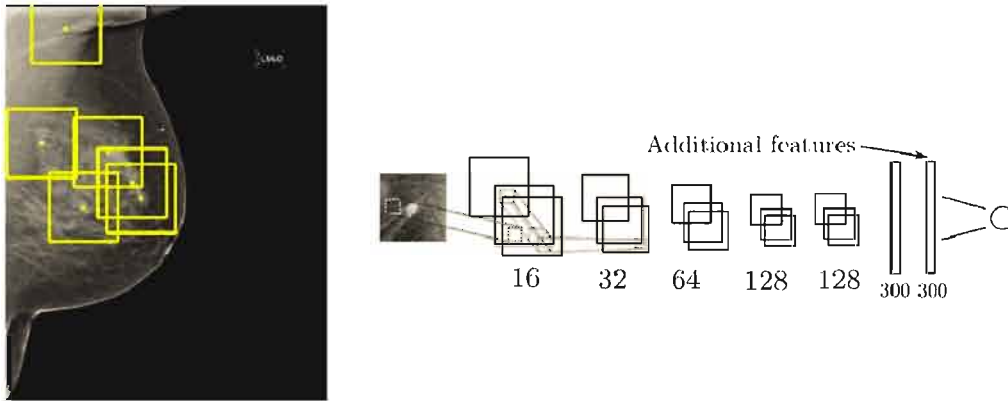
Figure 3.13: Breast cancer mammography screening using a convolutional neural network. Image adapted from Kooi et al. [176] © 2016, with permission from Elsevier.

The first UK company receiving a CE mark for deep learning in radiology is Kheiron Medical Technologies[4]. Their mammography screening system called Mia (Mammography Intelligent Assessment) is allowed to be used as a second reader in breast cancer screening. The deep learning algorithm was trained on more than one million screening mammography images.

Google researchers have presented an AI system that is able to outperform human experts in breast cancer prediction [177]. They proposed an ensemble of three deep learning models operating on different levels of analysis (lesion level, individual breast and the full case). The lesion level model consists of a detection (RetinaNet [178]) and classification stage (MobileNetV2 [179]). The lesion level scores are combined to produce a case level score. For the breast and case level models, ResNet-50 image feature extractors were employed followed by per-breast and per-case concatenation of the feature vectors and further classification. The models were trained and evaluated on data from the UK and USA. Compared to the average radiologist, the AI system achieved a greater AUC score with an a margin of 11.5%. Including the AI system in a double reading process showed that the performance was maintained and led to a reduction in workload of the second reader by 88%. The work

---

[4]https://www.kheironmed.com/meet-mia

by McKinney et al. [177] has, however, been criticised for the opacity and lack of reproducibility of their methods [180].

Cardiovascular diseases

Various imaging techniques play an important role in the diagnosis and management of cardiovascular diseases (CVD) including echocardiography, CT, MRI and nuclear medicine [181]. Artificial intelligence techniques are applied to many cardiac diagnostic applications including myocardial infarction, cardiomyopathies, coronary artery diseases, valvular heart diseases etc. [182, 183]. An important step in the detection and diagnosis of CVD is motion tracking and segmentation of the main chambers [184–193]. This allows quantification of cardiac morphology (e.g. ventricle volumes) and cardiac function (e.g. ejection fraction and wall thickening). Therefore, continuing progress is made for cardiac segmentation enabled by several ongoing challenges such as the Left Ventricle Full Quantification Challenge (LVQuan)[5] and the Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms)[6].

Zheng et al. [191] proposed an automatic method to classify cardiac pathologies such as dilated cardiomyopathy, hypertrophic cardiomyopathy, myocardial infarction and right ventricle abnormality based on cine MRI (see figure 3.14). Given two MR images from a 2D+t cine MRI sequence, apparent flow is estimated using a U-Net type network. Through combination with segmentation, time series of the radius and thickness of myocardial segments are extracted describing cardiac motion. These features are then used to diagnose cardiac pathologies with binary logistic regression classifiers. The model was trained and evaluated on the Automatic Cardiac Diagnosis Challenge (ACDC) dataset [194] and achieved an accuracy of 94%.

The use of machine learning for per-vessel prediction of early coronary revascularisation after fast myocardial perfusion SPECT imaging is studied in Hu et al. [195]. A total of 1980 patients were included from 9 centres in the REFINE SPECT registry. A LogitBoost classifier used 18 clinical, 9 stress test and 28 imaging features to predict early coronary revascularisation. Compared to standard quantitative analysis (total

---

[5]https://lvquan19.github.io
[6]https://www.ub.edu/mnms/
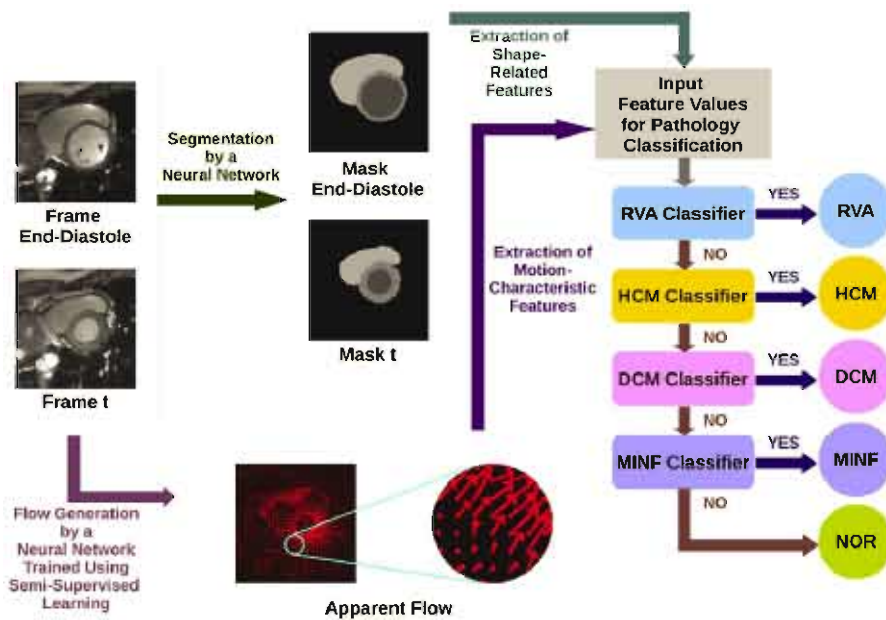
Figure 3.14: Cardiac pathology classification on cine MRI with motion characterisation. Image from Zheng et al. [191] © 2019, with permission from Elsevier.

perfusion deficit), an improvement is achieved with the ML classifier (AUC of 79% versus 71%). The ML algorithm also outperforms expert interpretation by nuclear cardiologists.

In Betancur et al. [196] the potential of deep learning is investigated for prediction of obstructive coronary artery diseases from SPECT myocardial perfusion imaging. The study population comprised of 1638 patients from different institutions. Compared to standard quantitative analysis, the CNN performed better with a per vessel AUC score of 76% versus 73%.

Abdominal diseases

There has been accelerating progress in automated segmentation, detection and diagnosis of abdominal anatomies and diseases [197–203]. This is facilitated by large public datasets like the Medical Segmentation Decathlon [151] and DeepLesion [197] databases.

A universal lesion detector in abdominal CT was developed by Yan et al. [197]. They collected a large-scale dataset composed of CT scans from 4,427 patients containing 32,120 lesions from various anatomical sites including lung, liver, lymph nodes, kidney, bone and so on. Their

proposed lesion detector based on a VGG-16 backbone [45] achieved a sensitivity of 81% with five false positives per image. Figure 3.15 includes an example output of the lesion detector.
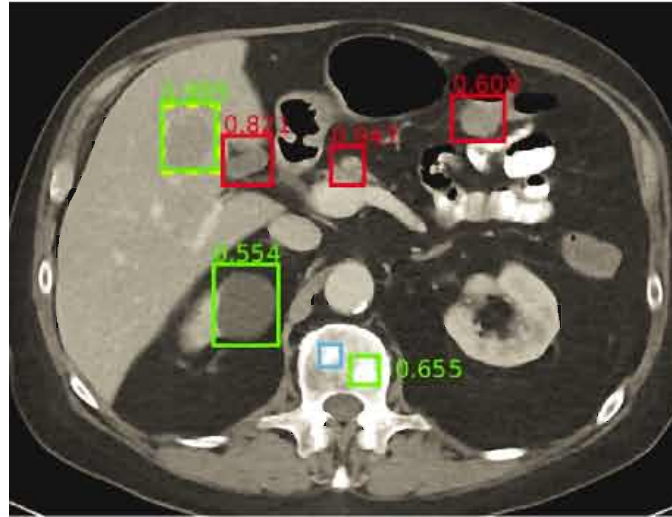
Figure 3.15: Example of the universal lesion detector proposed by Yan et al. [197]. The liver lesion, renal cyst and bone metastasis were correctly identified (green). A bone metastasis was missed (blue). False positives (red) include normal pancreas, gallbladder and bowel. Image reproduced with permission from Yan et al. [197].

AppendixNet, an 18-layer 3D ResNet for detection of appendicitis on CT exams, has been proposed by Rajpurkar et al. [200]. They showed that pre-training the network on a large collection of YouTube videos called Kinetics improved the performance from an AUC of 72% to 81%. The potential of deep learning for non-invasive and automatic kidney function estimation based on ultrasound has been demonstrated by Kuo et al. [201].

Neurological diseases

Application of AI to neuroimaging has seen a lot of interest [204]. Possible tasks include brain age prediction [205, 206], cortical and cerebellum parcellation [207, 208], Alzheimer's disease classification [209, 210], schizophrenia classification [211, 212], intracranial haemorrhage detection [213, 214], aneurysm detection [215–217] and others. A large number of studies address brain tumour analysis which will be covered in detail in part II of this work.

Cerebral aneurysms can cause subarachnoid haemorrhages and early detection is critical for management guidance. Usually CT angiography is used for cerebral aneurysm examination associated with high sensitivity. However, because of the small size of cerebral aneurysms, some may be overlooked during the initial assessment. Yang et al. [217] proposed a deep learning system for aneurysm detection with CT angiography. The detector based on an encoder-decoder architecture with convolutional block attention modules (see figure 3.16) was developed on a large dataset of 1,068 CT angiograms and evaluated on an external test set of 400 CT angiograms. They achieved a sensitivity of 97.5% and conclude that the overall detection performance of radiologists increased with the help of the algorithm.
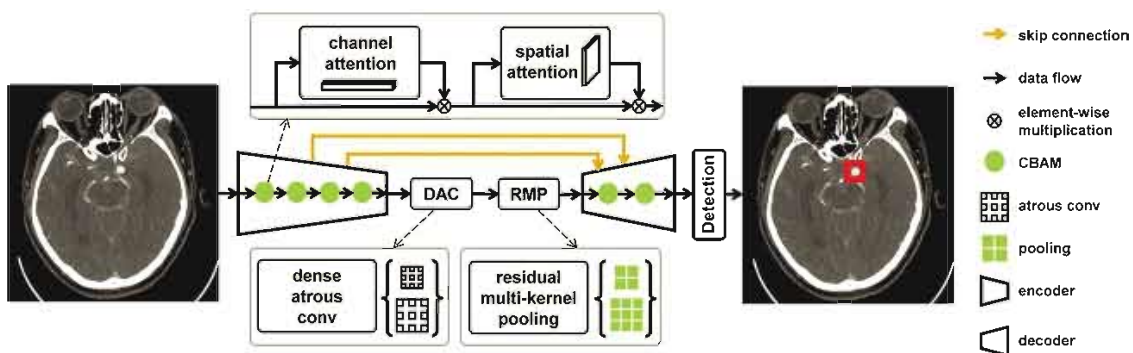
Figure 3.16: The aneurysm detection network proposed by Yang et al. [217]. Image reproduced with permission from supplemental material in Yang et al. [217] © 2021, RSNA.

A deep learning model to predict Alzheimer disease using ^{18}F-FDG PET of the brain was developed by Ding et al. [218]. An InceptionV3 architecture was trained on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [219]. The algorithm achieved an AUC of 98% with a 100% sensitivity and 82% specificity at average of 75.8 months prior to the final diagnosis. The recent approval by the FDA (Food and Drug Administration) of Aducanumab, a drug designed to lower the amyloid plaque burden in the brain should renew the interest of the medical community for amyloid plaque PET imaging. In this regard, DL developed for quantifying amyloid burden with increased accuracy may prove of great value. Further, as several radiotracers are available for that purpose, the approach proposed by Kang et al. [220] for translating the results obtained with [^{11}C]PIB and [^{18}F]Florbetapir into one another, appears highly attractive [220, 221].

Whole-body imaging

Deep learning algorithms are also applied to analyse whole-body PET/CT scans [222]. In Sibille et al. [223], different CNNs were evaluated to detect, localise and classify ^{18}F-FDG-avid foci in whole-body ^{18}F-FDG PET/CT images of patients with lung cancer and lymphoma. The CNNs were trained and evaluated on a dataset of 629 patients (302 with lung cancer and 327 with lymphoma). On the test set, the CNN was able to classify ^{18}F-FDG-positive foci as suspicious or not suspicious of cancer with an AUC of 99% for lung cancer and 98% for lymphoma. The overall localisation accuracy was 96.4% for the body part, 86.9% for the specific region (i.e. organ), and 81.4% for the subregion. A follow up study evaluated the usefulness and performance of the above CNN in research and clinical routine [224]. Automatically segmented total metabolic tumor volumes of diffuse large-B cell lymphoma lesions was predictive for clinical endpoints such as disease-free survival and overall survival. Yet the Dice coefficients between manual and automatic segmentations was only 0.65 in a research cohort and 0.48 in a routine cohort.

Microscopy imaging

Next to the application on non-invasive imaging data, AI techniques are increasingly applied to digital pathology data [225, 226]. Applications range from identifying and segmenting individual primitives such as cancer nuclei and lymphocytes [227, 228] to slide or patient level detection, diagnosis and prediction of mutation status [229–234].

A deep learning system to automatically grade prostate cancer biopsies according to the Gleason grading standard was proposed by Bulten et al. [232]. Biopsies were collected from 1,243 patients. The whole-slide images were pre-processed by previously developed tumour detection [235] and epithelial tissue segmentation models [236]. After patch extraction, a U-Net model was trained to classify each pixel into background, stroma, benign epithelium, Gleason 3, Gleason 4, or Gleason 5. Volume percentages were calculated to assign the final Gleason score and Gleason Grade group. The automated deep-learning system achieved a performance similar to pathologists. Figure 3.17 shows some example grading results.

In 2020, the Prostate cANcer graDe Assessment (PANDA)[7] challenge

---

[7]https://www.kaggle.com/c/prostate-cancer-grade-assessment/overview

was organised on kaggle using data from Bulten et al. [232] and Ström et al. [237]. This resulted in the largest publicly available whole-slide imaging dataset containing almost 11,000 biopsies. Very high performances were achieved with quadratic weighted kappa scores up to 94%.

Kather et al. [234] show that deep learning algorithms can predict mutations, molecular tumour subtypes and immune-related gene expressions directly form routine histology images of tumour tissue. They applied their model on many different cancer types including breast, prostate, head and neck, lung, pancreatic, colon and rectal, melanoma and gastric cancer. Architectures pre-trained on ImageNet were fine-tuned on data from The Cancer Genome Atlas (TCGA)[8] project.

Overlay legend: Benign   Gleason 3   Gleason 4   Gleason 5
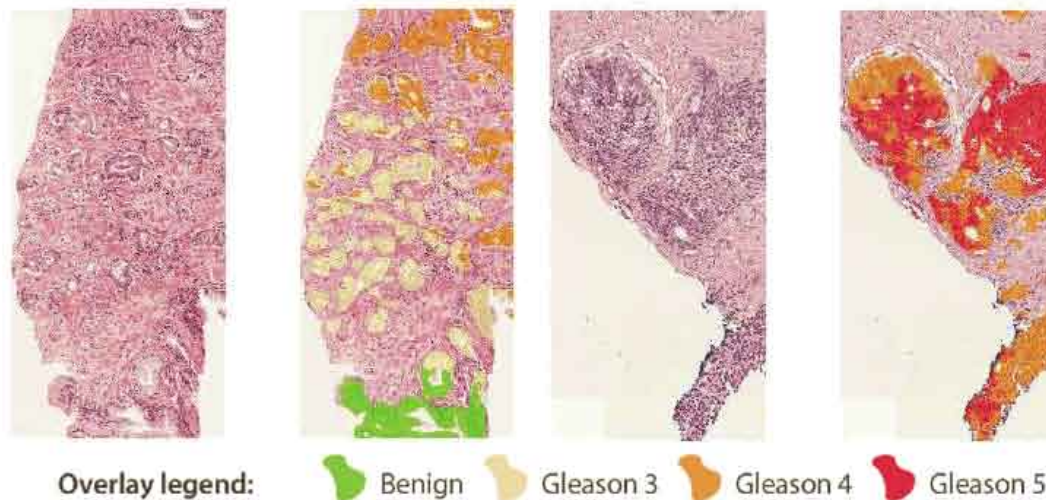
Figure 3.17: Example results from the deep learning prostate grading system proposed by Bulten et al. [232]. Image adapted from supplementary appendix in Bulten et al. [232] © 2020, with permission from Elsevier.

3.5 Conclusion

In this chapter we have provided an overview of medical imaging and the role of artificial intelligence. We started by giving a brief outline of different medical imaging modalities and the principles behind them. Positron emission tomography was explained in more detail as PET detector calibration is the topic of part I of this work. Afterwards, we

---

[8]https://www.cancer.gov/tcga

included an overview of state-of-the-art applications of AI throughout the entire medical imaging chain. Starting with image formation we saw how AI can improve the quality of the raw acquisition data, advance the image reconstruction process and further enhance image quality through post-processing. Furthermore, AI can transform medical image analysis, helping radiologists meet the rising demand for imaging examinations and leverage these large amounts of data towards precision medicine. While challenges remain regarding training data availability, variability of image quality and interpretability, AI systems already achieve high performances in segmentation, detection and diagnosis across numerous anatomical application areas that match or even outperform human radiologists. We can conclude that AI will have a profound impact on radiology and will become and important tool supporting radiologists in their daily work but not replace radiologists.

PART I:

AI IN IMAGE ACQUISITION: PET DETECTOR CALIBRATION

4 | Neural networks for positioning of gamma interactions

In this chapter we perform a comprehensive evaluation on the use of artificial neural networks to estimate the 3D first interaction position (Compton or photoelectric) in a monolithic crystal with a size and thickness typical for clinical systems. Using data from optical simulations, as in a previous study by Stockhoff et al. [80], results are compared with the nearest neighbour algorithm. Performance is evaluated as a function of network complexity and amount of training data and we identify and address potential pitfalls related to neural network training and evaluation. The content of this chapter is published in an A1 publication: Milan Decuyper et al. "Artificial neural networks for positioning of gamma interactions in monolithic PET detectors". In: *Physics in Medicine and Biology* 66 (7 Mar. 2021), p. 075001. ISSN: 0031-9155. DOI: 10.1088/1361-6560/abebfc.

4.1 Introduction

In section 3.2.2 we have seen that detecting gamma rays with high sensitivity and good spatial, timing and energy resolution is the key challenge in PET detector design. To improve these properties, the use of monolithic PET detectors is investigated as an alternative to the pixelated design (see figure 3.7). The main challenge of monolithic detectors is to develop an accurate and efficient algorithm able to determine the gamma interaction position from the measured light distribu-

tion with limited degradation in spatial resolution towards the detector edges. Several positioning algorithms reaching high spatial resolutions were already discussed in section 3.3.1. Nearest neighbour positioning algorithms for example reach state-of-the-art spatial resolutions [60, 79, 80, 239]. However, positioning of an event is computationally expensive as a distance metric needs to be calculated with a large set of reference light distributions.

Our aim is to investigate the use of artificial neural networks to design a gamma positioning algorithm that achieves a superior resolution and is computationally more efficient.
One could argue on the necessity of further improving the spatial resolution of detectors as there are fundamental limitations due to the physics of PET (see section 3.2.2). Positron range is the main factor limiting the resolution of small animal PET systems to 0.54 mm. The resolution of clinical systems is due to their large bore (diameter of 60 to 80 cm) limited by acollinearity to 1.5-2 mm. Taking into account the fundamental PET limitations, an intrinsic detector spatial resolution of 1.6 mm is sufficient for a bore diameter of 70 cm [240]. On the other hand, considering possible applications in organ dedicated and paediatric imaging, systems with smaller bore diameters down to 35 cm could benefit from intrinsic resolutions as good as 0.9 mm [241]. Moreover, when achieving better spatial resolutions than required, there is room to trade resolution for other parameters e.g.: less readout channels, inexpensive materials with less light output, detector thickness, etc.
Besides superior resolution, we also aim for computational efficiency which is necessary to process all events from a large number of detectors at a sufficient rate (which is rather high in PET). This is important to limit the compute time and could be a first step towards live reconstruction.

In this regard, neural networks pose several advantages. First, as universal function approximators, they can directly infer the 3D position from the measured light distribution. No feature extraction is necessary as the networks automatically learn the optimal features from example data. Second, neural networks with continuous output are not restricted to a discrete calibration grid. Finally, once trained, inference by forward propagating events through the network is fast and parallelisable.
Promising applications of neural networks for gamma positioning have been reported (see section 3.3.1). However, when training neural networks, there are many hyperparameters, design choices and vulnerabili-

ties related to overfitting that should carefully be investigated.

Comparison of spatial resolutions reported in different studies with varying positioning algorithms is difficult. The measured resolution is strongly dependent on detector geometry and there is no standardised practice to evaluate the performance of monolithic detectors. Moreover, accurate determination of the intrinsic detector spatial resolution of monolithic detectors is challenging and is for example strongly influenced by the width of the calibration beam. Some groups model the source dimension as a Gaussian distribution and use deconvolution of the beam width to report the intrinsic spatial resolution [69, 242]. A method to measure the intrinsic spatial resolution in monolithic PET detectors based on the convolution of a Gaussian shaped distribution and a modified Lorentzian distribution was proposed in González-Montoro et al. [242].
We evaluate the performance of neural networks on the same geometry and simulation data, acquired with a perfect calibration beam, as in a previous study using mean nearest neighbour positioning. Moreover, we employ the same evaluation methods. This allows a direct comparison between the two positioning algorithms.

4.2 Materials and methods

4.2.1 Detector setup and simulation data

The data used to train the networks is a subset of the data used in Stockhoff et al. [80]. For convenience of reading, some key aspects will be repeated here. For full detail we refer the reader to Stockhoff et al. [80]. A 50 mm × 50 mm × 16 mm LYSO ($Lu_{1.8}Y_{0.2}SiO_5$) crystal is simulated with GATE v8.0 (see figure 4.1). The readout of the crystal is a pixelated 8 × 8 array of 6.07 mm × 6.07 mm SiPM photodetectors. The photon detection efficiency (PDE) of the SiPMs is set to 75% as in Stockhoff et al. [80]. Although a PDE of 75% cannot be obtained with existing commercial SiPMs we opted to use this efficiency as the main goal of this work is to investigate how neural networks can be used to position gamma interactions and how they compare to nearest neighbour positioning. In Stockhoff et al. [80] only a small loss in spatial resolution is reported with combined readout (16 channels) compared to individual readout (64 channels). As combined readout (where rows and columns

are summed) strongly reduces the cost of electronics, we chose to use the 16 channel readout in this work as well. The surface reflections of the crystal are defined by the LUT Davis model [243, 244] (entrance face: polished + optical grease + specular reflector, 3MTMESR; sides: rough + black paint; readout/SiPM face: polished + optical grease). The light distributions are generated by irradiating the detector with a perfectly perpendicular mono-energetic 511 keV pencil beam source at specific positions. Two datasets are acquired. First a 'training dataset' by traversing the entire detector in 1 mm steps. Hence events are acquired in a 49 × 49 grid (blue positions in figure 4.2). In total 10,000 train events and 2,000 evaluation events per position are acquired. The second dataset is an independent 'overfitting dataset' of 10 × 10 intermediate grid points in the detector centre with an offset of 0.5 mm from the 'training dataset' (see figure 4.2). This dataset will be used to evaluate overfitting of the network on the discrete training grid and contains 2,000 events per position split into a validation and test set of 1,000 events per position. All events are individually standardised to zero and unit variance. The x, y and z (depth of interaction) coordinates are given as the position of the first gamma interaction (photoelectric or Compton interaction) as this is the position we want to recover from the given light distribution. Additionally, flood source data was acquired as a rectangular 50 mm × 50 mm 511 keV source with perpendicular irradiation. A total of 840,000 events was collected for this dataset.
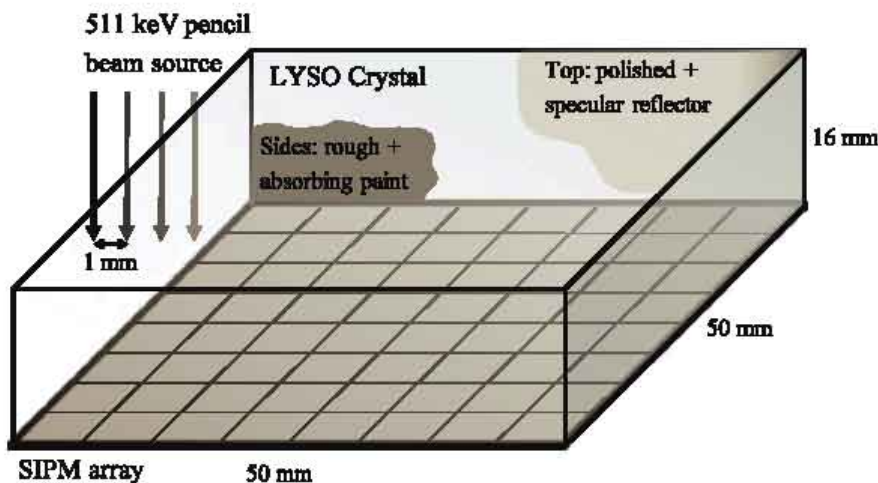
Figure 4.1: Illustration of the simulated detector setup.

On a single core each optical simulation takes around 15 hours with 35000 simulated gamma events per calibration position. The Gate .root-
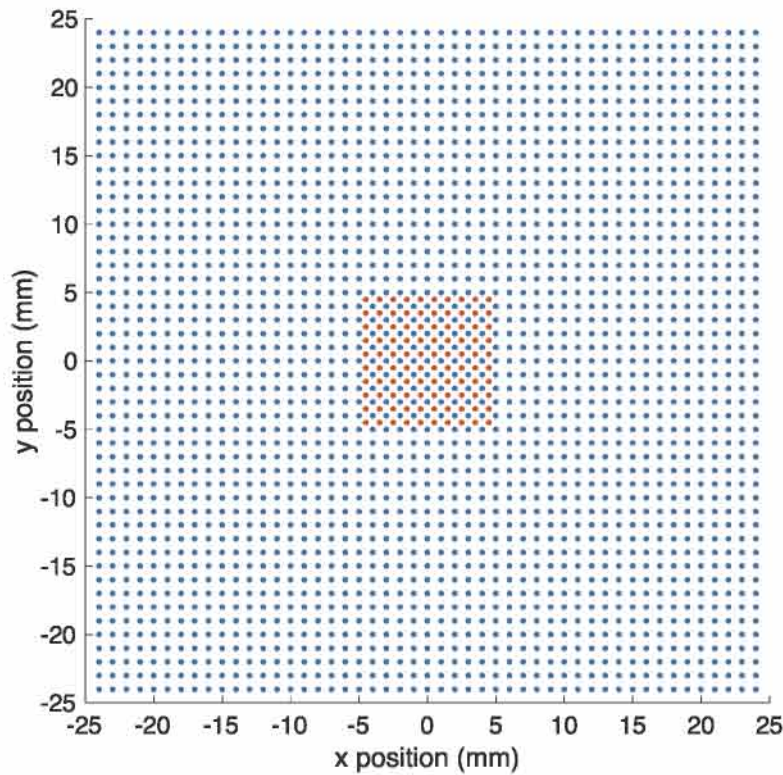
Figure 4.2: Irradiated positions for the two acquired datasets: a 'training dataset' with positions in 1mm steps across the entire detector (blue) and an 'overfitting dataset' in the detector centre with positions at an offset of 0.5 mm from the training positions (red).

file size is about 5GB and includes information on every single photon interaction, energy, time, etc. After extracting only the number of photons per photodetector pixel each file can be reduced to a file size of approximately 20 MBs. The simulation ran on multiple cores on a supercomputer (2 × 18-core Intel Xeon Gold 6240) and could be sped up by taking advantage of the symmetry of the detector. Only 20% of the calibration positions needed to be simulated while the rest of the signals were generated in post-processing steps (see Stockhoff et al. [80]).

4.2.2 Network architecture

A regression artificial neural network is trained to learn the underlying mapping between the measured light distribution and the first (Compton or photoelectric) interaction position. The architecture is a multi-layer perceptron with an input layer, several fully connected hidden layers and finally the output layer. We train one network to predict both

coordinates at once instead of separate networks for every direction. Consequently, the architecture has 16 inputs (channels) and two outputs (x- and y-coordinates) as illustrated in figure 4.3. A third z-coordinate can optionally be added to allow DOI prediction as will be explained in section 4.2.4.

The optimal network architecture should be complex enough to learn the underlying function. On the other hand, a too complex network might overfit on the noise or information specific to the training set and not generalise to new unseen events across the entire detector. To find the optimal network complexity, different architectures are evaluated by varying the number of hidden layers (from 2 to 5) and the number of neurons in each layer (64, 128, 256, 512 and 1024). Every hidden layer is followed by a leaky ReLU activation function to introduce non-linearity as it is computationally efficient and does not suffer from the vanishing gradient problem (see section 2.3.1) [22]. No dropout or batch normalisation is applied. All networks are trained on a training set of 1,000 events/position except in section 4.3.2 where the influence of amount of training data is investigated.
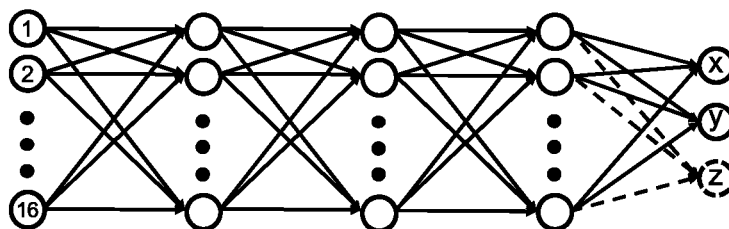
Figure 4.3: Neural Network architecture with 16 inputs, three hidden layers and two (three) outputs x, y (and z).

4.2.3   Training procedure

As briefly discussed in previous section and in chapter 2, the risk of overfitting is one of the main challenges when training deep neural networks. Besides choosing the optimal network complexity, overfitting can be reduced by acquiring enough training data. One of the advantages of using simulation data is that a lot of data can be generated. This allows us to investigate the required amount of training data by evaluating the performance of networks trained on a varying number of training events per calibration position (100 to 8,000 events/position). There is, however, an additional pitfall related to the calibration setup and

corresponding training set. Only data from a discrete set of calibration positions is fed to the neural network. This leads to the potential risk of overfitting on those discrete positions in the training grid. In that case, the network will only output those positions and no intermediate coordinates. When evaluating on an independent test set with data acquired at the same positions, a very small FWHM would be achieved and this overfitting could remain unnoticed. Hence very good spatial resolutions would be reported while the overall spatial resolution is in reality inhomogeneous which is unacceptable. As acquiring ground truth data from all possible positions is not feasible, especially in an experimental setup, training data remains limited to a discrete set of positions. To assess and potentially limit this form of overfitting, we use data acquired at intermediate positions, the red grid in figure 4.2, for validation and testing. By stopping the training early and by choosing the optimal epoch based on the validation loss on these intermediate positions, we can select the best network state before overfitting starts to occur.

Below we report the training strategy used to train all networks. Different hyperparameters were optimised for performance and fixed to properly evaluate the effect of network complexity and training set size. All networks are trained on data that includes both Compton scattered and non-scattered events. The first (Compton or photoelectric) interaction position is used as ground truth. One epoch is defined as an iteration over 100 events/position (240,100 events in total) randomly selected from the total training set. This way, irrespective of the training set size (varying from 100 to 8,000 events/position), the same number of events are processed per epoch. This allows a regular check of the validation performance (after each epoch). Validation after iterating over the entire train set, potentially 8000 events/position, might be too late to assess and prevent overfitting. The network weights are optimised through backpropagation using the Adam optimisation algorithm [25], mini-batches of size 256 and L1 loss. The initial learning rate was set to 0.001 which was halved every 10 epochs that the validation loss did not improve. Early stopping was applied after 40 epochs without improvement. The networks are implemented in PyTorch [245] and trained on an 11 GB NVIDIA RTX 2080 Ti GPU.

4.2.4   DOI estimation

An advantage of monolithic detectors is that depth-of-interaction information is present in the measured light distribution. Hence algorithms can be trained to infer DOI and thereby more precisely position gamma interactions. We extended the 2D positioning network from section 4.2.2 to predict DOI by adding a z-coordinate output as illustrated in figure 4.3. The optimal architecture and number of training events is chosen based on the results of previous two sections. This architecture is then trained to predict the 3D first interaction position.

4.2.5   Evaluation

Several metrics are used to characterise the performance of the positioning networks. All parameters are calculated on a test set with 2,000 events per position in the training grid (blue grid in figure 4.2) and on a test set with 1,000 events/position for the intermediate positions in the detector centre (red grid in figure 4.2). This data includes both Compton scattered (around 60%) and non-scattered events. A 2D histogram of the flood source predictions with a bin size of 0.2 mm was created to assess uniformity. Each metric below is calculated per beam position. Mean and median values are reported over the entire detector and the centre region. Results and discussion will focus on the median values to assess overall positioning performance as outliers can have a large influence on the average performance metrics.

FWHM: For every beam position a 2D point spread function (PSF) of the predicted positions was created with a discretisation bin size of 0.1 mm. A Gaussian was fitted to the line profiles through the maximum along the x- and y- directions from which the full width at half maximum was calculated as a measure of the detector's spatial resolution. For DOI resolution, the FWHM was calculated from a Gaussian fit to the DOI error distribution over all events with a bin size of 0.1 mm.

Euclidean Distance: The mean and median 2D or 3D Euclidean distance between the ground truth beam position and the predicted positions is calculated for every grid location.

Bias: The bias vector contains both distance as directional information on where the algorithm tends to position events from a certain beam location. Hence it illustrates whether the PSF is centred at the correct

location or if, for an edge position for example, it is more directed towards the centre. It is defined as the vector between the ground truth beam location and the average predicted position. The bias vector magnitude is reported, and directions are illustrated in a quiver plot.

4.3 Results

4.3.1 Network complexity

In order to find the optimal network complexity, different neural network architectures are evaluated as a function of number of hidden layers and neurons in each layer. Figure 4.4a shows the obtained median distance measures, calculated over the entire detector on the training grid. In figure 4.4b, the median distance is reported on the intermediate positions in the detector centre to assess overfitting of the networks on the training grid. The median FWHM values are shown in figure 4.5 for the same two regions. When evaluating on the training grid, the distance and FWHM keep decreasing for increasing network complexity. A median Euclidean distance of around 0.48 mm and FWHM of 0.44 mm are obtained with three or more hidden layers and 1024 neurons. On the intermediate positions, the lowest median FWHM of 0.40 mm is achieved with three hidden layers of 256 neurons. For more complex networks, the distance measure saturates and FWHM slightly increases. Overall a median positioning distance lower than 0.51 mm from the ground truth location is achieved starting from a complexity of three hidden layers and 256 neurons. A lower distance and FWHM is observed in figure 4.4b and figure 4.5b compared to figure 4.4a and figure 4.5a. This effect is because subfigures (b) only include points in the detector centre, where we expect better performance, while subfigures (a) include points across the entire detector.

To more closely inspect potential overfitting on the training grid, median FWHM values over the centre [-5 mm, 5 mm] region are reported in table 4.1 for both the training grid positions and the intermediate positions. Additionally, the effect of early stopping through validation on intermediate positions is investigated. When using training grid data for validation, the difference in FWHM between train and intermediate positions is larger, especially for the more complex network with five layers. The FWHM on intermediate positions increases from 0.41 mm to
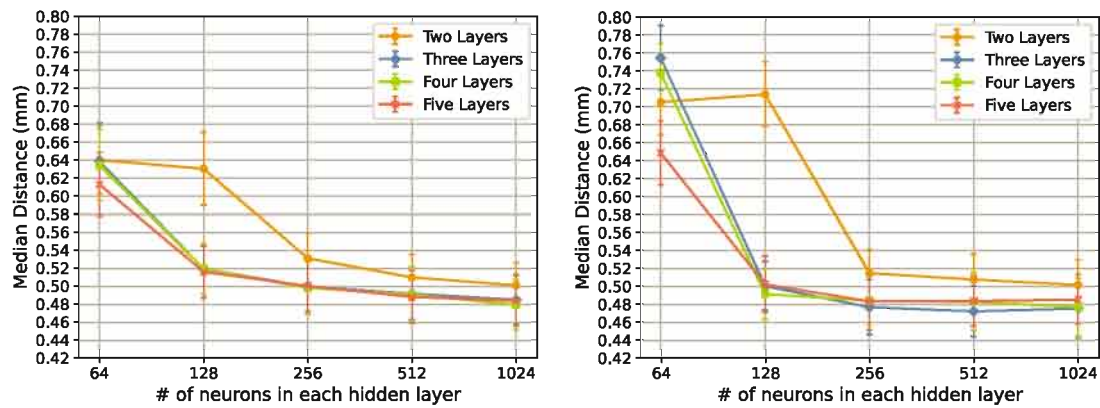
(a) Results calculated over the entire detector on training positions (blue grid in figure 4.2).

(b) Results calculated on intermediate positions in the detector centre 9 × 9 mm^2 (red grid in figure 4.2).

Figure 4.4: Median positioning distance [mm] of networks with varying number of hidden layers and neurons in each layer. Error bars represent the interquartile range.
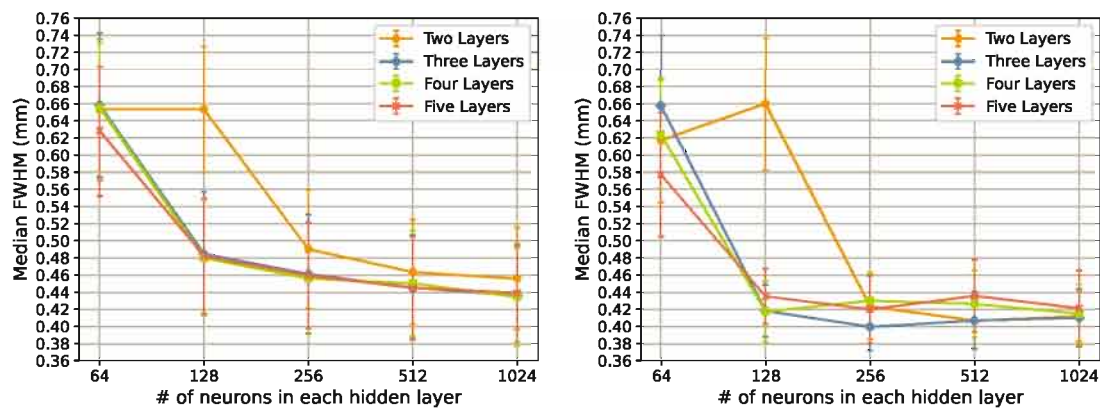
(a) Results calculated over the entire detector on training positions (blue grid in figure 4.2).

(b) Results calculated on intermediate positions in the detector centre 9 × 9 mm^2 (red grid in figure 4.2).

Figure 4.5: Median FWHM [mm] values of networks with varying number of hidden layers and neurons in each layer. Error bars represent the interquartile range.

0.45 mm while a resolution of 0.37 mm FWHM is observed on the train positions. In case of using intermediate grid points for validation and early stopping, the FHWM remains 0.39 mm on train positions and varies between 0.40 mm and 0.43 mm on the intermediate points. Figure 4.6 shows a 2D histogram of the reconstructed flood source positions with the five-layer network of 256 neurons when using training grid positions for

validation (figure 4.6a) and when validating on intermediate positions (figure 4.6b). Without early stopping on intermediate positions, non-uniform positioning can be observed around $x = 0$ and/or $y = 0$.

Table 4.1: Median FWHM [mm] measures for different network complexities illustrating the effect of using training grid positions or intermediate positions for validation and early stopping. The included networks have three, four or five layers with 256 neurons each. Values are calculated in the detector centre [-5 mm, 5 mm] on training grid positions (blue grid in figure 4.2) and intermediate positions (red grid in figure 4.2).

Network	Training positions for validation		Intermediate positions for validation	
	Train Grid	Intermediate Grid	Train Grid	Intermediate Grid
3L - 256	0.37	0.41	0.39	0.40
4L - 256	0.38	0.42	0.39	0.43
5L - 256	0.37	0.45	0.39	0.42
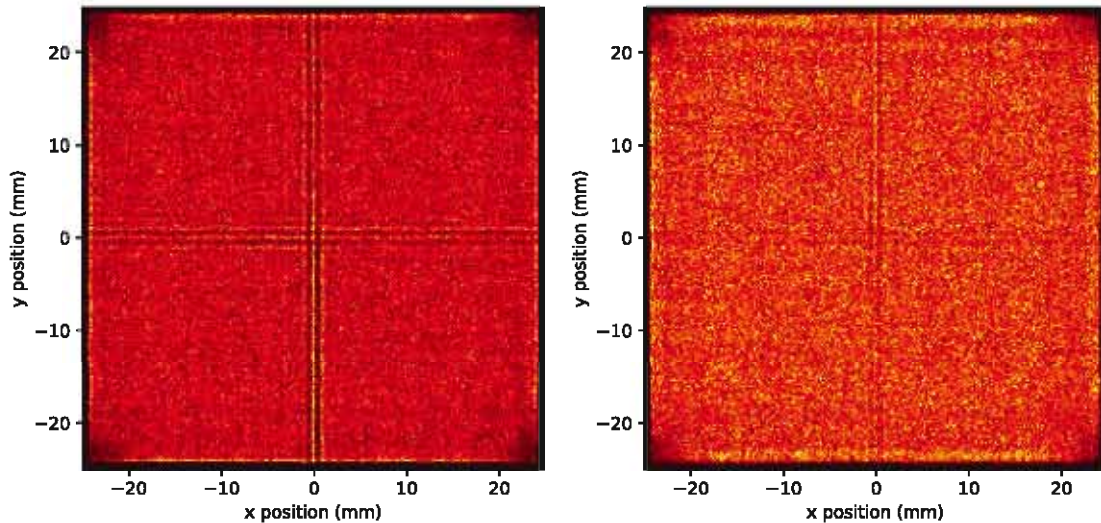
(a) Results when using training grid (red grid in figure 4.2) positions for validation and early stopping.

(b) Results when using intermediate positions (blue grid in figure 4.2). for validation and early stopping.

Figure 4.6: 2D histogram of predicted positions from a flood source for a neural network with five layers of 256 neuron each.

4.3.2 Amount of data

The positioning performance is evaluated as a function of amount of training events per position for several network architectures. This allows us to investigate the optimal training set size. Figure 4.7a indicates the obtained median distance measures calculated on the training grid with the number of training events varying from 100 to 8,000. The median distance on the intermediate grid is reported in figure 4.7b. There is a significant reduction in positioning error from 100 to 1,000 training events after which the median distance remains stable between 0.48 mm and 0.52 mm on the training positions and between 0.47 mm and 0.50 mm on the intermediate positions. More complex networks with 512 neurons in each layer achieve slightly lower distances on the training grid while performance on the intermediate grid remains very similar for all architectures.
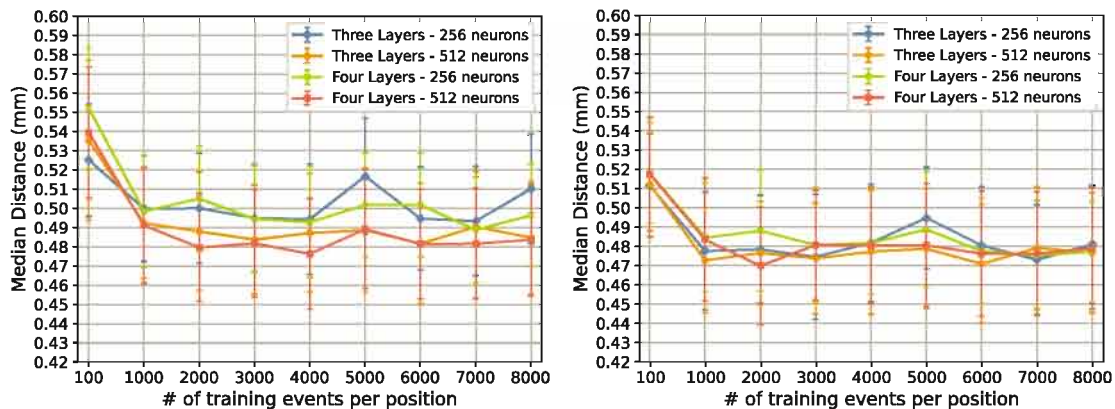
(a) Results calculated over the entire detector on training positions (blue grid in figure 4.2).

(b) Results calculated on intermediate positions in the detector centre 9 × 9 mm^2 (red grid in figure 4.2).

Figure 4.7: Median positioning distance [mm] of networks trained on varying number of training events per position. Error bars represent the interquartile range.

4.3.3 2D Positioning

This section includes a full evaluation of the positioning performance for a network with three hidden layers of 256 neurons trained on 1,000 events per position. The mean and median values of the performance measures explained in section 4.2.5 are reported in table 4.2 on the training grid

(both for the entire detector and centre region) and on the intermediate
grid. Over the entire detector, a mean and median spatial resolution
of 0.50 mm and 0.46 mm FWHM is achieved. In the detector centre,
spatial resolution improves to 0.40 mm. The 2D positioning distance
from the ground truth location is 1.10 mm on average with a median
value of 0.50 mm. In the detector centre and the whole detector, a mean
bias vector magnitude of 0.05 mm and 0.17 mm is observed respectively.
The performance on the 2,000 test events per position in the training
grid is visualised in figure 4.8. Figure 4.8a shows the 2D histogram of
the predicted positions and figure 4.8b denotes a quiver plot with spatial
resolution as a colour scale and bias vectors as arrows. Higher FWHM
values and a larger bias directed towards the centre is observed near the
edges of the detector. A 2D histogram of the reconstructed flood source
positions obtained with the trained network is shown in figure 4.9.

Table 4.2: 2D positioning performance measures [mm] calculated over
different regions of the detector on the training grid (blue grid
in figure 4.2) and on the intermediate grid (red grid in figure 4.2).

	Train Grid Entire Detector		Train Grid Centre 30×30 mm^2		Intermediate Grid Centre 9×9 mm^2	
	Mean	Median	Mean	Median	Mean	Median
FWHM	0.50	0.46	0.41	0.41	0.40	0.40
2D Distance	1.10	0.50	1.11	0.48	1.12	0.48
Bias	0.17	0.09	0.05	0.05	0.05	0.04

4.3.4   Including DOI estimation

The network architecture of the previous section is extended with an
additional output to estimate the depth-of-interaction (z-coordinate).
This network is again trained on 1,000 events per position. The 2D
positioning performance is shown in table 4.3. With a mean FWHM, 2D
distance and bias vector magnitude of 0.50 mm, 1.09 mm and 0.17 mm
respectively over the whole detector, performance is very close to that
of the 2D positioning network (see section 4.3.3). The 3D positioning
performance, calculated over the entire detector, is included in table 4.4.
A mean and median 3D distance of respectively 1.53 mm and 0.77 mm is
achieved. The obtained mean and median absolute DOI error is 0.87 mm

(a) 2D histogram plot of predicted positions.

(b) Quiver plot illustrating FWHM [mm] (colour scale) and bias vectors (arrows).
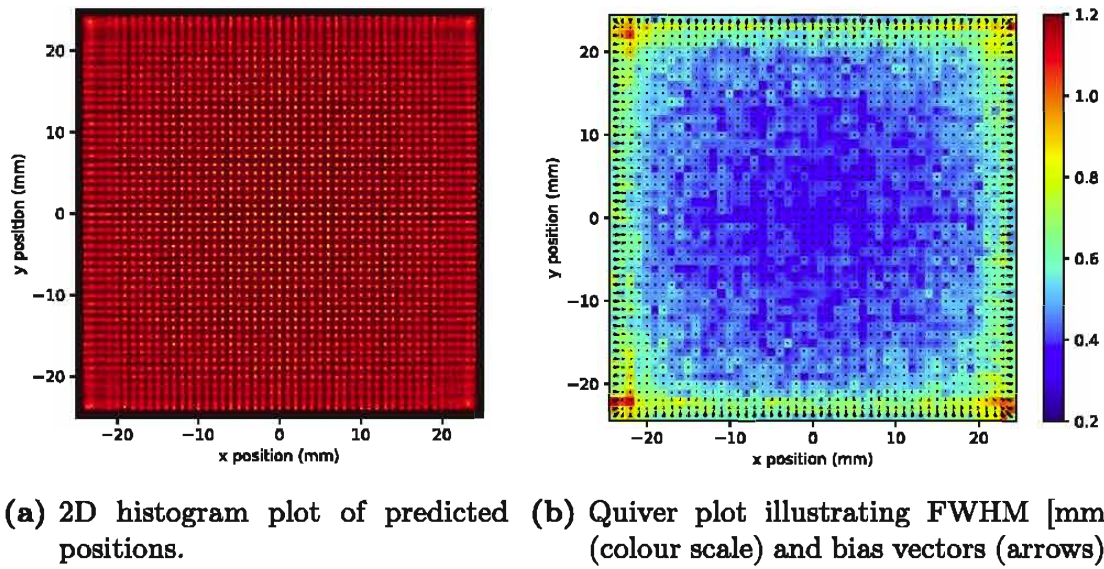
Figure 4.8: Histogram and Quiver plot illustrating 2D positioning performance over the entire detector (training grid).
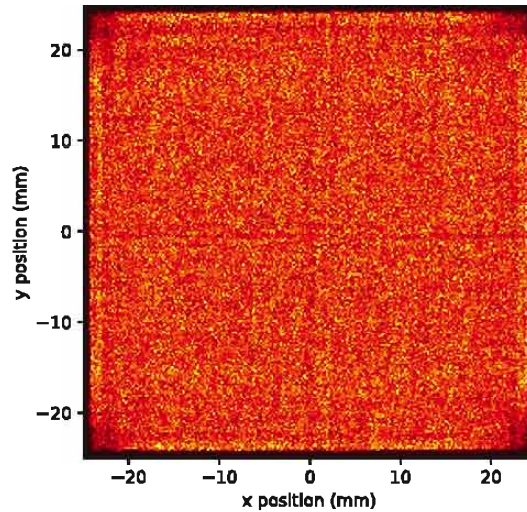
Figure 4.9: 2D histogram of predicted positions from a flood source demonstrating uniform positioning of the network.

and 0.39 mm respectively. Figure 4.10 shows the DOI error distribution with a FWHM of 0.99 mm, calculated through a Gaussian fit (red curve). A negative DOI error indicates that the event is positioned closer towards the SiPM array.

Table 4.3: 2D positioning performance measures [mm] of the 3D positioning network, different regions of the detector on the training grid (blue grid in figure 4.2) and on the intermediate grid (red grid in figure 4.2).

	Train Grid Entire Detector		Train Grid Centre 30×30 mm^2		Intermediate Grid Centre 9×9 mm^2	
	Mean	Median	Mean	Median	Mean	Median
FWHM	0.50	0.46	0.42	0.42	0.41	0.41
2D Distance	1.09	0.51	1.11	0.49	1.11	0.49
Bias	0.17	0.11	0.07	0.06	0.05	0.05

Table 4.4: 3D and DOI positioning performance measures [mm] of the 3D positioning network. Values are calculated over the entire detector on training positions (blue grid in figure 4.2).

	Mean	Median
3D Distance	1.53	0.77
Absolute error DOI	0.87	0.39
FWHM	0.99	

4.3.5 Computational complexity

Computational complexity of the positioning algorithm is important to process events at a sufficient rate. Measured light distributions from incoming gamma rays are processed by forward propagation through the network. This process is parallelisable and can be very fast, especially when using powerful GPUs. For a network with 16 input channels, three hidden layers of 256 neurons with ReLU activation and three outputs, positioning one event takes 272,640 FLOPs. To assess the event rate that can be achieved with the Nvidia RTX 2080 Ti GPU, we propagated 100 million events through the network in batches of 100,000 events. The compute time required by the GPU was 4.7 s resulting in an event rate of over 21 million events per second. Training the above network took around 14 minutes on the GPU.
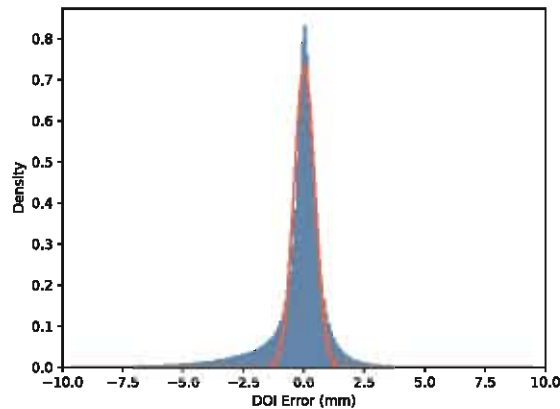
Figure 4.10: DOI error [mm] distribution (blue) and Gaussian fit (red) calculated from events over entire detector on training positions (blue grid in figure 4.2).

4.4   Discussion

4.4.1   Network complexity

The optimal network architecture is assessed in section 4.3.1 through evaluation of spatial resolution as a function of network complexity. Figure 4.4a and figure 4.5a suggest that the best performance measures are achieved with the most complex network architectures. Hence when evaluating on data acquired using the same grid as the network was trained on, a network with four or five hidden layers of 1024 neurons would be chosen as the optimal architecture. On the other hand, figure 4.4b and figure 4.5b reveal that performance saturates and even slightly degrades for the most complex networks on test data from intermediate positions. In terms of median FWHM, a network with three hidden layers of 256 neurons is now the optimal architecture. This illustrates the potential pitfall of overfitting on the training grid positions. While the spatial resolution on the training grid keeps improving for increasing network complexity, it degrades on intermediate positions resulting in non-uniform positioning. The optimal network complexity in relation to overfitting depends on the studied detector geometry and/or the training grid size. If the network is trained on a smaller grid, it is easier for the network to start memorising those positions. Consequently, the network architecture should be carefully chosen depending on the studied detector setup and overfitting should be examined.

While training the network, we used validation data on intermedi-

ate positions to select the best network state and stop training when overfitting starts to occur. This way, we tried to limit the amount of overfitting even if the chosen architecture would be too complex for the task at hand. The effect of validation on intermediate positions is shown in table 4.1 and figure 4.6. When validating on the training grid, the difference in FWHM between train and intermediate positions is larger, especially for the complex network with 5 hidden layers of 256 neurons. Validation on an intermediate grid limits the amount of overfitting but there still remains some difference for complex networks. A reason for this can be that during training, performance on intermediate points might still improve while overfitting already starts to occur. Instead of choosing the network state where the validation loss is lowest, one could select the epoch where training and validation loss is still close. This way, an overall lower performance is achieved but resolution would be more uniform for all positions. A second cause is that L1 loss does not penalise discreteness of the predictions. The loss for discrete or uniform predictions distributed over the same range can be the same. Hence choosing the optimal network state and early stopping based on performance on intermediate positions limits but not entirely prevents overfitting on the training grid. This is also illustrated in figure 4.6. Strong overfitting is observed in the detector centre without validation on an intermediate grid (figure 4.6a). Using intermediate points for validation considerably reduces the non-uniform positioning in the centre as seen in figure 4.6b. This also shows that overfitting occurs earlier in the detector centre than at the edges. Acquiring data at intermediate positions in the detector centre is therefore sufficient to notice overfitting which limits the additional calibration time needed to acquire the overfitting dataset.

4.4.2 Amount of training data

The performance as a function of the amount of training data is denoted in figure 4.7. An improvement in positioning distance from the ground truth is observed from 100 to 1,000 training events per position. For more training events, performance remains stable with small variations within a range of 0.02 mm. These variations are similar to those expected due to differences in initialisation and the stochasticity of training neural networks. For more complex networks with hidden layers of 512 neurons we again see a lower median distance on the training grid (figure 4.7a). Performance on the intermediate grid (figure 4.7b) however is similar for

all networks. These results show that 1,000 training events per position are sufficient to reach the limit of what the neural network is able to infer from data of the simulation setup in this study. Furthermore, taking into account the higher acquisition time when requiring more training events, we have selected 1,000 as the optimal number of training events per position. In an experimental setup, more data might be required due to additional noise from various sources (e.g. electronic noise, scintillator intrinsic activity etc.) in the data.

Based on the results in sections 4.3.1 and 4.3.2, the architecture with three hidden layers of 256 neurons trained on 1,000 events per position was chosen as the optimal network. Results included in sections 4.3.3 to 4.3.5 are therefore obtained using this network.

4.4.3   2D Positioning

A full 2D performance evaluation of the chosen network is included in section 4.3.3. A very good average spatial resolution of 0.5 mm FWHM is achieved across the whole detector and 0.41 mm without the 10 mm border region. As illustrated in figure 4.8, performance degrades towards the edge of the detector where a higher FWHM and a larger bias directed towards the centre is observed. Estimating the interaction position for gamma rays at the edge is more difficult due to truncation of the light distribution. Additionally, as the training data ranges between -24 mm and 24 mm, the network learns to only make predictions within this range. Consequently, the bias is not compensated by predictions beyond the edge positions and therefore directed towards the centre. Evaluation on the intermediate grid shows that the network is not overfitting on the training positions. This is also illustrated by the flood map in figure 4.9, showing uniform positioning across the detector.

A previous paper by Stockhoff et al. [80] evaluates the spatial resolution with a mean nearest neighbour algorithm for the same optical simulation setup. In total 20,000 events per training grid position were used to calculate the reference light distributions which were interpolated to a 0.25 mm grid. Performance was evaluated for different SiPM readouts with 3 mm (32 channels) and 6 mm (16 channels) pixels. With the same readout as in this study, the reported mean FWHM and 2D distance measures are 0.48 mm and 1.73 mm respectively in the central 10×10 mm^2 region of the detector. The performance over the whole

detector is reported for the 32 channel readout with 3 mm pixels. A mean
FWHM of 0.56 mm and 2D distance of 1.41 mm is achieved. Hence with
neural networks, a performance improvement of 11% in FWHM can be
obtained even with less input channels. In the detector centre, with
the same SiPM readout, the mean FWHM improves with 17%. When
comparing 2D histogram plots of positioned events across the 49 × 49
calibration grid, a much more uniform positioning and less artefacts are
observed with the neural network.

4.4.4 Including DOI

The 2D positioning network is extended with an additional z-coordinate
output. We believe that the features or parameters learned by the
network to accurately predict the 2D position are also appropriate for
DOI estimation and therefore no additional complexity or amount of
data is necessary when adding a third coordinate as output. Evaluation
of the 2D positioning performance (see table 4.3) indicates that adding
DOI almost has no effect on the achieved 2D spatial resolution. Perfor-
mance measures are very close to those obtained with the 2D positioning
network (table 4.2). Events are on average positioned 1.53 mm from the
ground truth 3D first interaction position. The mean absolute DOI error
is 0.87 mm and the DOI error distribution has a FWHM of 0.99 mm as
depicted in figure 4.10. The negative DOI error tail is longer indicating
a bias towards deeper DOI estimation. This is possibly due to Compton
scatter as will be discussed in next chapter. Compton scatter could
also explain the large differences between the mean and median distance
metrics suggesting the presence of strong outliers.

In Stockhoff et al. [80], DOI is incorporated by calculating the ref-
erence light distributions from each calibration position for six different
depth layers. To estimate the 3D interaction position, test events are
compared with reference signals from all positions and all six depth
layers. Consequently, DOI estimation is limited to a discrete set of 6
possible depths. For the central 10 × 10 mm² region of the detector, a
mean absolute DOI error is reported of 1.6 mm. With neural networks,
no discretisation into a set of possible depths is necessary and a contin-
uous DOI value can be inferred. This way an improvement in DOI error
is achieved of almost 46%. Moreover, the continuous DOI coordinate
can be estimated by only adding one additional output neuron to the
architecture used for 2D positioning. To increase the precision of DOI

estimation with nearest neighbour more DOI layers have to be added. This, however, increases the computational cost and amount of events that need to be acquired as every layer should contain enough events [80]. The number of DOI layers also influences the planar 2D positioning accuracy with the nearest neighbour approach while for neural networks the 2D resolutions remains very similar as shown in table 4.3.

These simulation results demonstrate the ultimate achievable resolution. In an experimental setup, several factors can degrade resolution such as the calibration beam diameter, lower photon detection efficiency, intrinsic activity of L(Y)SO, additional noise etc. However, it illustrates how neural networks can be trained and used to position gamma rays with very good spatial resolutions that surpass current state-of-the-art algorithms such as mean nearest neighbour. In chapter 6, neural network positioning performance will be evaluated on experimental data and contributing factors degrading spatial resolution will be assessed in chapter 5. The approach of DOI estimation adopted in this study is not directly transferable to an experimental setup. No accurate ground truth depth information of the first interaction can be determined to train the neural network on. Different techniques are investigated to derive DOI information for experimental data [81, 246, 247]. For example, similar to the nearest neighbour approach, DOI could be inferred by dividing events into different depth layers based on variance of the measured light distribution. Furthermore, ground truth DOI data can also be acquired through side irradiation. However, the strong attenuation of the large crystal would result in events being mainly located at the irradiated crystal edge. This can significantly influence DOI estimation performance for events located in the detector centre.

4.4.5   Computational Complexity

The very high event rate of more than 21 million events per second shows that positioning with neural networks can be very fast with appropriate hardware. Using GPUs, positioning is fast enough to process all events from a large number of detectors. This is especially beneficial in total body PET systems or to allow live reconstruction. With mean nearest neighbour positioning, an event is positioned by calculating a distance metric with a large set of reference signals. In the detector setup of this work there are 193 × 193 reference positions per DOI layer after interpolation to a step size of 0.25 mm. This results to in

total 223,494 reference signals. When positioning events on an HPC cluster (2 × 18-core Intel Xeon Gold 6140) with the python scikit-learn *KNeighborsClassifier* algorithm, an event rate of around 20,000 events per second is achieved. Although the event rate is strongly dependent on the used hardware and algorithm implementation, this shows that positioning with neural networks is significantly faster than with mean nearest neighbour positioning.

4.5 Conclusion

In this study, we investigated the use of neural networks for 3D gamma ray positioning in a large monolithic crystal using optical simulation data. Performance was assessed as a function of network complexity and amount of training data. Results show that networks should not be too complex to avoid overfitting and be designed with respect to the calibration setup. Through the use of validation data acquired at intermediate positions that are not in the training set, we could recognise and limit the risk of overfitting on the training grid. Optimal performance was achieved with a network containing three hidden layers of 256 neurons trained on 1000 events per position. Results show that a very high spatial resolution can be achieved, superior to mean nearest neighbour positioning. Finally, forward propagation of events through the network is fast, especially when using GPUs.

5 | Degrading factors

In previous chapter, we have described how neural networks can be trained to position gamma interactions in monolithic PET detectors. Results on simulation data indicate state-of-the-art spatial resolution, superior to mean nearest neighbour positioning. It remains to be investigated which factors influence and potentially degrade the 3D positioning performance of neural networks. Some of these factors are already mentioned such as light truncation and reflections near the crystal edges resulting in degradation of spatial resolution towards the detector borders. This effect was visualised in figure 4.8b, showing the spatial resolution in FWHM across the entire detector. Other potential factors are related to non-idealities which are not modelled in the simulation data like lower photon detection efficiencies, varying SIPM gain, intrinsic activity of L(Y)SO, additional noise in the electronics etc. In this chapter we investigate two factors influencing the spatial resolution: intra-crystal Compton scatter and width of the calibration beam.

5.1 Introduction

Compton scatter

In PET detectors, gamma rays are absorbed by the scintillation crystal and the absorbed energy is re-emitted in the form of light. There are several mechanisms in which photons can interact with matter where the most important types are photoelectric absorption, Compton scattering and pair production [55]. The predominant interaction types depend on the atomic number of the material and the energy of the incident photon. The energy of annihilation gamma rays (511 keV) is below the energy threshold for pair production (1022 keV). We will therefore only consider

photoelectric absorption and Compton scattering.

In case of photoelectric interaction, the total energy of the photon is absorbed by an atom. The absorbed energy results in the ejection of an electron with a kinetic energy equal to the difference between the incident photon energy and the binding energy of the electron. This is the desired form of interaction in the crystal as the photon disappears and only undergoes one interaction that generates the measured light distribution.

Compton scatter refers to the collision of a photon with an electron. As a result, the photon is deflected with a scattering angle θ. A part of the photon energy, related to the scattering angle θ, is transferred to the electron which relaxes by emitting the energy in the form of light (in case of scintillation material). The probability distribution of the photon scattering angle is described by the Klein-Nishina formula [248] and is illustrated in figure 5.1. The relation between scattering angle and the transferred energy is shown in figure 5.2 and given by equation:

$$E_{sc} = \frac{E_0}{1 + \frac{E_0}{m_e c^2}(1 - \cos \theta)}$$

with E_{sc} the energy of the scattered photon and E_0 the energy of the incident photon. It is observed that forward scattering with small angle is more likely than backward scattering for 511 keV photons. Moreover, a higher energy is transferred to the electron with increasing scattering angle. Hence the amount of visible light that is emitted in the crystal through Compton interaction depends on the scattering angle and small angle scattering with little light output is more probable.

Intra-crystal Compton scattering complicates the determination of the correct line of response. Gamma rays can Compton scatter one or multiple times before final photoelectric absorption as depicted in figure 5.3. In L(Y)SO detectors, typically around 60% of the gamma rays undergo Compton scatter [250]. The light yield from Compton interaction is often small compared to photoelectric interaction due to the likelihood of small angle scattering (see figures 5.1 and 5.2). Discerning this small Compton light yield within the total light distribution is difficult and recovering the first interaction position, necessary to resolve the correct LOR, is therefore more challenging for these events. Especially when photoelectric absorption occurs underneath the Compton interaction position as the light distributions of these interactions
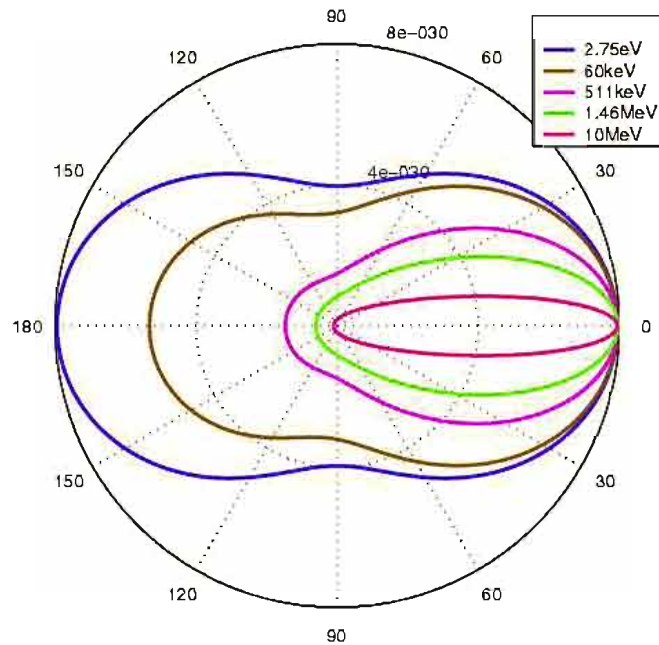
Figure 5.1: The Klein-Nishina distribution of photon scattering angles for different incident photon energies. Incident photon arriving from 180° angle. Image from [249].
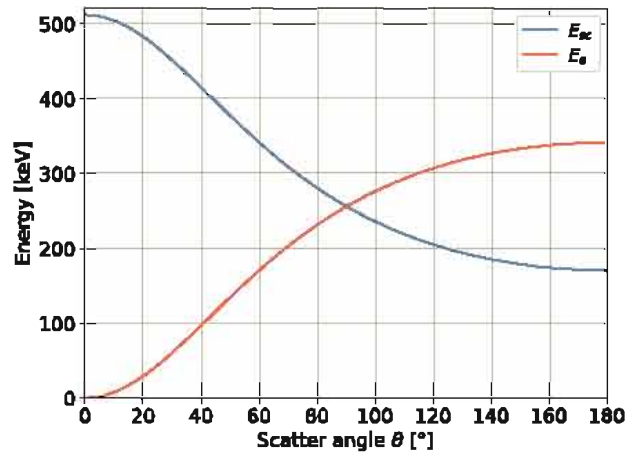
Figure 5.2: Relation between scattering angle and the energy E_{sc} of the scattered photon and kinetic energy E_e of the recoiling electron after Compton scattering of an incident photon with energy E_0 = 511 keV.

overlap. In this case, the 2D coordinates (x,y) remain the same but DOI estimation on the other hand will be affected considerably. The photoelectric interaction closer to the photodetector causes the light

distribution to be more narrow resulting in a deeper DOI estimation.
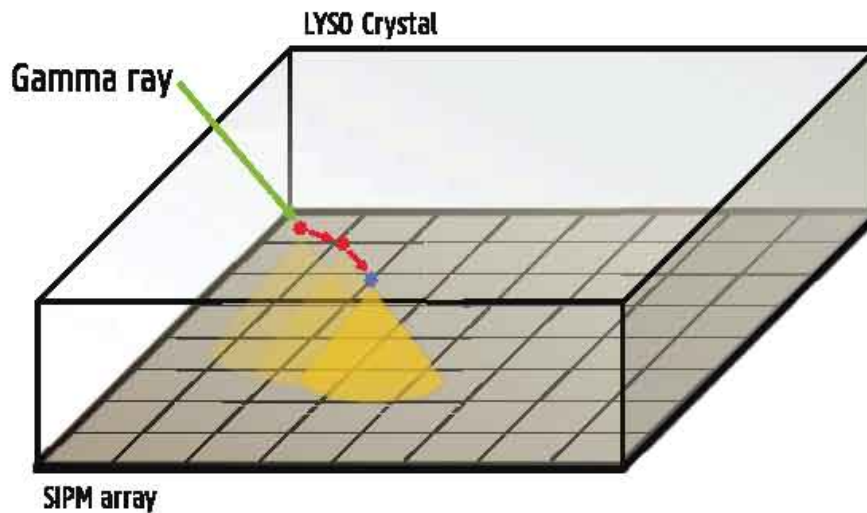
Figure 5.3: Illustration of a gamma ray undergoing two Compton interactions (red stars) with small light yields before final photoelectric absorption (blue star).

Mitigating the effect of Compton scatter on positioning accuracy has been investigated in pixelated PET detectors through intercrystal scatter identification [251–253]. Intercrystal scatter refers to the proces where a photon first Compton interacts in a certain crystal element before absorption in a different (neighbouring) crystal leading to misidentification of the true LOR.

There is only limited research on assessing and mitigating effects of Compton scatter in monolithic PET detectors and most existing positioning algorithms are not optimised to handle Compton scatter.
In the mean nearest neighbour approach, for example, reference light distributions are calculated using the mean signal over all events acquired at a beam location (both Compton scattered and pure photoelectric) [80]. Hence effects of Compton scatter are averaged out and MNN does not take Compton scatter into account. Stockhoff et al. [80] report much lower DOI errors when excluding Compton scattered events from the dataset.
In Li et al. [254], spatial resolution is evaluated separately with and without Compton scatter using maximum likelihood estimation. Simulation data was acquired for a $49.2 \times 49.2 \times 15$ mm^3 LYSO crystal with a 12×12 SiPM array placed at the entrance side of the detector. They show that Compton scatter degrades 2D spatial resolution from 0.80 mm to 0.86

mm FWHM and DOI resolution from 1.05 mm to 1.19 mm.

Iborra et al. [84] trained neural networks to predict the 3D location of the photoelectric interaction. They state that poor results were obtained when evaluating the entire process (training + testing) on predicting the first (Compton or photoelectric) interaction position. Attempts were made to train classification networks that separate Compton scattered from pure photoelectric events with no success. They concluded that the deviations in light distribution caused by Compton scatter are too weak to determine the first interaction position and therefore trained the networks to predict the final photoelectric interaction location.

The neural networks in chapter 4 are trained to predict the first (Compton or photoelectric) interaction position. Hence the used simulation data consist of both Compton scattered and pure photoelectric events. In the simulated LYSO crystal, around 60% of the gamma rays undergo one or more Compton interactions before the final photoelectric absorption. Hence a majority of events are Compton scattered. Using simulation data, the exact number of interactions and their positions are known which allows to investigate the degrading effect of Compton scatter on the positioning performance of neural networks (see section 5.2). Moreover, we examine whether this degradation can be mitigated using Compton scatter detection neural networks and positioning networks, specifically trained for Compton scattered events.

Calibration beam width

In the simulation setup of chapter 4, data was acquired by irradiating the detector with a 511 keV perfect beam source. This means that the emitted gamma rays do not deviate and exactly enter the crystal at the transverse beam location. In an experimental setup, however, it is difficult to replicate this perfect calibration beam with sufficient count statistics. Using a collimator, radiation from a source is collimated into a beam with a certain beam diameter (typically around 1 mm) as illustrated in figure 5.4. Smaller beam widths lead to more precise irradiation but increased acquisition time as it takes longer to acquire a sufficient amount of events.

This beam width results in a broader spread of events acquired at a certain beam location and therefore a measured spatial resolution (FWHM) that is larger than the detector's intrinsic spatial resolution.
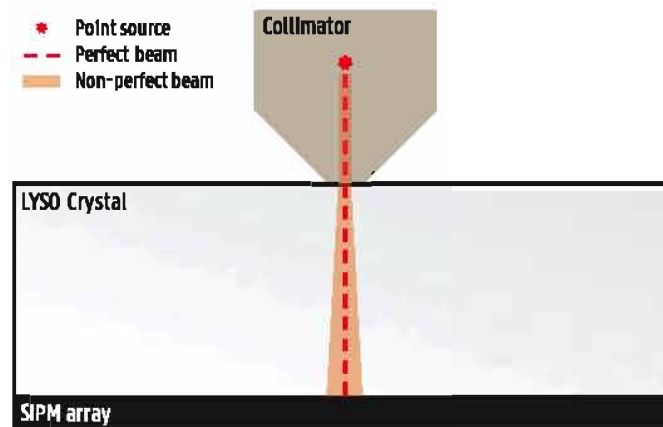
Figure 5.4: Illustration of an experimental calibration setup with a non-perfect collimated beam source. The collimated beam has a certain angular spread resulting in an increasing diameter deeper into the crystal.

For this reason, a degradation in 2D spatial resolution is measured in an experimental setup versus simulation while the intrinsic resolution could be the same. As a consequence, accurate determination of the intrinsic detector spatial resolution is difficult. Some research groups model the source beam profile as a Gaussian distribution and use deconvolution of the beam diameter to report the intrinsic spatial resolution [69, 242]. In González-Montoro et al. [242], the intrinsic spatial resolution in monolithic PET detectors is determined based on the convolution of a Gaussian shaped distribution and a modified Lorentzian distribution.

A non-perfect beam source can also have an effect on the calibration of the positioning algorithms. In case of mean nearest neighbour positioning, the average detector response is calculated over many events acquired at a beam location. Since the beam spread is symmetrical, a broader peak will be observed but the peak location will remain the same. It is expected that this will not have a big influence on the obtainable intrinsic spatial resolution [255]. For neural networks, on the other hand, the beam width could have an influence on the 2D positioning accuracy. All events acquired at a beam location receive the same label (the beam coordinates) even though they possibly enter the detector at a position that deviates within the range of the beam width diameter becomes broader deeper into the crystal (see figure 5.4). Training neural networks on this data with slightly 'wrong' labels introduces additional noise in the training updates which can degrade the obtained positioning

performance.

In section 5.3, the effects of calibration beam width on the spatial resolution of neural networks will be examined. To this end, a source beam with a diameter of 0.6 mm is modelled in the same simulation setup of chapter 4 and simulation data with a non-perfect beam is acquired.

5.2 Compton Scatter

In this section, we investigate the degrading effect of Compton scatter on the positioning performance of gamma interactions in monolithic PET detectors with neural networks. Additionally we examine whether this adverse effect can be mitigated using scatter specific positioning networks or scatter identification networks. The contents of section 5.2.1 were published in Milan Decuyper et al. "Artificial neural networks for positioning of gamma interactions in monolithic PET detectors". In: *Physics in Medicine and Biology* 66 (7 Mar. 2021), p. 075001. ISSN: 0031-9155. DOI: 10.1088/1361-6560/abebfc

5.2.1 Influence on spatial resolution

Methods

The same 3D positioning network is used as obtained in section 4.2.4. This network was trained on the entire simulation dataset containing both Compton scattered (~60%) and pure photoelectric (~40%) events. The positioning performance on all data is included in section 4.3.4. From simulations, the 3D positions from all (Compton and photoelectric) interactions are known. To assess the adverse effect of Compton scatter on the positioning accuracy, the spatial resolution is separately evaluated for Compton scattered and pure photoelectric events. The same evaluation metrics are used as in section 4.2.5.

We additionally evaluate the evolution of the positioning performance (3D distance) as a function of 3D scatter distance. The scatter distance is calculated as the 3D Euclidean distance between the first and final interaction position. For pure photoelectric events this distance is set to zero. It is expected that Compton scattered events with a photoelectric interaction very close to the first interaction will only result in a small

degradation in positioning performance. On the other hand, events that have a final photoelectric interaction position far from the first Compton interaction will have the worst positioning accuracy.

Results

Table 5.1 includes the 2D and 3D positioning performance of the same neural network as in section 4.3.4, separately evaluated for Compton scattered and non-scattered events. A large difference in spatial resolution is observed. For Compton scattered events the median FWHM is 0.66 mm compared to 0.42 mm for non-scattered events. The median 3D distance increases from 0.40 to 1.59 mm and the FWHM of the DOI error distribution increases by 150% (from 0.71 to 1.80 mm). Both DOI error distributions are included in figure 5.5. A much larger negative tail is observed for scattered events indicating that they are positioned deeper than their first interaction position. A visual illustration of the influence of Compton scatter on spatial resolution is shown in figure 5.6 through separate 2D histogram plots for scattered and non-scattered events.

Table 5.1: 2D and 3D positioning performance of the 3D positioning neural network with three hidden layers of 256 neurons for Compton scattered and non-scattered events. Values calculated over the entire detector on training positions (blue grid in figure 4.2).

	Non-scattered		Compton scattered	
	Mean	Median	Mean	Median
2D FWHM	0.46	0.42	0.68	0.66
2D Distance	0.33	0.27	1.66	1.16
2D Bias	0.09	0.07	0.28	0.16
3D Distance	0.49	0.40	2.29	1.59
Absolute Error DOI	0.30	0.22	1.30	0.69
FWHM DOI Error	0.71		1.80	

Figure 5.7 shows the evolution of the positioning performance (3D distance) as a function of 3D scatter distance. The x-axis denotes the scatter distance threshold where events are discarded with a scatter distance beyond the threshold. The right y-axis and blue curve denote the percentage of events that is discarded at each threshold. When removing events that are scattered further than 11 mm (~5%) or 8 mm

(a) Only for non-scattered events.   (b) Only for Compton scattered events.
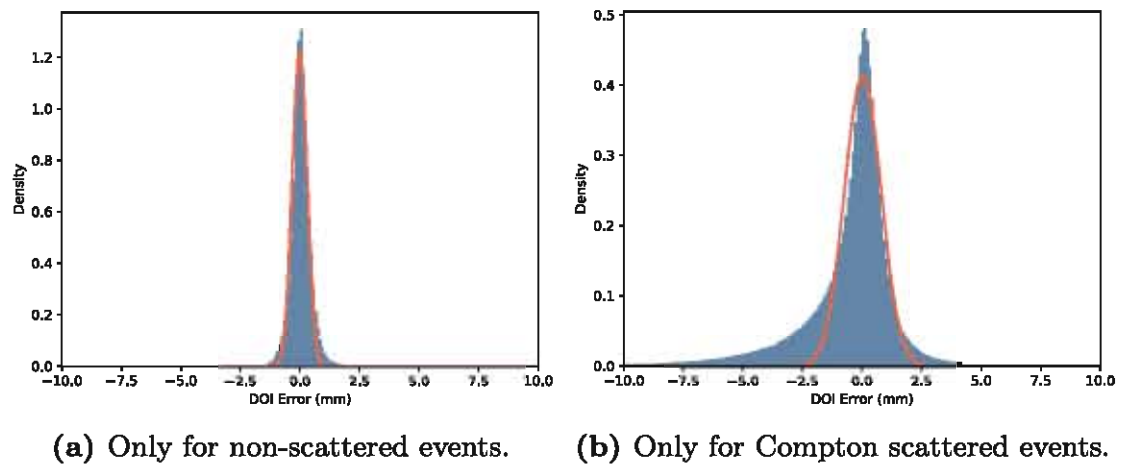
Figure 5.5: DOI error (mm) distribution (blue) and Gaussian fit (red) calculated from events over entire detector.

(a) Positioning only non-scattered (pure photoelectric) events.   (b) Positioning only Compton scattered events.
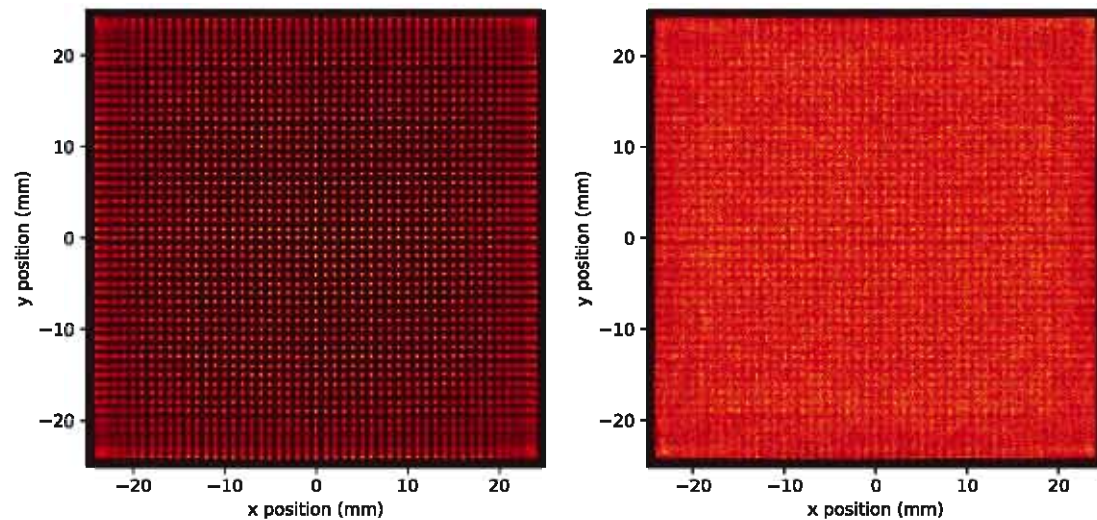
Figure 5.6: 2D histogram plots of predicted positions over the entire detector (training grid).

(~10%), the 3D positioning distance improves from 1.53 mm to 1.36 mm and 1.18 mm respectively.

Discussion

A considerable degradation in positioning accuracy is observed due to Compton scatter. The median FWHM of 0.42 mm for non-scattered
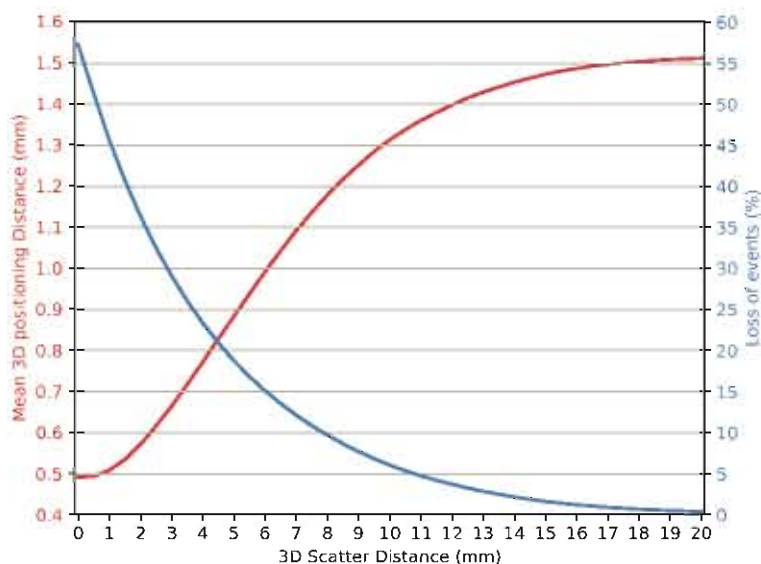
Figure 5.7: Evolution of 3D positioning performance and event loss as a
function of 3D scatter distance. Events are removed that are
scattered with a 3D distance between first and final interaction
position that exceeds the scatter distance threshold.

events increases to 0.46 mm for all events (see section 4.3.4) and to 0.66
mm for only Compton scattered events. Furthermore, the FWHM on
scattered data should be considered carefully as the PSFs show much
larger tails, reducing the goodness of the Gaussian fit which makes
the calculated FWHM values less reliable. These tails cause an over-
all blurring of the image that can be visually perceived in figure 5.6.
Compton scatter strongly influences the mean 3D positioning distance
(0.49 mm versus 2.29 mm). When comparing performance on non-
scattered events (table 5.1) with performance on all (scattered and non-
scattered) events in table 4.4, Compton scatter increases the positioning
error with 93%. The difference between mean and median values is also
much smaller for non-scattered events. Most events undergo small-angle
forward scattering resulting in a final photoelectric interaction that is
closer to the SiPM array than the first interaction. As a narrower light
distribution is measured, the network will position the event deeper into
the crystal. This explains the more prominent tail at the negative side
of the DOI error distribution for scattered events in figure 5.5b and
corresponds with other works that assess the effect of Compton scatter
[80, 254]. Note that data is acquired with a perfectly perpendicular
calibration source. In a realistic setup where gamma rays arrive at

different angles, the degrading effect of Compton scatter on the 2D positioning accuracy could be even larger.

Discriminating Compton from photoelectric events and removing them to improve image quality would considerably decrease sensitivity since the majority of events (around 60%) are Compton scattered. However, the evolution of 3D positioning distance as a function of scatter distance (see figure 5.7) reveals that performance remains similar when adding events with small scatter distances (within 2 mm). The largest fraction of events scatter within a relatively small distance as shown by the distribution (blue curve) in figure 5.7. Around 64% of the events are pure photoelectric or scattered within a distance of 2 mm. Events that scattered further away have a larger degrading effect on the overall positioning accuracy. As only few events have a very large scatter distance, the performance eventually saturates to an overall 3D positioning distance of 1.53 mm. Figure 5.7 illustrates that discarding a small fraction of Compton scattered events (corresponding with the worst positioning performance) can already improve the overal spatial resolution. Training neural networks to identify (far) Compton scatter and potentially processing them with different positioning networks is the topic of next sections.

5.2.2 Scatter specific positioning network

Methods

We now examine whether a neural network specifically trained to position Compton scattered events can improve the positioning accuracy for these events. Two separate 3D positioning networks are trained: one only on pure photoelectric events and a second only on events that Compton scattered at least once. Otherwise the exact same methodology was used as in sections 4.2.3 and 4.2.4, i.e. the same network architecture (three hidden layers of 256 neurons), the same number of training events (1000 events/pos.) and intermediate positions for validation. The same metrics explained section 4.2.5 are used to evaluate the positioning performance.

Results

Table 5.2 includes the positioning performance of the network only trained on pure photoelectric events. The metrics are separately evaluated on

non-scattered and Compton scattered events. This allows comparison with the results in table 5.1 when training a network on all data. On pure photoelectric events, a median 2D FWHM and 3D positioning error of 0.39 mm and 0.36 mm is achieved respectively. On Compton scattered events this degrades to a 2D FWHM of 0.59 mm and 3D Distance of 1.75 mm. Due to the large tails in point spread functions caused by Compton scatter, the Gaussian fits and consequently the FWHM measures are less reliable. The distance metrics are therefore more meaningful. The same evaluation is contained in table 5.3 for the network trained on only Compton scattered events. With this network a median 3D distance is attained of 0.48 mm on non-scattered events and 1.61 mm on scattered events.

Table 5.2: 2D and 3D positioning performance of the 3D positioning neural network with three hidden layers of 256 neurons only trained on pure photoelectric events. Performance is included both on Compton scattered and pure photoelectric (non-scattered) events. Values calculated over the entire detector on training positions (blue grid in figure 4.2).

	Non-scattered		Compton scattered	
	Mean	Median	Mean	Median
2D FWHM	0.43	0.39	0.61	0.59
2D Distance	0.29	0.24	2.16	1.29
2D Bias	0.05	0.04	0.33	0.24
3D Distance	0.43	0.36	2.93	1.75
Absolute Error DOI	0.26	0.19	1.66	0.75
FWHM DOI Error	0.64		1.66	

Discussion

Compared to the results in table 5.1, a similar or even slightly better performance is achieved on pure photoelectric events with the network only trained on these events. Performance on Compton scattered events, on the other hand, is worse. This is not reflected in the (less reliable) FWHM measures but can be noticed by the distance metrics. As expected, the network did not learn to take Compton scatter into account as it is not trained on these events.

The network trained specifically for scattered events achieves a better

Table 5.3: 2D and 3D positioning performance of the 3D positioning neural network with three hidden layers of 256 neurons only trained on Compton scattered events. Performance is included separately for Compton scattered and pure photoelectric (non-scattered) events. Values calculated over the entire detector on training positions (blue grid in figure 4.2).

	Non-scattered		Compton scattered	
	Mean	Median	Mean	Median
2D FWHM	0.51	0.48	0.77	0.77
2D Distance	0.39	0.31	1.65	1.17
2D Bias	0.11	0.08	0.30	0.18
3D Distance	0.59	0.48	2.27	1.61
Absolute Error DOI	0.38	0.28	1.27	0.72
FWHM DOI Error	0.87		1.97	

positioning accuracy on these events. The obtained performance is, however, similar to the network trained for all events. On non-scattered events, the positioning is less accurate compared to the results in tables 5.1 and 5.2.

We can conclude that training a separate network for Compton scattered events does not improve positioning performance. A network trained on all data already learns to optimally process both scattered and non-scattered events. The reduced performance without Compton scatter in the training data does show that the network learns to identify Compton scatter and take this into account to obtain a better estimation of the first interaction position. This can be a reason why neural networks achieve a better positioning performance than mean nearest neighbour algorithms as shown in chapter 4. In mean nearest neighbour positioning, the effects of Compton scatter are averaged out when calculating the reference light distributions while neural networks are trained on the individual events.

5.2.3 Scatter identification

In previous section, we learned that training scatter specific positioning neural networks does not improve the positioning performance of Compton scattered events. Figure 5.7 shows that the overall positioning error can be reduced by discarding events with the largest distance between

first (Compton) and final (photoelectric) interaction, associated with the worst positioning accuracy. For this, an algorithm is necessary that is able to identify these far scattered events. One could tackle this problem as a classification or regression task.

In case of classification, a scatter distance threshold needs to be chosen that separates all events into two classes: non- or close-scattered events and far scattered events. This threshold determines a tradeoff between positioning accuracy and sensitivity. Performance improves when discarding more events but results in an increasing loss of sensitivity. As PET scanners rely on coincidence detection, a loss of detector sensitivity translates into a quadratic loss of sensitivity on scanner level. Choosing this threshold is therefore difficult and might be application dependent. If one wants to adapt the resolution-sensitivity tradeoff to specific applications, different classification algorithms need to be trained for each desired threshold.

To allow a more user friendly and more precise control of resolution versus sensitivity, a regression approach can be beneficial. Here the algorithm is trained to estimate the distance between the first and final interaction. Now the user can adapt the scatter distance threshold depending on the application and all events that are predicted by the algorithm to have scattered further than this threshold are discarded.

For these reasons we opted to tackle the scatter identification problem as a regression task and investigate whether a neural network can be trained to predict the scatter distance of incoming events in a monolithic PET detector.

Methods

Using the simulation data from chapter 4, we train a neural network to predict the 3D distance between the first (Compton) interaction and final (photoelectric) interaction position. We use the same optimal network architecture as obtained for event positioning, i.e. a network with 16 input channels and three hidden layers of 256 neurons with leaky ReLU activation (see figure 5.8). As this network architecture proved to be optimal for event positioning, we expect that it will be also suited for scatter distance prediction. For an initial proof of concept, we chose not to perform a compute intensive grid-search optimisation as in chapter 4. The network now has one output value, the predicted 3D scatter distance.
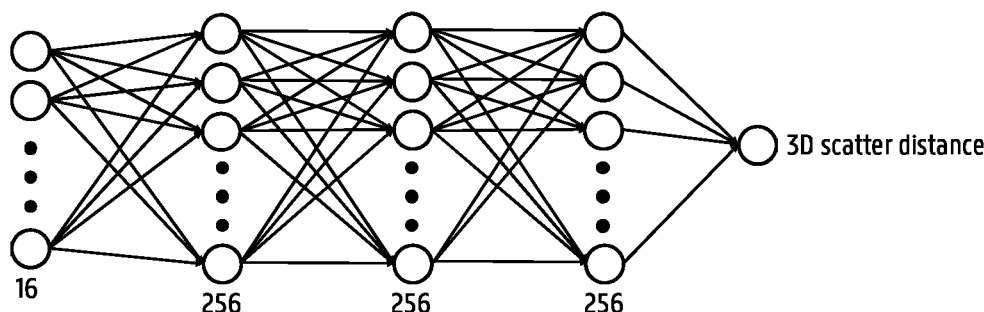
Figure 5.8: Neural Network architecture to predict scatter distance with 16
inputs, three hidden layers of 256 neurons and one regression
output.

The network is trained using the training dataset acquired across the
entire detector (see section 4.2.1 and figure 4.2). From simulations, the
first and final interaction positions in the crystal are known and the 3D
euclidean distance between them can the calculated. These distances
(in mm) are used as ground truth labels to train the network. The
dataset is strongly imbalanced in terms of scatter distance. Around 40%
of the events are pure photoelectric which corresponds with a scatter
distance of 0 mm. This strong data imbalance can cause the network
to systematically underestimate the scatter distance and always predict
values close to zero. To limit this, the network is only trained on
Compton scattered events, i.e. all pure photoelectric events with a
scatter distance of zero are removed from the training set. Since close-
scattered events have a similar light distribution to non-scattered events,
we expect that the network will also learn to attribute a scatter distance
that is close to zero to pure photoelectric events. From the remaining
events, 2,401,000 (1000 per calibration position) are used for training and
an additional same amount for validation. The performance is evaluated
on the same testset as used to evaluate the positioning performance in
chapter 4 and contains 2,000 events per calibration position. For the test
set, the pure photoelectric events are not removed and performance is
evaluated on all events.

Furthermore, the network is trained using the AdamW optimisation
algorithm, mini-batch size of 256 events, MSE loss and L2 weight decay
set to 0.01. One epoch is defined as an iteration over 240,100 events,
randomly sampled with replacement from the entire training set. The
initial learning rate was set to 0.001 and halved every 10 epochs that
the validation loss did not improve. Early stopping was applied after 60

epochs without improvement. The network was implemented in PyTorch [245] and trained on a MacBook Pro with a 2.8 GHz Quad-Core Intel i7 CPU.

Results

The scatter distance prediction network achieves a mean absolute error of 1.93 mm and a median absolute error of 1.82 mm. Similar to figure 5.7, the evolution of 3D positioning distance as a function of predicted scatter distance is shown in figure 5.9. For the same thresholds of 11 mm and 8 mm as in section 5.2.1, the 3D positioning distance reduces from 1.53 mm to 1.42 mm and 1.31 mm respectively. The corresponding loss of events is 4% at 11 mm and 8% at 8 mm. When discarding 5% (at 10 mm threshold) or 10% (at 7 mm threshold) of the events, the positioning distance decreases to 1.39 mm and 1.26 mm respectively. The classification performance at these two thresholds is included in table 5.4. High specificities are achieved of 98.7% at 7 mm and 99.3% at 10 mm. The sensitivities are lower: 70.4% and 73.9% respectively.
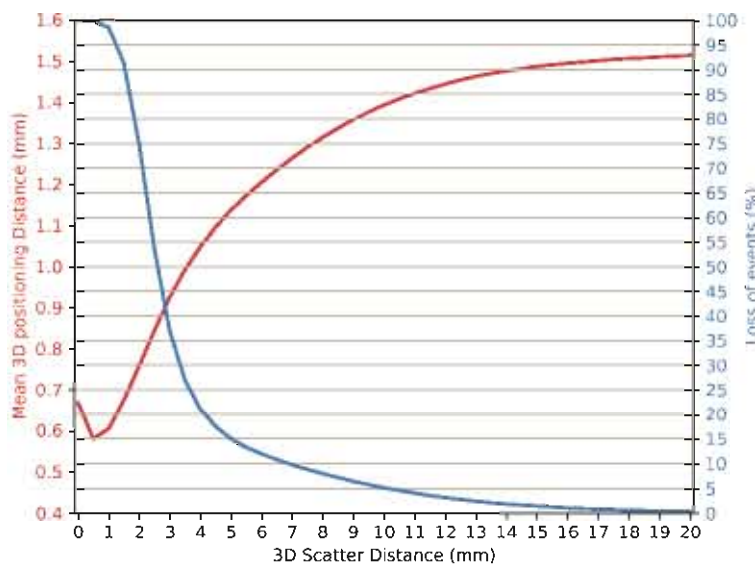
Figure 5.9: Evolution of 3D positioning performance and event loss as a function of predicted 3D scatter distance with scatter distance prediction network. Events are removed that are predicted as scattered with a 3D Euclidean distance that exceeds the scatter distance threshold.

Table 5.4: Classification performance of the 3D scatter distance prediction
network for two scatter distance thresholds.

Threshold	Event loss	Accuracy	MCC	Sensitivity	Specificity
7 mm	10%	95.21%	76.05%	70.35%	98.65%
10 mm	5%	97.66%	78.25%	72.85%	99.27%

Discussion

From the above results we can conclude that neural networks are able
to estimate the scatter distance between the first and final interaction
position of an incoming gamma ray in a monolithic crystal. This demon-
strates that the network is able to identify the Compton and final pho-
toelectric light yield, especially for far scattered events. The evolution
of positioning performance as a function of predicted scatter distance
(figure 5.9) shows a similar trend as observed in figure 5.7 when using
the ground truth scatter distances. Note that now all events (Compton
scattered and pure photoelectric) are included and the event loss goes
from 0% to 100% instead of 60% resulting in a steeper curve for small
scatter distances. When removing the 5% or 10% furthest scattered
events, an improvement in 3D positioning distance of around 9% and
18% can be attained respectively. Overall, the network slightly under-
estimates the scatter distances of far scattered events (illustrated by
the lower distance threshold for the same event loss percentages). This
results in a very high specificity and lower sensitivity (see table 5.4). A
high specificity is preferred as we want to be sure to only discard the far
scattered events associated with worst positioning performance.

These results on simulation data demonstrate the possibility of using
neural networks to identity Compton scatter and to tune the spatial
resolution versus sensitivity tradeoff to specific application dependent
situations. For applications that require high resolution with margin
to reduce sensitivity, more far scattered events can be removed. The
question remains how to implement this in an experimental setup. No
information on scatter distance is known for experimental calibration
data so we cannot train a scatter distance prediction network on real
detector data. A network trained on simulation data could be applied to
experimental data but factors that are not modelled in simulation such
as variable SiPM gain, electronic noise, lower PDE etc. would reduce

the effectiveness. Application of neural networks on experimental data will be investigated in chapter 6.

5.2.4   Bayesian positioning network

Discerning the small light yield of the first Compton interaction within the total light distribution is challenging as it overlaps with other light spreads of later interactions. The position estimation of an intra-crystal Compton scattered gamma ray should therefore be associated with a higher uncertainty than of a pure photoelectric event. This leads to a possible different approach to (far) scatter identification through uncertainty modelling with Bayesian neural networks (see section 2.3.4). The uncertainty related to Compton scattering is input dependent and can thus be identified as heteroscedastic uncertainty.

In this section we investigate whether heteroscedastic uncertainty modelling with the predictive variance method explained in section 2.3.4 allows to identify events associated with uncertain and less accurate position estimation. Furthermore, we asses whether the obtained uncertainty scores are correlated with scatter distance.

Methods

We employ the same optimal network architecture, number of training, validation and test events and training procedure that was found in chapter 4. An additional output is added to the network to predict the variance as illustrated in figure 5.10. We train the network to predict the log variance $s = \log \hat{\sigma}^2$ as it is numerically more stable and avoids potential division by zero [32]. The adopted training loss over N samples is:

$$L_{BNN} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \exp(-s_i) \, \text{MAE}(p_i, \hat{p}_i) + \frac{1}{2} s_i$$

with p_i and $\hat{p}_i$ the ground truth and predicted positions of the first interaction. To evaluate the achieved spatial resolution, we use the same evaluation metrics as in section 4.2.5.
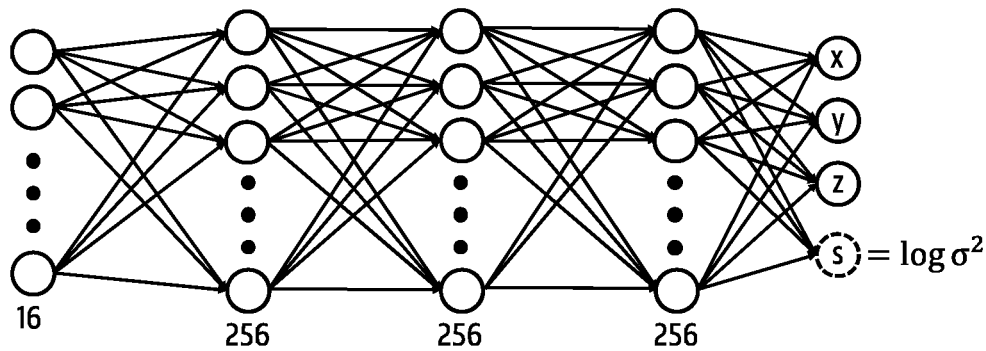
Figure 5.10: Neural Network architecture to predict the 3D first interaction position and heteroscedastic uncertainty. The network consists of 16 inputs (measured light distribution), three hidden layers of 256 neurons, three position outputs x, y and z and one output s representing the log variance.

Results

The 2D and 3D positioning performance of the 3D positioning neural network trained with heteroscedastic uncertainty modelling is included in table 5.5. A median FWHM of 0.44 mm and mean 3D positioning distance of 1.54 mm is achieved across the entire detector. The evolution of mean 3D positioning distance as a function of predicted variance is plotted in figure 5.11. When discarding 5% or 10% of the events (with highest variance), the mean distance reduces to 1.36 mm and 1.23 mm respectively. Figure 5.12 shows how the 3D scatter distance changes in relation to the predicted variance. The mean scatter distance decreases when events with high variance are removed. A Spearman's correlation coefficient of 0.64 is measured between 3D scatter distance and predicted variance of Compton scattered events.

Discussion

The performance measures in table 5.5 indicate a spatial resolution that is very close to that of the original network without uncertainty modelling (see section 4.3.4). So adding an additional uncertainty output and adapting the loss does not alter the achieved positioning accuracy.

A similar trend is observed in 3D positioning distance when removing highly uncertain events compared to removing far scattered events. This shows that events with a high positioning error are also predicted with a

Table 5.5: 2D and 3D positioning performance of the 3D positioning neural network trained with heteroscedastic uncertainty modelling (see figure 5.10). Values calculated over the entire detector on training positions (blue grid in figure 4.2).

	Mean	Median
2D FWHM	0.47	0.44
2D Distance	1.10	0.49
2D Bias	0.17	0.11
3D Distance	1.54	0.74
Absolute Error DOI	0.88	0.38
FWHM DOI Error	0.94	
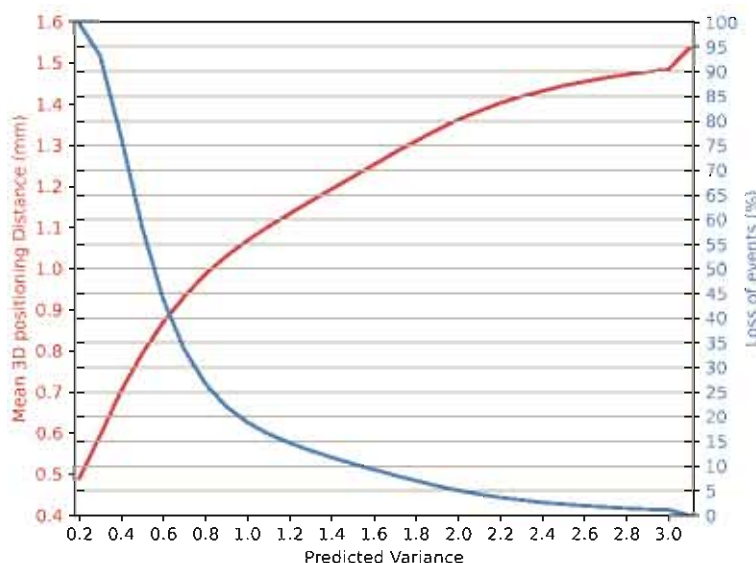
Figure 5.11: Evolution of 3D positioning distance and event loss as a function of predicted variance. Events are removed associated with a high position estimation uncertainty that exceeds the variance threshold.

high uncertainty. When removing 5% or 10% of the events with highest predicted uncertainty, an improvement in 3D positioning distance of around 12% and 20% can be obtained respectively. Training a positioning neural network with predictive variance to model heteroscedastic uncertainty therefore allows to tune the spatial resolution versus sensitivity tradeoff in monolithic PET detectors.

Figure 5.12 and the measured Spearman's correlation coefficient reveal
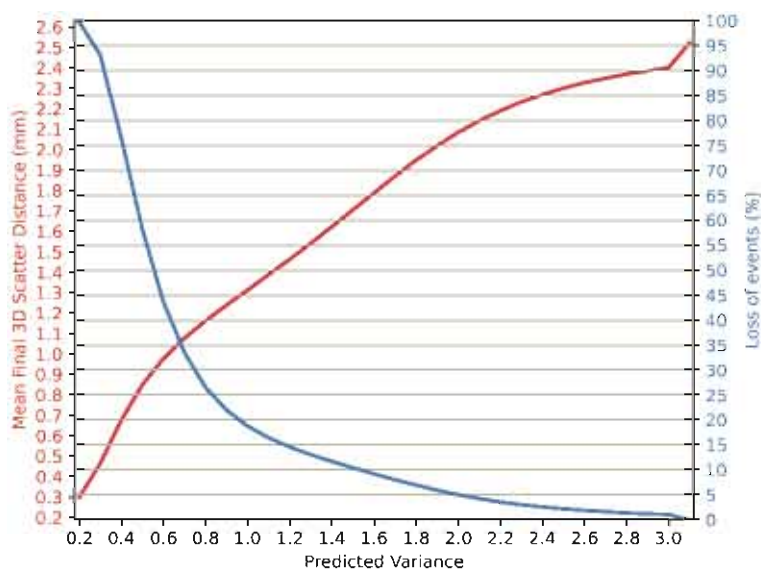
Figure 5.12: Evolution of 3D scatter distance and event loss as a function of predicted variance. Events are removed associated with a high position estimation uncertainty that exceeds the variance threshold.

that the predicted uncertainty scores are correlated with scatter distance. As expected, the modelled heteroscedastic uncertainty, which is input dependent, incorporates ambiguity related to Compton scattering. The network learned to recognise Compton scattered events and associate this with a high uncertainty in the position estimation.

Filtering events associated with high positioning error mainly reduces the long tails of the point spread functions. This has limited influence on the Gaussian fits and consequently the measured FWHM values. For this reason, we used the mean 3D Euclidean distance as metric to asses the improvement in performance when filtering uncertain events.

Next to Compton scatter, heteroscedastic uncertainty could also incorporate position-related uncertainty related to the edge effect. This type of uncertainty is also input dependent as events near the detector border are more difficult to position than events in the centre. However, plotting the uncertainty as a function of position and calculation of the correlation coefficient between uncertainty and calibration position did not reveal a significant correlation. The contribution of event location to the predicted uncertainty is therefore much smaller than the contribution of Compton scatter.

In this section we demonstrated that state-of-the-art positioning ac-

curacy can be achieved with Bayesian neural networks and that the predicted uncertainty scores allow to further improve the resolution by trading sensitivity. No pre-filtering of events with a different neural network or other algorithm is necessary and only one network needs to be trained. Furthermore, the uncertainty is learned from the available position labels and loss so no additional information on Compton scattering is required. This methodology is therefore applicable to an experimental setup as well which will be investigated in chapter 6.

5.3 Calibration beam width

In this section, we evaluate the effect of calibration beam source width on the evaluation and training of neural networks for positioning of gamma rays in monolithic PET detectors.

5.3.1 Materials and methods

Instead of using simulation data collected with a perfect beam source as in chapter 4, we now use simulation data acquired with a beam source that has a diameter of 0.6 mm at the entrance side of the crystal. The simulation model for the non-perfect beam is illustrated in figure 5.13. Except for the calibration beam, the same PET detector setup and data acquisition methodology is used as in section 4.2.1. The data was acquired by Mariele Stockhoff and the non-perfect beam model is described in Stockhoff et al. [255].

Two 2D positioning neural networks are trained with different ground truth labels. For one network, the x- and y-coordinates are set to the true first interaction (Compton or photoelectric) position as obtained from simulation. Because the source beam has a diameter of 0.6 mm, these coordinates can deviate slightly form the beam positions (grid locations as shown in figure 4.2). The other network is trained on position labels where the x- and y-coordinates are set to the beam location, similar to what is done in an experimental setup. Through comparison of the positioning performance of both networks, we can assess the influence of beam width on the evaluation and training of neural networks. Both networks are evaluated on data acquired with perfect beam (see section 4.2.1) as well as data with the non-perfect beam. The same optimal network architecture (three hidden layers of
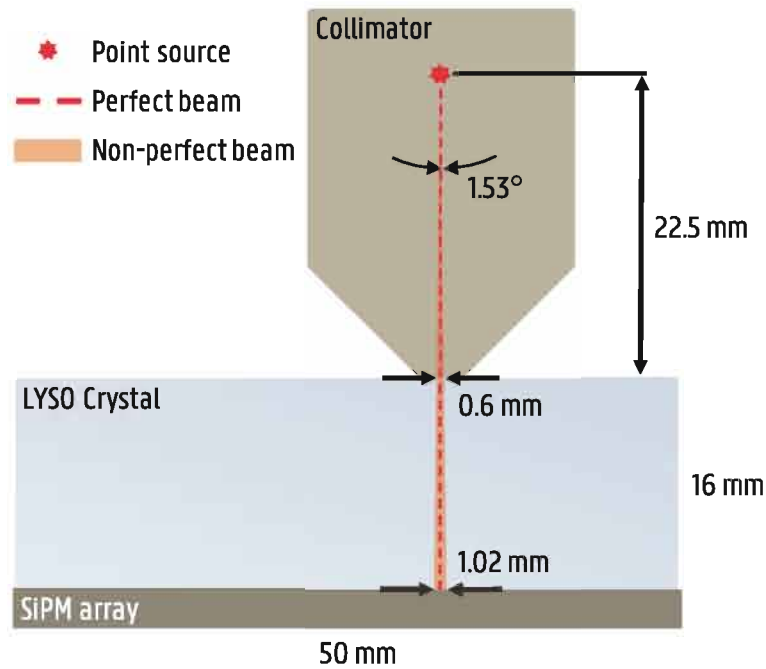
Figure 5.13: Illustration of a non-perfect beam modelled as a collimated point source with an angular spread of 1.53° and collimator diameter of 0.6 mm [255].

256 neurons), training procedure and number of training events (1000 per position) was used obtained from chapter 4.

5.3.2   Results

Table 5.6 shows the 2D positioning performance of the network trained on non-perfect beam data with true first interaction coordinates as ground truth labels. Results are included both on the perfect beam data as on the 0.6 mm beam data. On the perfect beam data a median FWHM is achieved of 0.46 mm over the entire detector and 0.41 mm when excluding the 10 mm border region (only detector centre). On the 0.6 mm beam data, the FWHM measures are higher: 0.72 mm and 0.69 mm on the entire detector and centre region respectively.

The 2D positioning performance of the network trained on non-perfect beam data with beam location as ground truth labels for x- and y-coordinates is included in table 5.7. Now a spatial resolution on perfect beam data is obtained of 0.48 mm FWHM over the entire detector and 0.43 mm FWHM in the centre region. These values are slightly higher

Table 5.6: 2D positioning performance of the neural network trained on 0.6 mm beam data with true first interaction positions as ground truth labels. Values calculated on training positions (blue grid in figure 4.2).

		Perfect beam		0.6 mm beam	
		Mean	Median	Mean	Median
Entire Detector	Distance	1.10	0.51	1.14	0.58
	FWHM	0.51	0.46	0.74	0.72
	Bias	0.16	0.09	0.18	0.11
Detector Centre 30×30 mm²	Distance	1.13	0.49	1.16	0.56
	FWHM	0.41	0.41	0.69	0.69
	Bias	0.05	0.05	0.05	0.04

(0.02 mm) compared to the results in table 5.6. The attained median FWHM scores on the 0.6 mm beam data are 0.72 mm (overall) and 0.69 mm (centre). Figure 5.14 visualises the performance on the non-perfect beam data across the detector through a histogram and quiver plot.

Table 5.7: 2D positioning performance of the neural network trained on 0.6 mm beam data with beam location as ground truth labels for x- and y-coordinates. Values calculated on training positions (blue grid in figure 4.2).

		Perfect beam		0.6 mm beam	
		Mean	Median	Mean	Median
Entire Detector	Distance	1.11	0.52	1.15	0.60
	FWHM	0.52	0.48	0.74	0.72
	Bias	0.16	0.10	0.18	0.11
Detector Centre 30×30 mm²	Distance	1.14	0.50	1.16	0.57
	FWHM	0.43	0.43	0.69	0.69
	Bias	0.07	0.06	0.07	0.06

5.3.3   Discussion

Results in table 5.6 show that when using the true first interaction positions for training, the network achieves the same intrinsic spatial

(a) 2D histogram plot of predicted positions.

(b) Quiver plot illustrating FWHM [mm] (color scale) and bias vectors (arrows).
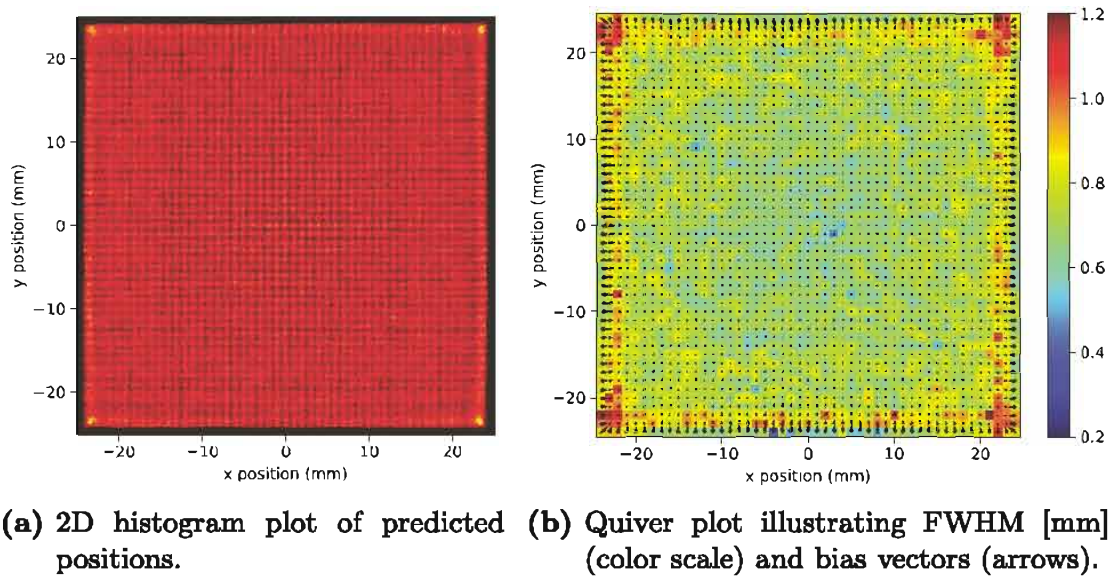
Figure 5.14: Histogram and Quiver plot illustrating 2D positioning performance over the entire detector (training grid) on non-perfect beam data with diameter of 0.6 mm. Results are obtain with neural network trained on 0.6 mm beam data with beam location as ground truth labels for x- and y-coordinates.

resolution as in table 4.2 when training on perfect beam data. When using the beam location instead of the true interaction coordinates for training, a slight degradation in intrinsic spatial resolution of 0.02 mm can be observed (median FWHM of 0.48 mm versus 0.46 mm). The difference is, however, very small and we can conclude that neural network training is robust to the 'noisy' labels associated with non-perfect beam data. The acceptable width of the calibration beam has of course also a limit. We expect that the beam width should be comparable or smaller than the intrinsic spatial resolution.

As expected, higher FWHM values are measured on the 0.6 mm beam data due to the broader spread of events. This can visually be observed when comparing figure 5.14 with figure 4.8. Evaluating the intrinsic resolution in an experimental setup with a non-perfect source beam is therefore difficult. The obtained FWHM measures are an underestimation of the true intrinsic resolution and should be taken into account when evaluating an experimental setup which will be done in chapter 6.

These observations correspond well with a similar study by Stockhoff et al. [255] when using mean nearest neighbour positioning.

5.4   Conclusion

In this chapter we have assessed the effect of intra-crystal Compton scatter and calibration source beam width on the positioning performance of neural networks.

A considerable degradation in spatial resolution was observed due to Compton scatter. It was noticed that the positioning error depends on the scatter distance and that a small percentage of far scattered events are associated with the highest positioning error. We therefore investigated whether scatter-specific positioning networks or networks to identify far scattered events could help to improve performance. A network specifically trained to position scattered events did not result in an improvement. This shows that a network trained on all (pure photo-electric and Compton scattered) events already learns to take Compton scatter into account. The fact that neural networks can learn to identify (far) Compton scatter was further supported by training a network to predict the 3D scatter distance. This network could be used to filter far scattered events in order to improve spatial resolution with a tradeoff in sensitivity which can be justified in certain applications. Considering the limited practicality of training a scatter prediction network in an experimental setup (no available labels), a different approach was investigated using a Bayesian neural network. This method allows to train one network to predict both the position as the positioning uncertainty related to Compton scatter without requiring additional information on Compton scattering.

Comparison between a network trained on data acquired with a perfectly narrow beam versus a calibration source with a realistic beam width showed no significant difference in achieved intrinsic spatial resolution. The beam diameter does, however, influence the measured spatial resolution which should be taken into account when evaluating and comparing spatial resolution of different PET detectors.

6 | Application on experimental data

In this chapter, we validate the methodology of training neural networks for positioning of gamma interactions in monolithic PET detectors developed in chapter 4 on experimental data. The spatial resolution is evaluated for two PET detector setups. One has the same design as the simulated detector in chapters 4 and 5. The other design is a PET detector with a smaller size (35 mm × 35 mm) and thickness of 12 mm. Results on the first PET detector design are published in Mariele Stockhoff et al. "High-resolution monolithic LYSO detector with 6-layer depth-of-interaction for clinical PET". in: *Physics in Medicine and Biology* 66 (15 Aug. 2021), p. 155014. ISSN: 0031-9155. DOI: `10.1088/1361-6560/ac1459`.

6.1 Introduction

The optimal training procedure and architecture of neural networks for positioning gamma interactions in a monolithic PET detector was investigated in chapters 4 and 5 using simulation data. The methodology and spatial resolution remains to be validated on experimental data. Factors that are not taken into account during simulation can complicate neural network training and reduce performance. Potential degrading factors are: background activity of LYSO, electronics noise, variable SiPM gain, lower photon detection efficiency, none mono-energetic calibration beam, irregularities in surface finish etc.

Evaluation of the gamma interaction positioning performance of neural networks for real PET detectors is the topic of this chapter. The

methodology developed in previous chapters is applied on two detector designs with different geometries. Performance is again compared with mean nearest neighbour positioning.

6.2　Materials and methods

Overall the same methodology is used as developed in chapter 4 with some alterations in data acquisition and DOI estimation. Below we first explain the geometry, data acquisition en data pre-processing steps of the two detector setups. Afterwards we briefly repeat the neural network training procedure.

6.2.1　Experimental setup

An experimental setup was built with the same design as the simulated PET detector in chapter 4. The design and data acquisition was performed by Mariele Stockhoff and details can be found in Stockhoff et al. [256]. For convenience of reading, some key aspects are repeated here.

Detector design

The detector consists of a $50 \times 50 \times 16$ mm^3 LYSO crystal (Epic Crystal). A rough black painted finish is applied on the sides and a black painted specular reflector on top. The crystal is coupled with optical grease to an 8×8 array of 6×6 mm^2 SiPMs. The 64 SiPM signals are combined to a readout of 16 (8+8) channels by summing rows and columns. The signals are amplified, digitised and pulse integration is performed on an FPGA after which the signal is transmitted to a computer for further processing.

Data acquisition and pre-processing

For calibration, a ^{68}Ge source is placed in a tungsten collimator. Data is acquired with two different collimators, one with a diameter of 0.6 mm and an other with a 1 mm diameter. The beam source perpendicularly irradiates the detector which is mounted on a robot stage to acquire data at discrete positions. Similar to the simulation setup, a training data set
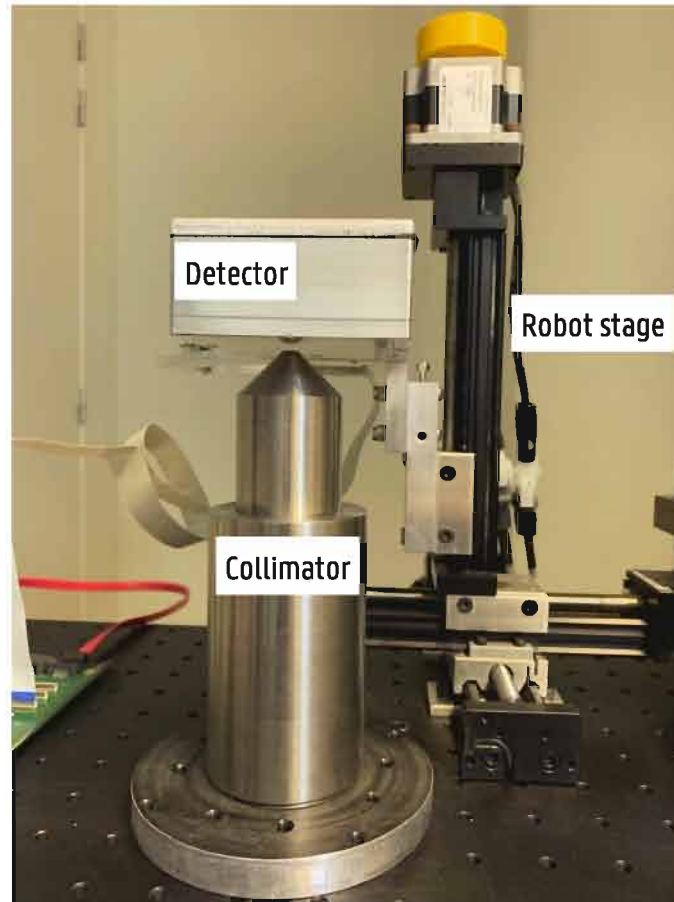
Figure 6.1: Experimental calibration setup consisting of collimator and a detector mounted on a 3D robot stage [256].

is acquired in a 49 × 49 grid across the entire detector (blue grid in figure 4.2) and an overfitting data set at 10 × 10 intermediate positions in the detector centre (red grid in figure 4.2) for validation.

During data acquisition, events are collected that originate from the calibration source and from the intrinsic activity in LYSO. The network should only be trained on true gamma events from the calibration beam. To this end, the events are pre-positioned with Anger logic and a region of interest is drawn around the calibration beam position. The ROI is defined as the 109 connected highest intensity pixels in the Anger histogram around the ground truth position. This resulted in an ROI with a diameter of about 3 mm. Only events that are pre-positioned within this region of interest are extracted. For evaluation, this ROI filtering is not applied and performance is evaluated on all events. Furthermore, energy filtering is applied with a 20% energy window and each event is standardised to zero mean and unit variance.

The training and validation sets contain 1,000 events/position (blue grid and red grid in figure 4.2 respectively) and the test set contains 30,000 events/position. For the training and validation sets, two versions are acquired: one with the 0.6 mm calibration beam and another with the 1 mm beam. Consequently, two networks are trained to evaluate the influence of calibration beam width on the achievable spatial resolution as in section 5.3. Both networks are evaluated on the 0.6 mm beam data as this data is closer to the ground truth beam location.

The x- and y-coordinates are set to the beam position. In contrast to the simulation setup, the exact z-coordinates of the first interaction position is not known for experimental data. Therefore, a similar approach is used to obtain DOI information as proposed in Stockhoff et al. [80] for mean nearest neighbour positioning. The events are divided into six depth layers based to their standard deviation across the 16 channels. For MNN positioning, the amount of DOI layers also influences the 2D spatial resolution and six DOI layers was determined as the optimal number. According to the depth distribution derived from the Beer-Lambert attenuation law, the 28.85% events in the dataset with smallest standard deviation across the channel values belong to layer 1 (furthest from the SiPM array, see figure 6.2). The next 22.03% belong to layer 2, 17.4% to layer 3, 13.8% to layer 4, 10.85% to layer 5 and 8.07% to layer 6 with largest standard deviation. This way the DOI layer can be determined for every event and is used as the z-coordinate. To evaluate DOI performance, evaluation data acquired with the 1 mm beam with ROI filtering is used to limit the influence of background events.
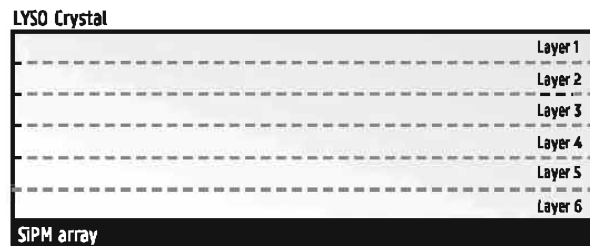
Figure 6.2: Illustration of virtual DOI layers defined to divide events into six depth dependent groups.

To assess uniformity of the positioning across the detector, a 'flood source' data set is acquired by placing the ^{68}Ge source at a distance of 52.5 cm of the detector. Finally, a bar phantom data set is collected to qualitatively asses the capability to resolve adjacent bars. The bar

phantom consists of four quadrants with bars that are separated with different spacings: 0.6 mm, 0.8 mm, 1.0 mm and 1.2 mm. Both flood source and bar phantom data sets are energy filtered with a window of 20%.
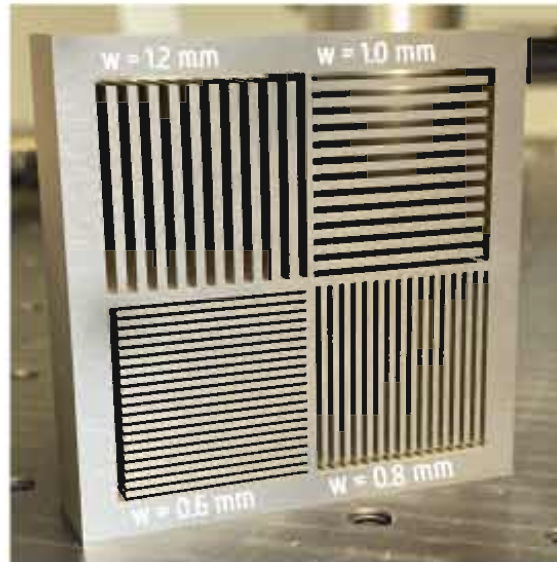
Figure 6.3: Picture of four-quadrant tungsten bar phantom. Each quadrant consists of bars separated with different spacings as indicated on the figure.

6.2.2   12 mm thick PET detector

The second PET detector design consists of a 34.9 mm × 34.9 mm × 12 mm LYSO crystal coupled to an 8 × 8 SiPM array. No multiplexing is performed so all 64 SiPM channels are read. To assess the effect of multiplexing, this is performed afterwards in software by summing rows and columns. The spatial resolution is then evaluated with and without multiplexing. Accordingly, two networks are trained, one with 64 input neurons (without multiplexing) and an other with 16 inputs.

Two datasets are acquired as depicted in figure 6.4. For the first dataset, events are collected with a ^{68}Ge source placed in a 1 mm beam collimator that traverses the detector in a 35 × 35 grid with positions spaced 1 mm apart. This dataset is used to train the network and 1,000 events were used per calibration position. The second dataset is acquired with a 0.6 mm collimator in a 17 × 17 grid with 2 mm spacing (red grid in figure 6.4). These positions are offset by 0.5 mm with respect to the positions of the first dataset. The second data set is split into a validation

set of 500 events/position (to prevent overfitting) and a test set of 1,000 events/position for final evaluation. Events are pre-processed through offset and gain correction, energy filtering, LYSO background activity filtering and standardisation to zero mean and unit variance.
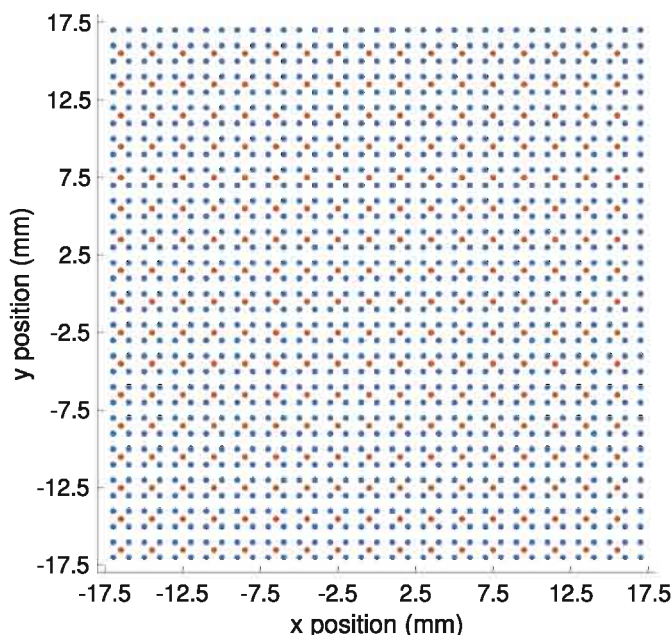
Figure 6.4: Irradiated positions for the two acquired datasets of the 12 mm thick detector. A 'training dataset' with positions in 1 mm steps across the entire detector (blue) and an 'overfitting dataset' with positions in 2 mm steps at an offset of 0.5 mm from the training positions (red).

The collected events are now grouped into seven depth-of-interaction layers based on signal standard deviation. All layers now contain an equal amount of events. As more events interact at the top of the crystal, the first layers correspond with a smaller DOI range than the subsequent deeper layers.

6.2.3 Neural network training

The same network architecture is used as in chapter 4 with three hidden layers of 256 neurons, leaky ReLU activations, and three outputs (x- and y- coordinates and DOI layer). The network is trained using the Adam optimisation algorithm with initial learning rate of 10^{-3}, mini-batch size of 256 events, L2 weight decay set to 10^{-2} and L1 loss. One

epoch is defined as an iteration over 100 events per calibration position when training for the 16 mm thick PET detector and 10 events/position for the second, 12 mm thick detector. As the training set of the second detector contains less discrete positions, it was opted to more regularly evaluate on the intermediate validation data to check for overfitting. Based on the validation loss, learning rate is halved every 10 epochs without improvement and training is stopped if the loss did not improve for 50 epochs. The networks en training procedure is implemented in python using PyTorch.

6.2.4 Bayesian positioning neural network

We also apply the methodology of heteroscedastic uncertainty modelling with predictive variance (see sections 2.3.4 and 5.2.4) to the experimental setup. The same data and training procedure is adopted as explained in sections 6.2.1 and 6.2.3 with an additional output and adapted loss to predict the variance. We will evaluate whether the obtained uncertainty measures can be used to further improve positioning performance by filtering highly uncertain events.

6.2.5 Evaluation metrics

To evaluate the 2D spatial resolution, the 2D FWHM and bias metrics are used similar to the evaluation measures in section 4.2.5. The distance metrics are not calculated as the evaluation dataset also contains many events originating from background activity that are distributed across the entire detector. These events strongly influence the measured average distance making this metric not useful. For the same reason, the Gaussian used to calculate the FWHM values is only fit to the PSF points in the 2D histogram that exceed 25% of the peak value. The threshold of 25% was empirically determined and resulted in the most consistent fit. Otherwise, the long tails due to background events cause an inaccurate fit of the Gaussian (not fitting the peak of the PSF) and thus less accurate estimation of FWHM.

To evaluate DOI performance, the predicted relative number of events in each layer is compared to the expected theoretical percentages according to the Beer-Lambert law. Furthermore the overall DOI accuracy and F1 score is calculated together with a confusion matrix.

6.3   Results

6.3.1   Experimental setup

2D spatial resolution

The spatial resolution obtained with the network trained on 0.6 mm beam data is included in table 6.1. A quiver plot visualising FWHM and bias values across the entire detector is shown in figure 6.5. Higher FWHM and bias measures are observed near the detector edges. A median 2D FWHM of 1.10 mm and bias of 0.12 mm is achieved over the entire detector. Without the 10 mm border region, the median FWHM and bias improve to 1.01 and 0.08 mm respectively. On the intermediate grid points in the centre 9×9 mm^2 region, a median FWHM is measured of 0.97 mm.

Table 6.1: 2D positioning performance [mm] of the network trained on 0.6 mm beam data. Measures are calculated over different regions of the detector on the training grid (blue grid in figure 4.2) and on the intermediate grid (red grid in figure 4.2).

	Train Grid Entire Detector		Train Grid Centre 30×30 mm^2		Intermediate Grid Centre 9×9 mm^2	
	Mean	Median	Mean	Median	Mean	Median
FWHM	1.15	1.10	1.02	1.01	0.97	0.97
Bias	0.20	0.12	0.09	0.08	0.09	0.08

Training the network on 1 mm beam data and evaluation on 0.6 mm data results in a median FWHM of 1.10 mm across the entire detector and 1.02 mm in the detector centre (see table 6.2). The attained 2D median bias values are 0.12 mm and 0.08 mm for the two regions respectively. On the intermediate grid, the median FWHM is 0.98 mm.

DOI performance

In terms of DOI layer prediction, the network (trained and evaluated on 1 mm beam data with ROI filtering) achieves an overall accuracy of 74% and an F1 score of 75%. The confusion matrix is depicted in figure 6.6a.
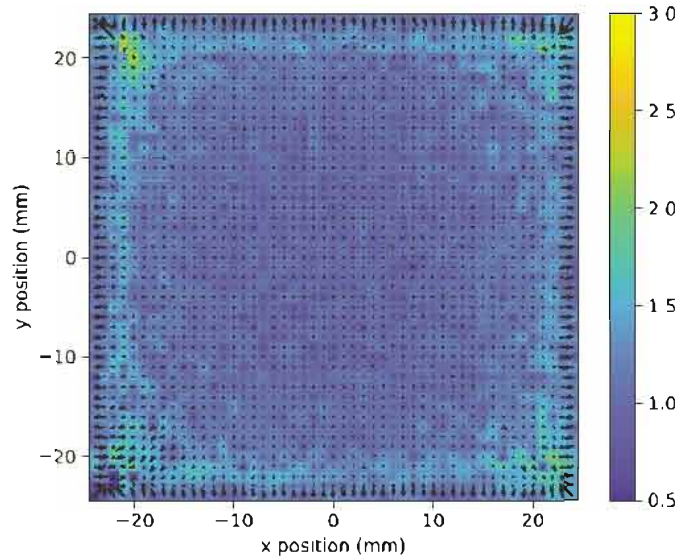
Figure 6.5: Quiver plot illustrating FWHM [mm] (colour scale) and bias
vectors (arrows) over the entire detector (training grid) obtained
with the network trained on 0.6 mm beam data.

Table 6.2: 2D positioning performance [mm] of the network trained on 1 mm
beam data. Measures are calculated over different regions of the
detector on the training grid (blue grid in figure 4.2) and on the
intermediate grid (red grid in figure 4.2).

	Train Grid Entire Detector		Train Grid Centre 30×30 mm^2		Intermediate Grid Centre 9×9 mm^2	
	Mean	Median	Mean	Median	Mean	Median
FWHM	1.15	1.10	1.02	1.02	0.97	0.98
Bias	0.19	0.12	0.09	0.08	0.08	0.07

Figure 6.6b shows the relative distribution of events across the six DOI
layers obtained with the neural network predictions and according to
the theoretical distribution expected from the Beer-Lambert attenuation
law. Most events are predicted in the correct layer with some confusion
mainly limited to the neighbouring layers.

Uniformity and bar phantom

The flood source uniformity plot is shown in figure 6.7. Bright hotspots
are observed in an 8 × 8 grid related to the 8 × 8 SiPM array. The four-
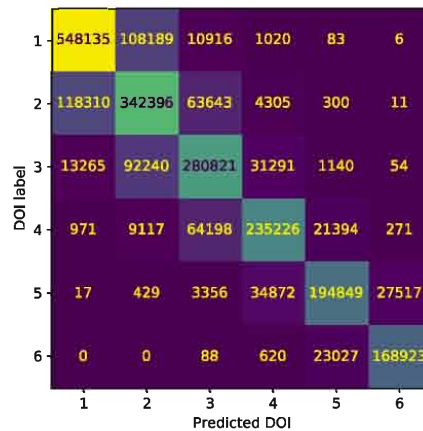
(a) Confusion matrix of DOI layer predictions.
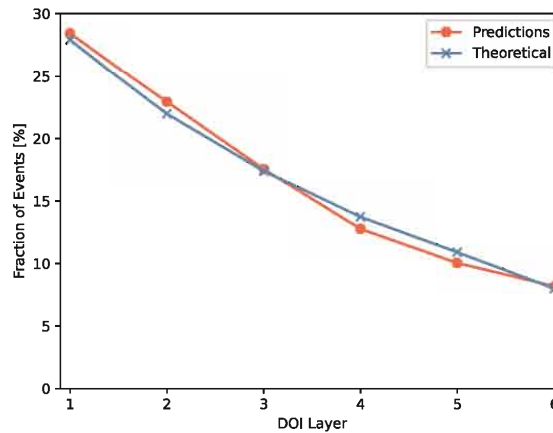
(b) Distribution of events positioned in each DOI layer.

Figure 6.6: DOI performance. The total number of included events is 2401000 (1000 events per calibration position) after ROI filtering.

quadrant bar phantom measurement is shown in figure 6.7. All four bar spacings can be distinguished with the smallest spacing being 0.6 mm (lower-left quadrant). The bars are more clear in the detector centre and are more difficult to resolve near the edges, especially in the corners.

MNN performance

For easy comparison, we include the results with mean nearest neighbour positioning as obtained from Stockhoff et al. [256]. With the MNN positioning algorithm calibrated and evaluated on 0.6 mm beam data, a median FWHM is achieved of 1.17 mm across the entire detector with a median 2D bias of 0.59 mm. Without the 10 mm border region, the
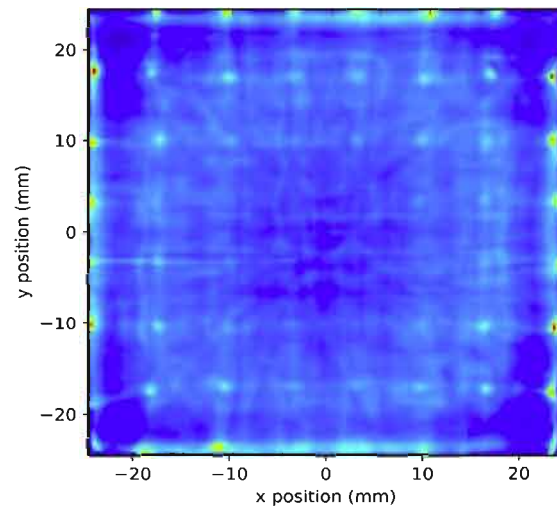
(a) 2D histogram of flood source predictions illustrating detector uniformity.
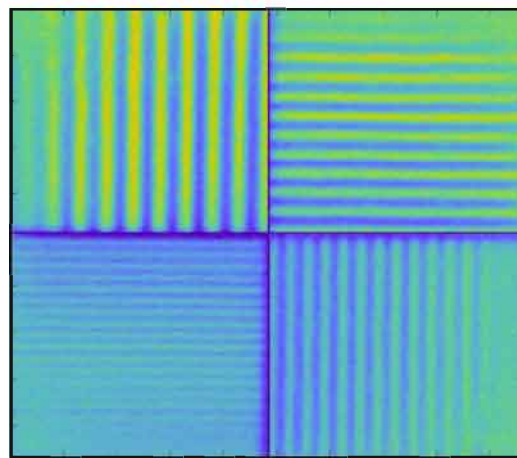
(b) Four-quadrant bar phantom measurement.

Figure 6.7: Detector uniformity and bar phantom measurement as a qualitative illustration of the achieved spatial resolution with the neural network.

resulting median FWHM and bias values are 1.14 mm and 0.26 mm. The quiver plot and bar phantom measurement is shown in figure 6.8.
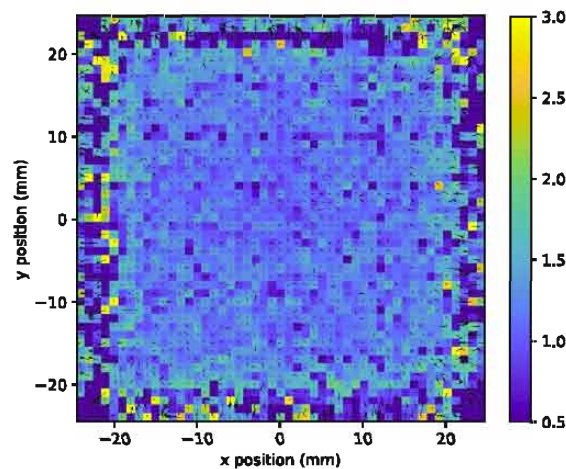
(a) Quiver plot illustrating FWHM [mm] (color scale) and bias vectors (arrows).
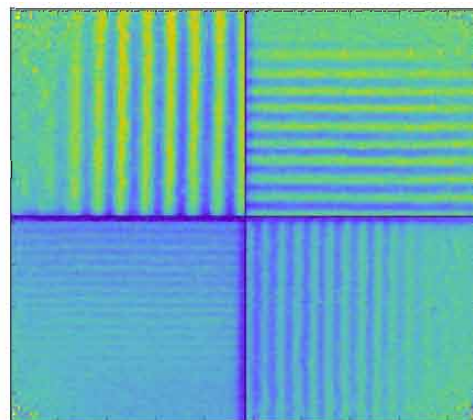
(b) Four-quadrant bar phantom measurement.

Figure 6.8: Quiver plot and bar phantom measurement illustrating 2D positioning performance of mean nearest neighbour positioning algorithm [256].

6.3.2   12 mm thick PET detector

The performance for the 12 mm thick PET detector is presented in table 6.3 with individual channel readout (64 input channels) and in table 6.4 with multiplexed readout (16 input channels). With the multiplexed readout, the resulting median FWHM is slightly higher: 1.03 mm versus 0.99 mm across the entire detector and 0.99 mm versus 0.96 mm in the detector centre. The FWHM bias measures are visualised in figure 6.9 for the two different readouts. Due to a problem with source

positioning during acquisition, no correct events are acquired for the first 7 positions at $y = 9.5$. These positions are not considered during evaluation.

Table 6.3: 2D positioning performance [mm] for the 12 mm thick PET detector with network trained on data with individual channel readout (64 channels). Measures are calculated on the intermediate grid (red grid in figure 6.4) over different regions of the detector .

	Intermediate Grid Entire Detector		Intermediate Grid Centre 22×22 mm²	
	Mean	Median	Mean	Median
FWHM	1.04	0.99	0.96	0.96
Bias	0.40	0.34	0.31	0.31

Table 6.4: 2D positioning performance [mm] for the 12 mm thick PET detector with network trained on data with multiplexed readout (16 channels). Measures are calculated on the intermediate grid (red grid in figure 6.4) over different regions of the detector .

	Intermediate Grid Entire Detector		Intermediate Grid Centre 22×22 mm²	
	Mean	Median	Mean	Median
FWHM	1.08	1.03	0.99	0.99
Bias	0.36	0.32	0.27	0.25

The DOI performance is depicted in figure 6.10 through a confusion matrix and the obtained distribution across the different DOI layers. Many events from the first layer are predicted into the second layer. Otherwise the distribution with the predictions and labels are similar with a majority of the events attributed to the same layer.

6.3.3 Bayesian positioning neural network

The 2D positioning performance of the network trained to predict both position and heteroscedastic uncertainty is presented in table 6.5. A median 2D FWHM is achieved of 1.09 mm across the entire detector
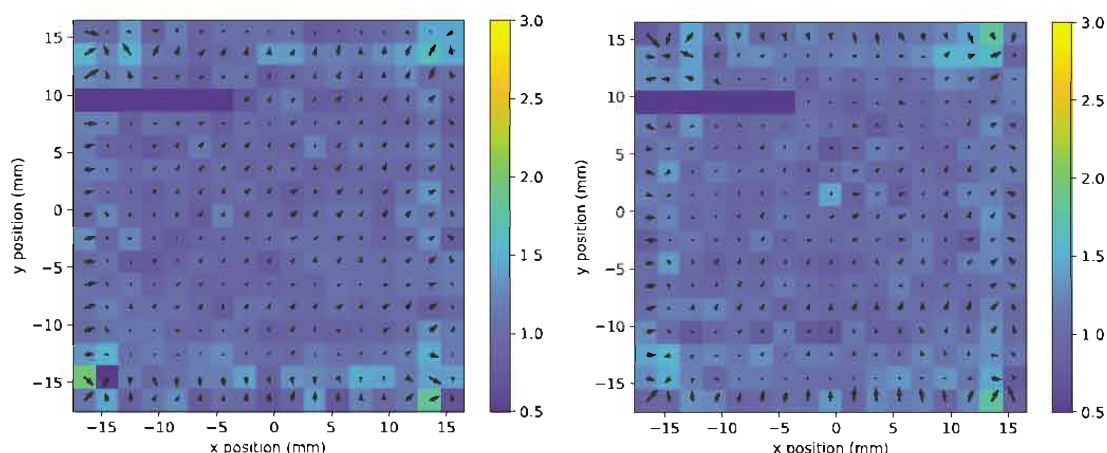
(a) Network trained on data with individ- (b) Network trained on data with multi-
ual channel readout.   plexed readout.

Figure 6.9: Quiver plots illustrating the obtained FWHM [mm] (color scale)
and bias vectors (arrows) for the 12 mm thick detector.

and 1.02 mm in the detector centre. The obtained median bias is 0.11
mm and 0.07 mm respectively. On the intermediate grid points in the
centre 9×9 mm^2 region, a median FWHM is measured of 0.97 mm.

Table 6.5: 2D positioning performance [mm] for bayesian positioning network.
Measures are calculated over different regions of the detector on
the training grid (blue grid in figure 4.2) and on the intermediate
grid (red grid in figure 4.2).

	Train Grid Entire Detector		Train Grid Centre 30×30 mm^2		Intermediate Grid Centre 9×9 mm^2	
	Mean	Median	Mean	Median	Mean	Median
FWHM	1.15	1.09	1.02	1.02	0.97	0.97
Bias	0.19	0.11	0.08	0.07	0.07	0.07

In chapter 5 it was already discussed that evaluating improvement
in positioning performance when filtering uncertain events is difficult
through FWHM measurement because reducing the long tails only has a
limited influence on the Gaussian fits to the PSFs. As we cannot calculate
a meaningful distance measure due to the presence of many background
events in the evaluation dataset, assessing whether the spatial resolution
improves is difficult. Plotting a histogram of all predictions without the
10% most uncertain events does reveal that mostly events in the corner
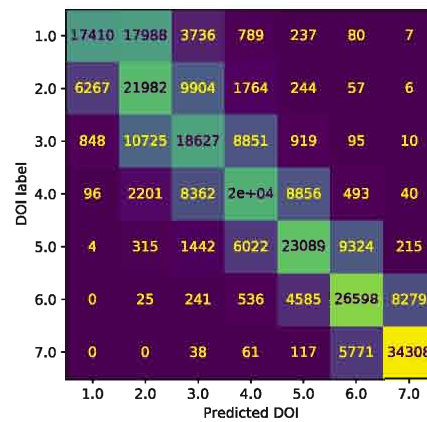
(a) Confusion matrix of DOI layer predictions.
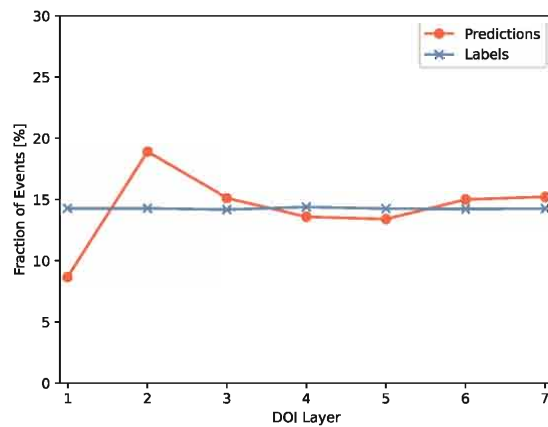
(b) Distribution of events positioned in each DOI layer.

**Figure 6.10:** DOI performance of 12 mm tick detector.

are filtered out (see figure 6.11).

6.4 Discussion

6.4.1 Experimental setup

The results in section 6.3.1 show that a high spatial resolution of around 1 mm FWHM is achieved in the detector centre region. Towards the detector border, the resolution and bias degrade which is typical for monolithic PET detectors. This is due to light truncation near the crystal edges resulting in a more difficult position estimation. Additionally, the measured PSFs near the borders show larger tails and are not well charac-
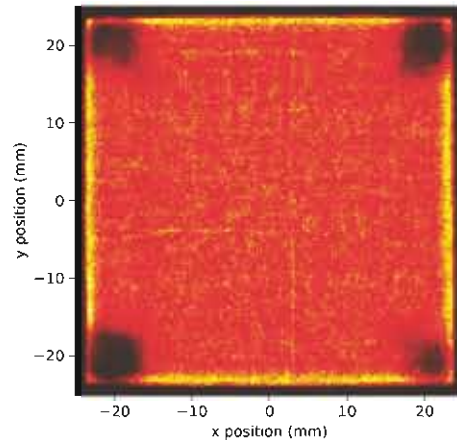
Figure 6.11: Histogram of all predictions without the 10% most uncertain events according to the bayesian positioning neural network with heteroscedastic uncertainty modelling.

terised by the Gaussian fit used to measure FWHM. The FWHM values near the detector edges should therefore be considered with caution. Evaluation on intermediate grid points, not used for training, shows no overfitting on the discrete 49 × 49 training grid.

Training the network on data acquired with a 1 mm collimated beam source instead of 0.6 mm and evaluating on the same 0.6 mm evaluation data shows no difference in performance. This corresponds with our observations on simulation data in section 5.3. The calibration beam diameter has little effect on neural network training and the achievable spatial resolution. Calibration with a broader beam allows data acquisition at a higher count rate and can therefore reduce the required calibration time.

No correction for the calibration beam width was included in the evaluation measures. In section 5.3, we have seen that a larger beam width for the evaluation data does result in a degradation in the measured FWHM values. The source beam used to acquire the evaluation data has a diameter of 0.6 mm at its smallest and spreads to more than 1 mm deeper into the crystal (see figure 5.13). We can therefore expect that the true intrinsic spatial resolution is even better than the FWHM measures indicate.

The DOI encoding capability of monolithic PET detector is an important asset compared to other detector designs. From the measured light

spread, the depth of interaction can be inferred. Using simulation data, neural networks can be trained to directly predict the z-coordinate of the first interaction position as demonstrated in chapter 4. For experimental data, however, it is much more difficult to obtain accurate depth of interaction information. In this work, DOI labels were used that were obtained based on the standard deviation of the measured light distribution and the Beer-Lambert attenuation law. Events were grouped into six DOI layers. These labels are not entirely accurate, especially when taking Compton scatter into account. As demonstrated in section 5.2, Compton scatter has a strong influence on the measured light distribution and consequently the light spread and DOI estimation. Events that scatter deeper into the crystal produce a sharper light peak and are potentially categorised into a deeper DOI layer than the true layer of first interaction. The problem of Compton scatter affecting DOI label accuracy persists when using other methods to obtain depth dependent data such as inclined or side irradiation. Quantifying the true DOI performance is therefore difficult which should be considered when interpreting the DOI results in section 6.3.1. The distribution of events across the DOI layers obtained with the neural network fits the theoretical trend as expected from the Beer-Lambert law. Most events are predicted into the same layer as the derived DOI labels with some events positioned into neighbouring layers. We can conclude that the neural network is able to approximately infer the depth of interaction which can be used to reduce parallax errors and improve time-of-flight estimation. Techniques to determine more accurate ground truth DOI information for experimental data could further improve the DOI performance.

The uniformity plot in figure 6.7a clearly shows the artefacts induced by the 8 × 8 SiPM array. For gamma rays that interact deeper into the crystal (close to the SiPMs), most of the light is captured by a single SiPM and as a consequence little information is included on the exact interaction position within the SiPM pixel. Most of these events are therefore predicted in the centre of that pixel.
The bar phantom measurement allows to qualitatively assess the overall detector resolution. Bars down to 0.6 mm can be resolved. The bars are less clear near the detector borders due to the edge effect as discussed previously. An additional reason is the source position and angle of the incoming gamma rays. The gamma rays originate from a point source and arrive in the border region with a certain angle resulting in increased absorption of these events by the tungsten bar phantom. Gamma rays

arriving in the centre below the point source are perpendicular and can pass through the slits.

Comparison with the results obtained with mean nearest neighbour positioning reveals superior positioning performance with neural networks. In the detector centre, an improvement of 11.4% is achieved with the neural network compared to MNN positioning. This improvement can also be visually perceived when comparing the quiver plots and bar phantom measurements. The quiver plot obtained with the neural network shows a more uniform positioning performance, especially towards the detector edges. The bars in figure 6.7b appear more linear and less noisy compared to figure 6.8b. Furthermore, the edge effect is more noticeable with the MNN algorithm. A plot of the DOI distribution in Stockhoff et al. [256], similar to figure 6.6b, shows that more events are predicted in the sixth (deepest) layer with the MNN algorithm that expected from the theoretical distribution. A similar result was observed in simulations in Stockhoff et al. [80] and is related to Compton scatter. This shows that the neural network is possibly able to more reliably position these events and learn to take Compton scatter into account, even when trained on imprecise labels.

As already discussed in chapter 5, comparison of spatial resolution between different studies is difficult and strongly influenced by crystal size and thickness, SiPM readout, calibration beam width and evaluation procedure. Many different detector designs have been evaluated with different positioning algorithms (see section 3.3.1). Reported FWHM measures range between 1.2 mm for a $50 \times 50 \times 10$ mm^3 LYSO crystal [78] and 1.7 mm with a $32 \times 32 \times 22$ mm^3 crystal [60]. This illustrates that, in this work, state-of-the-art spatial resolution is achieved with neural network positioning. A further improvement in resolution can potentially be achieved by reducing SiPM pixel size as was shown in Stockhoff et al. [80] on simulation data.

6.4.2   12 mm thick PET detector

For the second, smaller PET detector design, we also achieve a good spatial resolution of 0.96 mm FWHM in the detector centre region. The measures are calculated on intermediate positions, not used during training, and no overfitting on the training positions is observed. When looking at figure 6.9a, the edge effect appears less present. This can be

explained by the intermediate positions that are slightly further from the edges (see figure 6.4), especially at the right and top edges in figure 6.9a. Comparing performance between individual and multiplexed readout shows only a minor degradation of around 0.4 mm in achieved FWHM. This is in line with previous observations on simulation data in Stockhoff et al. [80]. Little difference was reported between individual and multiplexed SiPM readout. A combined readout can therefore be used to greatly reduce the cost of readout electronics with only a minor decrease in positioning accuracy.

The DOI performance shown in figure 6.10 indicates that the network is able to infer the assigned DOI label. A majority of the events are classified in the correct layer. Only for the first layer, many events are predicted in the second layer. For this detector, the events are grouped into seven DOI layers according to signal standard deviation, with an equal amount of events in each layer. The different layers therefore correspond with different thicknesses as more events interact at the top of the crystal. Consequently, events attributed to the deeper layers have a larger variety in light spread than events in the top layers and the relation between DOI and the assigned label is less interpretable. Assigning events to DOI layers with equal thickness according to the Beer-Lambert law, as for the first detector, allows a more direct relation between signal standard deviation and DOI label. This potentially results in a better DOI estimation of the network.

6.4.3 Bayesian positioning neural network

Results of the neural network with uncertainty modelling demonstrate that the same spatial resolution can be achieved as the original positioning network without variance prediction. When using the obtained variance measures to filter uncertain events, we observe that mostly events in the corners are discarded. This behaviour is not desired as uniform positioning across the detector is necessary. It appears that the contribution of event location (edge effect) to the predicted uncertainty is now higher than the contribution of Compton scatter which was not observed on the simulation data (see section 5.2.4).

One potential reason is that the edge effect can be stronger in the experimental setup than with simulations because of imperfections in the surface finish on the sides. A second reason could be that the ROI filtering applied to the training dataset to remove background events also

removes many of the (far) scattered events. When the training set does
not contain Compton scattered events, the network is not able to learn
uncertainty associated with Compton scatter and only the contribution
related to the edge effect remains. This problem could be diminished
through the use of a larger ROI, but this also introduces more background
events to the training data. The use of other crystals types without
intrinsic radioactivity such as BGO could also be a solution as no ROI
filtering would be required and all acquired events can be included in the
training dataset.

6.5   Conclusion

In this chapter, we have validated our methodology of training neural
networks for positioning of gamma interactions in monolithic PET de-
tectors on experimental data. Similar to the results on simulation data,
high spatial resolutions can be achieved with neural networks, superior
to the mean nearest neighbour positioning algorithm. Neural networks
are trained on individual events and directly learn to infer the interaction
position from the measured light distribution. This leads to an improved
positioning accuracy of Compton scattered events and less degradation
near the detector edges. Moreover, neural networks produce continuous
coordinate outputs, not restricted to a discrete calibration grid. Lastly,
positioning events with the network is fast and parallelisable, especially
when using powerful hardware like GPUs.

PART II:

AI IN IMAGE ANALYSIS: PRIMARY BRAIN TUMOUR DIAGNOSIS

7 | Computer-aided diagnosis of primary brain tumours

The second part of this dissertation focuses on the application of AI to medical image analysis, specifically on computer-aided diagnosis of primary brain tumours (PBTs). Based on medical images, AI algorithms (primarily deep learning networks) will be developed for segmentation of primary brain tumours and prediction of important characteristics regarding prognosis and therapy planning. To understand this part of the thesis, some background information is provided in this chapter. Starting with a brief overview of neuroanatomy, the different types and classification of primary brain tumours will be described followed by epidemiology, diagnosis and treatment of PBTs. Afterwards, a literature review is included on work related to PBT segmentation and diagnosis with AI.

7.1 Primary brain tumours

In contrast to secondary brain tumours or metastases which originate from other parts of the body and have spread to the brain, PBTs arise in the brain. Different types of PBTs are historically named after the cells or structures in the brain from which they originate. We therefore provide a brief overview on neuroanatomy based on the human brain book by Carter et al. [257] which contains a complete overview of the brain.

155

7.1.1 Neuroanatomy

A schematic overview of the brain anatomy is illustrated in figure 7.1. Together with the spinal cord, the brain forms the human central nervous system (CNS). The CNS is responsible for receiving information from all over the body, processing this information and in turn send signals to control the activity in all parts in the body. For protection, the brain is surrounded by cerebrospinal fluid (CSF), the meninges, skull and scalp. Cerebrospinal fluid is produced in the central cavities in the brain, called ventricles.

The brain consists of three major parts: the brain stem, cerebellum and cerebrum. The cerebrum is divided into a left and right hemisphere and the outer part is called the cerebral cortex. In order to increase the surface area that fits into the skull, the cerebral cortex is folded, forming patters of bulges (gyri), shallow grooves (sulci) and deep grooves (fissures).
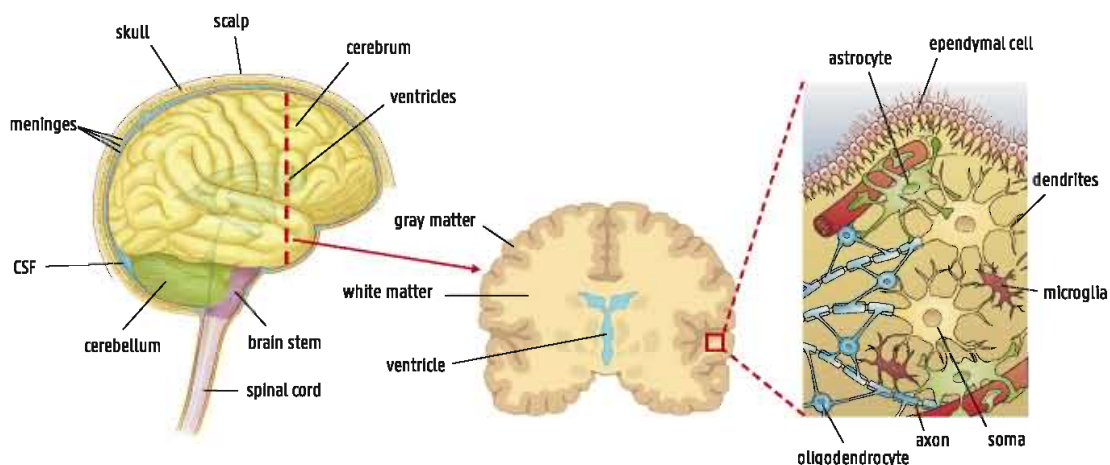
Figure 7.1: Overview of the anatomy of the central nervous system. Adapted from [258–260].

The basic components of the brain are the neurons or nerve cells. They are composed of a cell body or soma, dendrites and an axon. The somas process signals arriving through the dendrites and transit new signals along their axons to the dendrites of other neurons. The cell bodies mostly reside in the cerebral cortex, giving this structure the typical grey color which is why it is often referred to as grey matter. Axons, on the other hand, surrounded by the myelin sheath are mostly located in the inner part of the brain, called white matter.

Next to neurons, the brain also consists of supporting cells or glia.

The cells that produce the myelin sheath around the axons are called oligodencrocytes. Astrocytes maintain the chemical environment in the brain and are part of the blood-brain barrier. Furthermore, ependymal cells are responsible for secreting and circulating CSF. Finally, microglia destroy invading microbes and clear away cell debris.

7.1.2 The WHO classification

The World Health Organisation (WHO) historically classifies primary brain tumours based on histological findings and the microscopic similarity with their cells of origin [3]. The characterisation of histological similarities is primarily dependent on light microscopic features in hematoxylin and eosin-stained sections and immunohistochemical expression of proteins. Many different types of PBTs are defined, where glioma and meningioma are the most common forms. Glioma arise from glial cells and are further subdivided according to the glial cell types they share histological features with. The most common glioma are astrocytoma, oligodendroglioma and ependymoma. Glioma can be very heterogeneous and typically invade the brain tissue. They are the most frequently occurring primary brain tumours and show a large heterogeneity in treatment response and prognosis. Meningiomas originate form the meninges. In contrast to glioma, they are slow growing, more homogeneous and rarely invade the brain.

Next to cell type, primary brain tumours are also divided into different WHO grades (I-IV) in order of malignancy based on histopathological and clinical criteria [261]. Histological features used to determine malignancy are anaplasia, pleomorphism, mitotic activity, proliferation, necrosis etc. WHO grade I tumour show low proliferation and are possibly cured with surgery alone depending on location. Tumours with grade II are more infiltrative and tend to recur and progress into higher grades of malignancy. Grade III is assigned to neoplasms with histological evidence of malignancy such as nuclear atypia and mitotic activity. WHO grade IV is reserved for highly malignant tumours with high mitotic activity, necrosis and rapid progression.

In the most recent classification of PBTs from 2016, the WHO has put increased emphasis on the integration of molecular markers [3]. Classification based on histopathological analysis alone suffers from subjectivity and inter-observer variability [262, 263]. Moreover, tumours that were

classified in one group based on histology often showed highly varying prognosis and several studies report that gene expression profiles are a better predictor of survival [264, 265]. The integration of genotypic parameters for CNS tumour classification intends to add objectivity and yield more narrowly defined diagnostic entities. A complete discussion of all defined tumour types and nomenclature is out of the scope of this study. The interested reader is referred to "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary" by Louis et al. [3]. We limit ourself to the important example of the classification of diffuse glioma based on histological and genetic markers which is shown in figure 7.2.
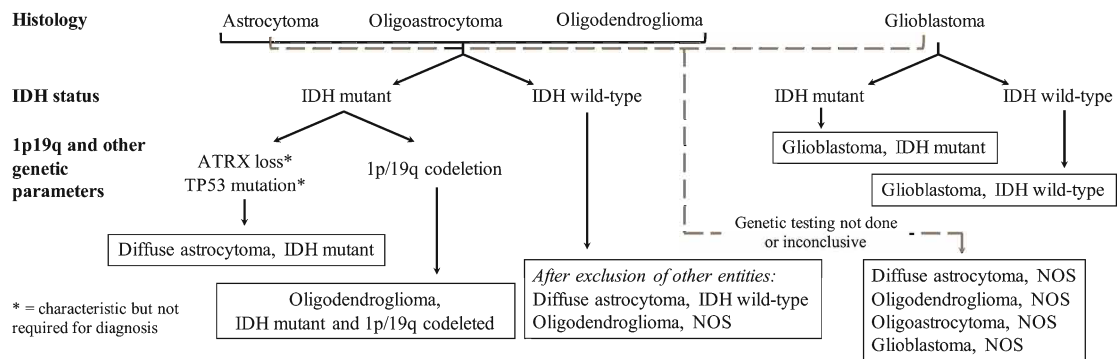
Figure 7.2: Classification of diffuse gliomas based on histological and genetic markers. Not otherwise specified (NOS) designates a group of lesions that cannot be classified into the more narrowly defined groups or for which insufficient information is available [3]. Adapted with permission from Copyright Clearance Center: Springer Nature, Louis et al. [3] © 2016.

In the classification scheme shown in figure 7.2, three markers play a central role: histological grade, isocitrate dehydrogenase (IDH) 1 and/or 2 mutation and co-deletion of chromosome arms 1p and 19q. Justification for the inclusion of these genetic markers is visualised in figure 7.3. Different tumour types, categorised according to IDH mutation and 1p/19q co-deletion, show distinct overall survival patterns.

In terms of WHO grade, one differentiates between Glioblastoma multiforme (GBM), the most aggressive type (WHO grade IV) of astrocytoma, and lower-grade glioma (LGG) including WHO grade II and III astrocytoma and oligodendroglioma. Glioblastoma is associated with very poor prognosis and a 5-year survival rate of only 5.6%. Lower-grade glioma, on the other hand, have more favourable survival rates up
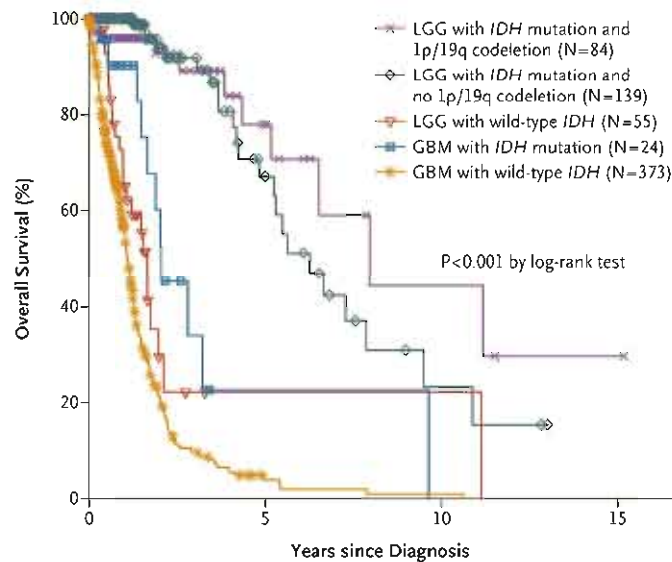
Figure 7.3: Kaplan-Meier curves showing overall survival for diffuse glioma classified according to IDH mutation and 1p/19q co-deletion status. Reproduced with permission from [266], Copyright Massachusetts Medical Society.

to 81.6% and 57.6% for WHO grade II and III respectively [267].

The first important genetic marker is IDH status playing a key role in the Krebs cycle and cellular homoeostasis [268]. IDH mutation occurs in more than 80% of lower-grade glioma cases and approximately 10% of glioblastoma cases, corresponding closely to so-called secondary glioblastoma [269, 270]. Secondary glioblastoma evolve from lower-grade astrocytoma whereas primary GBM are immediately formed from healthy tissue. Gliomas with IDH mutation are less aggressive and demonstrate better response to temozolomide chemotherapy than IDH wildtype gliomas. For example, glioblastoma patients with IDH mutation show a longer overall survival (OS) compared to patients with IDH wildtype glioblastoma (see figure 7.3) [270]. Moreover, reported OS of IDH wildtype LGG is only slightly longer than IDH wildtype glioblastoma [266]. Hence IDH mutation is associated with a significantly better prognosis and appears to be a more important predictor than WHO grade as Reuss et al. [271] reported little difference in survival between IDH mutant WHO grade II and III astrocytoma. IDH mutation can be detected through negative gene sequencing for the IDH1 codon 132 and IDH2 codon 172 gene mutations. Immunohistochemistry (IHC) can also be used to determine IDH1 mutation. However, a negative IDH status using IHC does not necessarily mean an IDH wildtype tumour and if no

sequencing is available the resulting diagnosis suggested by the WHO is astrocytoma, not otherwise specified (NOS) [3, 272].

The second important genetic marker is combined loss of chromosome arms 1p and 19q (1p/19q co-deletion). According to the 2016 WHO classification scheme, diagnosis of oligodendroglioma requires demonstration of both IDH mutation and 1p/19q co-deletion. Similarly to IDH mutation, 1p/19q co-deletion is linked to more favourable outcomes and oligodendrogliomas respond well to combined procarbazine, lomustine and vincristine chemotherapy [273].

7.1.3   Epidemiology

Primary brain tumours are a relatively rare type of cancer. A systematic review and meta analysis by Robles et al. [274] reports a worldwide incidence rate of primary brain tumours of 10.82 (95% CI: 8.63-13.56) people per 100,000 per year. The incidence varies significantly for different regions, with the highest rates reported in northern Europe, the United States, Canada, and Australia [275]. PBTs are, however, a significant cause of cancer morbidity and mortality [276], especially in children and young adults where they are the leading cause of cancer deaths. The most common types of CNS tumours in children are pilocytic astrocytoma (17%), malignant gliomas (17%) and embryonal tumours (15%). In adults, the most occurring CNS tumours are meningiomas (36%), pituitary tumours (15%), and glioblastoma (15%).

Many potential risk factors for primary brain tumours have been studied, but only few are well established [276]. The most established risk factor is ionising radiation linked to inducing primarily meningioma and glioma. Other identified factors are genetic factors and allergies or immune-related conditions. Allergies are reported to be inversely correlated with risk of developing CNS tumours [275, 276]. No evidence has been found of significant association between exposure to non-ionising radiation from mobile phones and brain tumour incidence.

7.1.4   Symptoms and diagnosis

Depending on the location, growth and size of the tumour, different symptoms can be present [277]. More general symptoms, not specific to an anatomical location, are epileptic seizures, headaches and symptoms due to increased intracranial pressure such as nausea, vomiting,

drowsiness and blurred vision. Tumours in certain functional areas of the brain can cause specific neurological deficits. Frontal lobe tumours, for example, might result in dysphasia (language disorders). Visual abnormalities can be caused by tumours that are involved in tracts that are connected to the primary visual cortex. Tumours located in the prefrontal or temporal lobe or in the corpus callosum can result in personality changes and mood disorders.

Tumours that cause obvious neurological deficits are often detected sooner using medical imaging. Slow growing tumours such as meningioma, on the other hand, often show no symptoms and are only discovered after years or by chance after a brain scan performed for other purposes (e.g. an accident or stroke).

To detect brain tumours, brain MRI is the gold standard, including T1-weighted sequences before and after application of a gadolinium based contrast agent, a T2 sequence and a T2-weighted fluid-attenuated inversion recovery (FLAIR) sequence [278]. An example of these four MRI sequences for a patient with an IDH wildtype GBM is included in figure 7.4. On a T1 weighted scan (figure 7.4a), fluids such as CSF have low intensities and white matter has a higher intensity than grey matter. The tumour appears hypo-intense. The T1 contrast-enhanced (T1ce) scan shown in figure 7.4b is acquired using the same scanning parameters, but after administration of a gadolinium based contrast agent. This contrast agent results in a bright signal in blood vessels and regions where the blood-brain barrier is disrupted. The ring shaped enhancing tissue around a necrotic core of the tumour seen in figure 7.4b is typical for glioblastoma. A T2 weighted sequence figure 7.4c highlights regions containing a lot of water (e.g. ventricles containing CSF). Now white matter appears hypo-intense compared to grey matter. In the tumour, the necrotic core and surrounding oedema, caused by fluid leakage and invasion of the tumour into healthy tissue, appear bright as they contain a lot of water. The FLAIR sequence (see figure 7.4d) is T2 weighted as well and has therefore similar characteristics. However, the signal of CSF is now attenuated. This improves the contrast between healthy and pathological tissues.

Diffusion and perfusion MRI and PET can aid to delineate metabolic hotspots to guide tissue sampling for biopsy or to assess tumour progression and treatment response.

Tumour type and molecular markers are determined based on tissue analysis (histological and genetic) extracted through biopsy or resection

(see next section).

(a) T1   (b) T1ce

(c) T2   (d) FLAIR
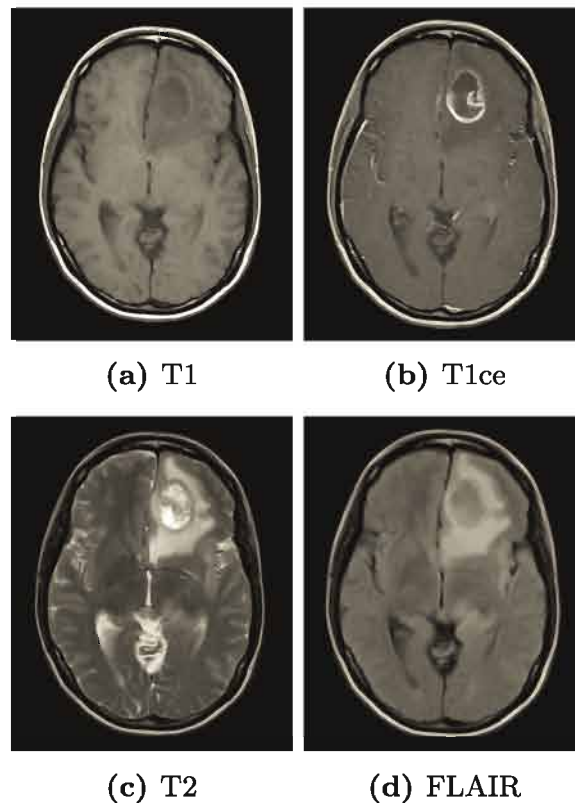
Figure 7.4: Example brain MRI sequences of a patient with an IDH wildtype Glioblastoma.

7.1.5   Treatment

Several therapy options exist to treat primary brain tumours. Optimal therapy planning and prognosis depends on tumour type, age, Karnofsky performance score and individual risks and benefits [278]. Different treatment options are briefly summarised below. For a more complete overview, we refer to the European Association of Neuro-Oncology (EANO) guidelines on the diagnosis and treatment of diffuse gliomas of adulthood [278].

Watch-and-wait

In some cases it might be beneficial to postpone (invasive) treatment and opt for a watch-and-wait approach. The patient is closely monitored with regular brain scans to evaluate tumour progression. Invasive procedures

involve risks and when there are indications that the tumour is benign (WHO grade I) or when there are few symptoms, one could choose to wait with performing surgery until signs of further growth. A study by Wijnenga et al. [279] reports no difference in survival between early resection and a wait-and-scan approach in low-grade glioma.

Surgery

The goal of surgery is to remove as much tumour tissue as possible without damaging healthy tissue and compromising neurological function (maximum safe resection). Advancements in microsurgical techniques, surgical navigation systems, medical imaging and awake surgery have contributed to reduce residual tumour volumes and risks of new neuro- logical deficits. The extent of resection and remaining tumour volumes are prognostic factors. However, whether and why extent of resection matters remains debated. Tumour that are better resectable often have a different, less malignant biology which complicates the relation between survival and extent of resection. For this reason, preventing neurological deficits has a higher priority than extent of resection.

Surgery, if possible, is the primary form of therapy for most PBT patients [273, 278]. It often results in an immediate relief of symptoms and allows histological and molecular analysis of the tumour tissue. In case resection is not possible due to location of the tumour or clinical condition of the patient, a stereotactic biopsy can be considered. A needle is used to extract small fractions of the tumour which can then be used for further diagnosis.

Radiotherapy

In radiotherapy, ionising radiation is used to kill cancer cells. The goal is to improve local control and increase survival without inducing toxicity to healthy regions [278]. The radiation can be applied using an external beam or through internal radiation sources (brachytherapy).
Optimising the location and dose of the radiation delivered to the tumour while avoiding healthy tissue requires careful planning. Especially sensi- tive structures such as the eyes, brain stem and optic nerves should be delineated and protected. The dose, timing and schedule of radiotherapy is planned based on prognostic factors and extent of resection if applied

after surgery. Often low doses of 1.8-2 Gy are administered daily until a certain total dose of 50-60 Gy is delivered. A single, high radiation dose can be administered as well which is called stereotactic neurosurgery. Brachytherapy, where a capsule containing a radiation source is placed near the tumour, can be an alternative to external beam radiation in children and in adults with deeply localised tumours [273].

Pharmacotherapy

Pharmacotherapy includes the administration of drugs to relieve symptoms and chemotherapeutics aiming to destroy cancer cells. In the first category, corticosteroids can be used to reduce oedema and anti-epileptic drugs to limit seizures.

The most commonly used drug in glioma treatment is temozolomide, a DNA alkylating agent that penetrates the blood-brain barrier and has a favourable safety profile [278]. EANO mainly recommends temozolomide for high-grade (WHO grade III and IV) astrocytoma and IDH wildtype glioma with O^6-methylguanine-DNA methyltransferase (MGMT) promotor methylation.
For IDH mutant and 1p/19q co-deleted glioma (oligodendroglioma) and astrocytoma WHO grade II, the use of alkylating agents from the nitrosourea class is recommended by EANO. More specifically a combination of lomustine, procarbazine and vincristine referred to as PCV.

7.2   Non-invasive computer-aided diagnosis

7.2.1   Importance of non-invasive diagnosis

From previous section we can conclude that determination of WHO grade (glioblastoma versus lower-grade glioma), IDH mutation and 1p/19q co-deletion status is key for optimal prognosis and therapy planning. Currently, genetic information of gliomas is derived from the analysis of tumour tissue obtained through biopsy or resection. However, biopsies involve risks and are subject to sampling error which can lead to misdiagnosis [280, 281]. Moreover, biopsies are related to reduced overall survival compared to a wait-and-scan approach followed by resection in low-grade glioma [279]. Tumour resection is standard of care for most glioma

types but is not always possible depending on tumour location and accessibility, the patient's clinical condition or when the patient refuses a surgical procedure. Therefore, non-invasive assessment of clinically relevant markers can aid in characterising glioma and guide therapy and surgery planning, especially when extraction of tumour tissue is not possible or genetic testing not available.

Correlations between MR phenotypes and glioma subtypes have been widely investigated. For example, presence of contrast enhancement and necrosis on T1 contrast enhanced (T1ce) MRI is associated with high-grade glioma [4]. IDH mutant glioma have been reported to demonstrate minimal enhancement, sharp tumour margins and homogeneous signal intensity [5, 6]. This contrasts with IDH wildtype glioma that is correlated with thick, irregular enhancement with necrosis and infiltrative oedema. Furthermore, increased enhancement, poorly circumscribed borders and heterogeneous signal intensity are characteristic MRI features related to 1p/19q co-deletion [7, 8]. However, visual interpretation and prediction of tumour properties remains very challenging and inaccurate. For instance, 40-45% of non-enhancing lesions are subsequently found to be highly malignant [282]. Conversely, 16% of WHO grade II glioma show contrast enhancement and this percentage is expected to be even higher for low-grade oligodendroglioma [283, 284].

To improve speed and accuracy of non-invasive tumour characterisation, there is an increasing interest to use machine learning techniques for medical image analysis. Below we provide an overview of existing state-of-the-art approaches for computer-aided brain tumour segmentation and diagnosis using AI.

7.2.2   Computer-aided segmentation

Brain tumour segmentation is not only an important pre-processing step to help further diagnosis, especially in radiomics (see section 3.4.1), it is also necessary for surgery planning, volume estimation and assessing tumour progression and treatment response. In clinical practice, brain tumour segmentation is still often done manually. An experienced radiologist delineates the different tumour tissues on multiple slices of a 3D MRI. This process is labour and time intensive and prone to inter- and intra-observer variability. Menze et al. [285] report a significant disagreement between delineations of different readers with Dice overlap

(a) T1   (b) T1ce   (c) T2

(d) FLAIR   (e) Segmentation
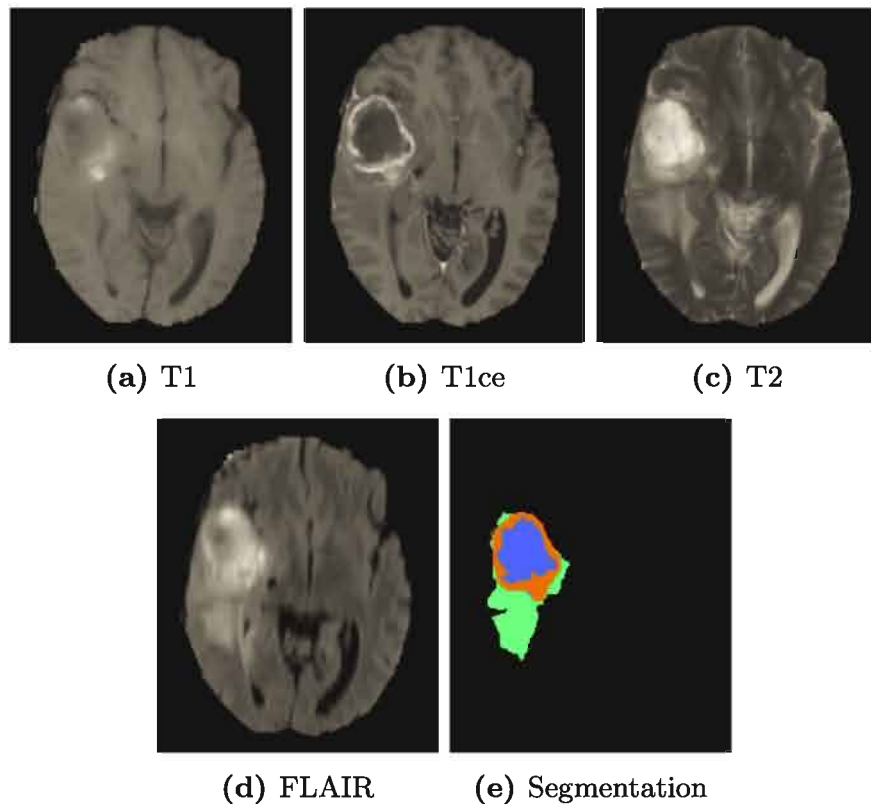
Figure 7.5: Example brain MRI sequences of BraTS dataset with manual annotations delineating different tumour tissues: necrosis (blue), enhancing (orange) and peritumoural oedema (green).

scores ranging between 74% and 85%.

For these reasons, there is a growing interest in automatic glioma segmentation. Automated delineation of different tumour tissues in multimodal MRI is challenging due to the large variety in imaging characteristics and tumour appearance, size, shape and location. A complete review of all studies in automated brain tumour segmentation is infeasible and out of the scope of this thesis. We therefore highlight several key approaches.

Encouraged by the annual Brain tumour Segmentation (BraTS) Challenges [285], a lot of research is performed on automatic glioma segmentation. BraTS provides a large heterogeneous dataset containing pretherapy MRI (T1, T1ce, T2 and FLAIR sequences) with corresponding manual annotations of three tumour tissue types: necrotic core, enhancing tissue and peritumoural oedema.

Early approaches often used thresholding or abnormality detection techniques to (semi-)automatically segment brain tumours [286–288]. For example, through the use of image registration, a pathological scan

can be aligned with a healthy atlas and abnormal tissues can be detected based on deviations in tissue appearance. Advantages of these techniques are that no training dataset with manual annotations is required and that they generalise well to scans acquired with various imaging characteristics. The accuracy of the segmentations is however often limited and deteriorates for very small lesions or large lesions that cause significant deformations that result in incorrect registration.

More successful results were achieved using discriminative, radiomics based approaches [285, 288–292]. Voxelwise classifiers such as random forests predict the tissue type of every voxel based on extracted intensity, texture, morphological and context features. Now annotated training sets are required which should be large enough to train the classifiers and, when applied, the scans need to match the characteristics of the training set to obtain accurate segmentation.

In recent years deep learning techniques, and in particular CNNs, have surpassed performance of more traditional radiomics methods [293]. Pereira et al. [294] proposed a top performing approach in the 2015 BraTS challenge using two 2D CNNs, one for low grade glioma and another for high-grade glioma. The CNNs were patch-based classification networks where 2D patches of 33 × 33 pixels are extracted from all four sequences and each 2D slice and classified into one of the tissue classes. Average Dice scores were achieved of 78%, 65% and 73% for the whole tumour (WT), tumour core (TC, necrosis+enhancing tissue) and enhancing regions (ET) respectively.
A multi-scale 3D CNN, called DeepMedic, was presented by Kamnitsas et al. [295]. DeepMedic consists of two parallel pathways incorporating local and contextual information and is fully convolutional. This allows to classify multiple voxels simultaneously. Additionally, post-processing was applied to the segmentation maps using a 3D conditional random field model. DeepMedic obtained state-of-the-art segmentation performance on the BraTS 2015 challenge data with reported Dice scores of 85% (WT), 67% (TC) and 63% (ET).
In the most recent BraTS challenges, (modified) U-Nets are the top performing architectures. A 3D U-Net with residual blocks in the encoder pathway was proposed by Isensee et al. [296]. The network was trained on patches of 128 × 128 × 128 voxels with extensive data augmentation and multi-class Dice loss was used as optimisation metric. On the 2017 BraTS test data, the attained mean Dice scores were 86% (WT), 78% (TC) and 65% (ET).

The winning approach in the BraTS 2017 challenge was proposed by Kamnitsas et al. [297] consisting of an ensemble of multiple CNN architectures including DeepMedic and U-Net. They reached average Dice scores of 89% (WT), 79% (TC) and 73% (ET).

Myronenko [298] achieved first place at the 2018 BraTS challenge with dice scores of 88% , 82% and 77% for whole tumour, tumour core and enhancing tumour volumes respectively. They used an ensemble of 10 encoder-decoder networks. Next to the segmentation decoder, an additional variational autoencoder branch was added to jointly reconstruct the input image during training to regularise the shared encoder.

The first place of the 2019 BraTS challenge was won by Jiang et al. [299] with Dice scores of 89% (WT), 84% (TC) and 83% (ET). Their approach consisted of a two-stage cascaded U-Net. The first stage produced a course prediction. This preliminary segmentation map is concatenated with the input MRI and fed to the second stage which refines the prediction. Two decoders are used in the second stage U-Net, one with deconvolution layers for upsampling and one with trilinear interpolation upsampling. The second decoder branch is only used during training for regularisation purposes.

Isensee et al. [300] applied the nnU-Net framework (see section 3.4.2) to the BraTS 2020 challenge and took the first place with Dice scores of 89%, 85% and 82% for whole tumour, tumour core and enhancing tumour respectively. The baseline configuration was improved through more data augmentation, region-based learning, post-processing, ensemble of 25 models and model selection based on the ranking scheme used by BraTS.

We can conclude that state-of-the-art performance is obtained in automatic brain tumour segmentation using CNNs, and more specifically U-Nets. The above methods require all four MRI sequences to be available as input. In clinical practice, however, it is common to have missing modalities. Due to time constraints, the acquisition of a T2 scan for example can be omitted. In chapter 9, we will design an automatic segmentation network based on the state-of-the-art U-Net architecture with increased robustness to missing modalities.

7.2.3   Computer-aided diagnosis

AI techniques are applied for a large variety of tasks in neuro-oncology including prediction of grade, molecular markers, survival, therapy re-

sponse, differentiation between pseudo-progression and tumour recurrence etc. [301, 302]. In this dissertation, we focus on non-invasive diagnosis of diffuse glioma according to the WHO 2016 classification scheme (see figure 7.2) based on routinely acquired MRI. A selection of studies on the prediction of grade, IDH mutation and 1p/19q co-deletion status of glioma is included in table 7.1.

A system for grade identification (low- versus high-grade) of astrocytoma from T2-weighted images was designed in the work by Subashini et al. [303]. Tumours were isolated with fuzzy c-means segmentation from which shape, intensity and texture features were calculated. A learning vector quantisation classifier trained on 164 images and evaluated on 36 images achieved an accuracy of 91%.

Hsieh et al. [304] proposed a computer-aided grading system using local and global MRI features. The most representative 2D slices from T1ce MRI were manually selected followed by manual delineation of the tumour contour. Histogram and texture features were fed to a logistic regression classifier reaching an AUC score of 0.89. In a follow up study [305], the same images were graded by three expert radiologists. They achieved AUC scores of 0.81, 0.87 and 0.84 without the help of the CAD system. Together with the CAD system, the AUC scores improved to 0.90, 0.90 and 0.88.

Yang et al. [109] differentiated LGG from glioblastoma with high accuracy (AUC of 0.97) based on T1ce MRI. The tumour was manually segmented followed by slice-level classification through the use of a 2D convolutional neural network (CNN), pre-trained on ImageNet, fine-tuned on 90 patients and evaluated on a test set of 23 patients.

An automated grading system on conventional MRI (T1, T1ce, T2 and FLAIR) was presented by Zhuge et al. [306]. Tumours were segmented using a U-Net. After segmentation, the tumour ROI was extracted and classification performance was compared between 2D mask-RCNN (applied on the slice with largest tumour contour) and a 3D CNN using a ResNet backbone. The networks were trained and evaluated on data from 315 patients collected from BraTS and TCIA. An accuracy of 96% was achieved with the 2D mask RCNN and 97% with the 3D CNN.

State-of-the-art performance on IDH mutation status prediction was reported by Chang et al. [307]. They predicted IDH mutation based on pre-operative MRI (T1, T1ce, T2 and FLAIR) of 496 patients. Tumours were manually delineated and classified by four 2D CNNs (one for each modality). Through the combination of the four probabilities with age

and a logistic regression classifier, an AUC of 0.95 was obtained.

Zhang et al. [308] used radiomics features and a random forest classifier to predict IDH status. Their dataset included pre-operative MRI, including T1, T1ce, T2, FLAIR and DWI sequences, of 120 patients with WHO grade III and IV. In total 2970 imaging features were extracted. On the validation cohort of 30 patients, they reached an AUC score of 0.92.

Application of a radiomics pipeline for IDH and 1p/19q co-deletion prediction using only routinely acquired T1ce and FLAIR MRI was investigated by Zhou et al. [309]. They collected a large training dataset of 538 glioma patients from multiple institutions. Imaging features describing shape, intensity and texture were combined with age and fed to a random forest classifier. The models were validated on public data from TCIA. An AUC score of 0.92 was achieved for predicting IDH status and 0.72 for 1p/19q co-deletion status. Age and shape features offered the highest predictive value.

Choi et al. [310] applied a recurrent neural network on dynamic susceptibility contrast perfusion MRI. Signal intensity-time curves were extracted from different tumour subregions that were segmented using a CNN followed by manual correction. They evaluated their approach for IDH prediction (AUC of 0.95) and for 1p/19q co-deletion prediction (AUC of 0.78). The 1p/19q co-deletion prediction model was only trained for IDH mutant glioma.

Yogananda et al. [311] trained a 3D dense U-Net for voxel-level IDH status prediction which can also be used for whole-tumour segmentation. Two networks were compared: one is only trained on T2 MRI and the other is trained on T1ce, T2 and FLAIR. Data from 214 glioma patients was acquired from the cancer imaging archive. The two networks attained a similar performance (AUC of 0.98). The same approach is applied for the 1p/19q co-deletion prediction task [312]. Now a dataset of 368 patients is used from TCIA for training and evaluation. For 1p/19q co-deletion prediction an AUC score of 0.95 is obtained.

Akkus et al. [313] analysed T1ce and T2 MRI of 159 LGG patients to predict 1p/19q co-deletion status. The tumour was delineated semi-automatically and each slice was classified using a multi-scale 2D CNN achieving an accuracy of 88%.

A radiomics pipeline to predict 1p/19q co-deletion in LGG trained on a private institutional dataset containing 284 patients and validated on an external dataset of 129 patients from TCIA was presented by van der Voort et al. [314]. Age, sex and imaging features were used to train an

SVM classifier. The algorithm achieved an AUC of 0.72 on the external validation dataset.

Table 7.1: Overview of studies on non-invasive prediction of grade, IDH mutation and 1p/19q co-deletion status of glioma.

Author	Task	Dataset	Method	Result
Zacharaki et al. (2011) [315]	distinguishing metastases, meningioma grade I and glioma grade II, III and IV	97 patients T1, T1ce, T2, FLAIR, DSC rCBV	manual segmentation age, shape, intensity features classification with kNN, decision tree, SVM	Accuracy = 76% Sensitivities = 82% (grade II), 29% (grade III), 82% (grade IV), 96% (metastases)
Subashini et al. (2016) [303]	distinguishing low-grade from high-grade astrocytoma	200 patients T2	Semi-automatic segmentation shape, intensity and texture features LVQ classifier	Accuracy = 91%
Skogen et al. (2016) [316]	predicting WHO grade II, III, IV	95 patients (grade II, III, IV) T1ce	manual 2D segmentation texture features ROC analysis	AUC = 0.91 (II, III vs. IV) AUC = 0.84 (II vs. III) AUC = 0.73 (III vs. IV)
Hsieh et al. (2017) [304]	distinguishing lower-grade (II, III) from high-grade GBM (IV)	107 patients (grade II, III, IV) T1ce	manual 2D segmentation histogram, texture features logistic regression	AUC = 0.89

Author	Task	Dataset	Method	Result
Tian et al. (2018) [317]	predicting glioma WHO grade II, III, IV	153 patients T1, T1ce, T2 + DWI, PWI	Manual segmentation histogram and texture features SVM	Accuracy = 97% (II, III vs IV)
Yang et al. (2018) [318]	distinguishing lower-grade (II, III) from high-grade GBM (IV)	113 patients (grade II, III, IV) T1ce	Manual ROI segmentation (pre-trained) 2D CNN: AlexNet, GoogleNet	AUC = 0.97
Zhuge et al. (2020) [306]	distinguishing lower-grade (II, III) from high-grade GBM (IV)	315 patients (grade II, III, IV) T1, T1ce, T2, FLAIR	Automatic segmentation U-Net 2D Mask-RCNN or 3D CNN	Accuracy = 97%
Chang et al. (2018) [307]	IDH mutant vs. IDH wildtype	496 patients (grade II, III, IV) T1, T1ce, T2, FLAIR	Manual ROI segmentation 2D CNN: ResNet34 Logistic regression combining age with probability output	AUC = 0.95

Author	Task	Dataset	Method	Result
Yu et al. (2017) [319]	IDH1 mutant vs. IDH1 wildtype	140 patients (grade II) FLAIR	Automatic segmentation with 2D CNN Location, shape, texture and histogram features SVM, AdaBoost	AUC = 0.86
Zhang et al. (2017) [308]	IDH mutant vs. IDH wildtype	120 patients (grade III, IV) T1, T1ce, T2, FLAIR, DWI (ADC)	Semi-automatic segmentation Anatomical, shape, texture and histogram features Random forest classification	AUC = 0.92
Arita et al. (2018) [320]	IDH mutant vs. IDH wildtype	199 patients (grade II, III) T1, T1ce, T2, FLAIR	Manual segmentation Location, shape, texture features LASSO regression	Accuracy = 87%
Chang et al. (2018) [321]	IDH mutant vs. IDH wildtype 1p/19q co-deleted vs. 1p/19q Intact MGMT methylated vs. unmethylated	259 patients (grade II, III, IV) T1, T1ce, T2, FLAIR	Automatic segmentation with 2D CNN 2D CNN: residual network	AUC = 0.91 (IDH) AUC = 0.88 (1p/19q) AUC = 0.81 (MGMT)

Author	Task	Dataset	Method	Result
Choi et al. (2019) [310]	IDH mutant vs. IDH wildtype 1p/19q co-deleted vs. 1p/19q Intact	463 patients (grade II, III, IV) T1, T1ce, T2, FLAIR, DSC perfusion MRI	Automatic segmentation with CNN followed by manual correction 2D convolutional LSTM	AUC = 0.95 (IDH) AUC = 0.78 (1p/19q)
Zhou et al. (2019) [309]	IDH mutant vs. IDH wildtype IDH mutant: 1p/19q co-deleted vs. 1p/19q Intact	744 patients (grade II, III, IV) T1ce, FLAIR	Manual segmentation Histogram, shape, texture and age features Random forest classification	AUC = 0.92 (IDH) AUC = 0.72 (1p/19q)
Yogananda et al. (2019) [311]	IDH mutant vs. IDH wildtype	214 patients (grade II, III, IV) T2	3D Dense U-Net	AUC = 0.98
Rathore et al. (2020) [322]	IDH mutant vs. IDH wildtype IDH mutant: 1p/19q co-deleted vs. 1p/19q Intact EGFRvIII in GBM	473 patients (grade II, III, IV) T1, T1ce, T2, FLAIR, DSC-MRI, DWI	Semi-automatic segmentation histogram, shape, anatomical, and texture features SVM	AUC = 0.85 (IDH) AUC = 0.75 (1p/19q) AUC = 0.87 (EGFRvIII)

Author	Task	Dataset	Method	Result
Akkus et al. (2017) [313]	1p/19q co-deleted vs. 1p/19q Intact	159 patients (grade II, III) T1ce, T2	Semi-automatic 2D segmentation 2D CNN	Accuracy = 88%
Han et al. (2018) [323]	1p/19q co-deleted vs. 1p/19q Intact	277 patients (grade II, III) T2	Manual segmentation shape, size, intensity and texture features Random forest	AUC = 0.76
Kim et al. (2019) [324]	1p/19q co-deleted vs. 1p/19q Intact	167 patients (grade II, III, IV) T1, T1ce, T2, FLAIR	Manual segmentation Texture, topological and pre-trained CNN features Random forest classification	AUC = 0.71
van der Voort et al. (2019) [314]	1p/19q co-deleted vs. 1p/19q Intact	284 patients + 129 from TCIA (grade II, III) T1ce, T2	Manual segmentation Intensity, texture, shape, texture, age and sex features SVM classifier	AUC = 0.72 (TCIA)
Yogananda et al. (2020) [312]	1p/19q co-deleted vs. 1p/19q Intact	368 patients (grade II, III, IV) T2	3D Dense U-Net	AUC = 0.95

Most of the studies included in Table 7.1 used manual or semi-automatic segmentations which might introduce variability and subjectivity to the classification pipeline and impede clinical adoption. However, as a lot of research is performed on automatic glioma segmentation with CNNs, manual delineation could be replaced with recent state-of-the-art automatic delineation algorithms.

An additional limitation is that existing studies often train and eval-

uate their models on a single, small dataset, often acquired from one institution. Hence their robustness to data from other clinical centres (with large variations in imaging protocols) remains to be evaluated on a completely independent dataset. Careful evaluation on external data that is not in any way used to train the algorithms is important to assess their generalisation performance.

Moreover, due to the limited amount of data, often radiomics methods are used where hand-engineered features are extracted that depend on expert opinion and are less robust to variations in image acquisition protocols. Convolutional neural networks, on the other hand, can automatically extract and classify features from complex imaging datasets with increased speed and without requiring human interaction resulting in a more objective computer-aided diagnosis tool. Several approaches in table 7.1 that use CNNs already illustrate this by achieving very high performances.

In this work we investigate the use of deep learning to develop an accurate, reproducible and fully automatic 3D pipeline to segment glioma and predict clinically relevant markers according to the most recent WHO guidelines based on routinely acquired pre-operative MRI. Automated diagnosis with deep learning remains a challenging task as large-scale and well-curated datasets of brain tumour scans comparable to ImageNet are unavailable. Existing datasets often include patients with missing image modalities and ground truth information on tumour characteristics.

7.3   Conclusion

This chapter started with explaining the basic anatomy of the brain, necessary to understand the different types of primary brain tumours that are defined by the World Health Organisation. We focused on the most common type of PBTs, glioma, and the most recent classification guidelines of the WHO to differentiate tumours based on malignancy (WHO grade) and molecular markers (IDH status and 1p/19q co-deletion). We then further discussed PBT epidemiology, symptoms, diagnosis, survival and different treatment options in relation to these important markers. After introducing the required background knowledge, we provided an overview of state-of-the-art literature on glioma segmentation and diagnosis with artificial intelligence techniques.

8 | Glioma grading with limited data: radiomics and pre-trained CNN features

As mentioned in previous chapter, determining the malignancy of glioma is highly important for initial therapy planning and prognosis. In this chapter, we investigate two feature extraction methods, radiomics features and features extracted using a pre-trained CNN, for the task of binary brain tumour grading in a limited data setting. This allows to evaluate whether we can design an accurate binary glioma grading system with limited data using hand-engineered features or features that are automatically learned by a convolutional neural network, pre-trained on natural images. Moreover, we compare the performance of pre-trained CNN features extracted from different input scales: one or multiple slices and with or without cropping to the tumour ROI.

Extraction of the radiomics features was performed by Stijn Bonte [325]. This work investigated the use of radiomics for primary brain tumour segmentation and classification. The results in this chapter have been presented during the 2018 Medical Imaging Summer School [326] and on the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2018 conference, published in Milan Decuyper et al. "Binary Glioma Grading: Radiomics versus Pre-trained CNN Features". In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2018. Ed. by Alejandro F. Frangi et al. Springer International Publishig, 2018. DOI: 10.1007/978-3-030-00931-1_57.

8.1 Introduction

In previous chapter we have seen that the optimal treatment strategy and prognosis of newly diagnosed glioma strongly relies on tumour malignancy (i.e. WHO grade). Whereas a watch-and-wait policy can be opted in case of low-grade glioma, maximum safe resection combined with appropriate chemotherapy and radiotherapy is recommended by EANO for high-grade glioma [273]. Biopsies for histopathological diagnosis negatively impact overall survival with a reported hazard ratio of 2.69 (95% CI 1.19-6.06; p=0.02) compared to wait-and-scan for low-grade glioma [279]. The invasive procedure involves risks, is subject to sampling error and the results may be subjective, depending on the neuropathologist performing the histopathological analysis [280]. Hence a biopsy to confirm diagnosis and grade of the tumour should be avoided and accurate non-invasive grading is preferred.

Non-invasive differentiation between low- and high-grade glioma is usually based on MRI with gadolinium-based contrast agents [4, 282]. The presence of contrast enhancement and necrosis are indicative of higher tumour malignancy, however 40-45% of non-enhancing lesions are subsequently found to be malignant (WHO grade III or IV) [282]. This results in reduced accuracy of non-invasive tumour grading (sensitivities ranging between 55% and 83%). Moreover, the ever-increasing amount and complexity of MR image data raises the burden of accurate data analysis and dramatically increases the workload of radiologists.

Computer-aided diagnosis may provide a way to handle this data explosion and increase diagnostic accuracy [328]. These systems can automatically process MR images, calculate quantitative features describing tumour characteristics and combine them to estimate tumour type and grade through the use of artificial intelligence. The time required for diagnosis can be reduced and accuracy and treatment planning enhanced while avoiding the need for biopsy.
Towards computer-aided brain tumour diagnosis, many studies investigate the use of radiomics [301, 302, 328]. High performances in binary tumour grading are achieved (see section 7.2.3). In current radiomics studies, however, often input of domain experts is required, such as manual segmentation data, making these methods not reproducible and not fully automatic. Additionally, most CADx methods are trained and evaluated on data from one clinical centre. Hence these systems are potentially not robust or applicable to data from other centres due to

large variations in imaging protocols.

Our goal is to investigate the use of deep learning to develop an accurate, reproducible and fully automatic CADx system. Training deep networks from scratch is however often not feasible when only a limited amount training data is available. In this chapter, we therefore use hand-engineered radiomics features and features extracted through a pre-trained CNN to discriminate GBMs from lower-grade glioma using a small dataset of 285 glioma cases. This allowed us to assess the predictive value of the radiomics features with pre-trained CNN features on the same heterogeneous dataset and to gauge the potential of deep learning for brain tumour diagnosis. Moreover, we investigate whether best performance is achieved when cropping the MRI to the tumour region-of-interest, which requires a prior segmentation or detection step, or if accurate classification can be obtained based on the entire MRI.

8.2 Materials and methods

8.2.1 Data

The data used in this work originates from the BraTS 2017 training database [285, 329]. It contains multi-institutional routine clinically acquired pre-operative MRI scans of 210 glioblastoma and 75 lower-grade glioma (WHO grade II and III) with pathologically confirmed diagnosis. For each case, multi-modal MRI are available including a T1-weighted, a post-contrast T1, a T2-weighted and a T2 fluid attenuated inversion recovery sequence. The MRI scans originate from multiple (n=19) institutions and were acquired with different clinical protocols and scanners resulting in a very heterogeneous dataset. All glioma cases are segmented manually by one to four raters and approved by experienced neuro-radiologists. Delineated tumour regions are the GD-enhancing, peritumoural oedema and the necrotic and non-enhancing tumour tissues. All subject's sequences are co-registered to the same anatomical template, interpolated to a 1 mm^3 voxel size and skull-stripped, i.e. only the brain area is preserved.

As non-invasive determination of tumour malignancy is mostly based on T1ce MRI, we only use this sequence and segmentation data in the current study to perform binary grade prediction. Example cases from the BraTS 2017 training dataset are shown in figure 8.1. For the

glioblastoma case (figure 8.1a), one clearly observes a necrotic core with thick, surrounding contrast enhancement. No contrast enhanced tissue is observed in the lower-grade glioma (figure 8.1b).
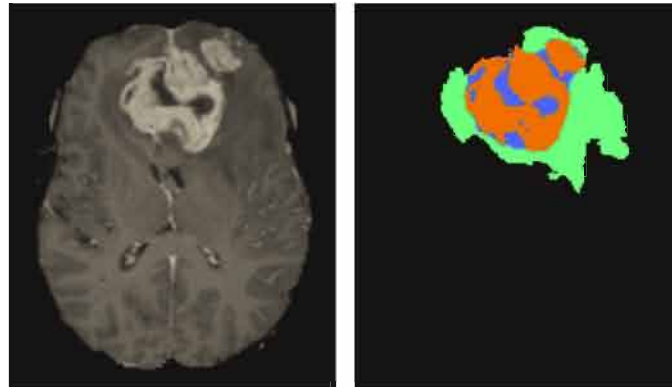
(a) Example of a glioblastoma case.
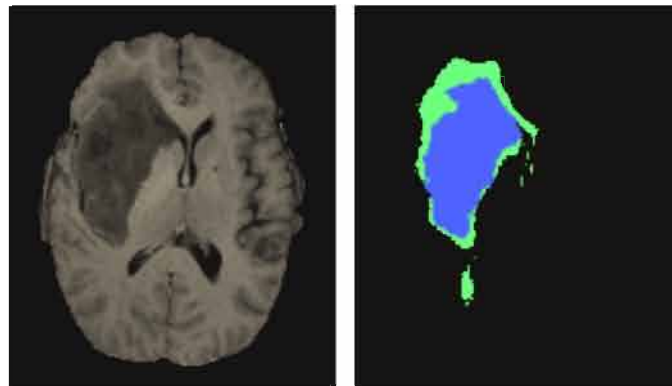
(b) Example of a lower-grade glioma case.

Figure 8.1: Example cases from the BraTS 2017 training dataset [285, 329]. An axial slice is shown from the T1ce sequence and the manual segmentation map. Blue denotes necrosis + non-enhancing tissue; orange indicates enhancing tissue and green peritumoural oedema.

8.2.2 Feature Extraction: Radiomics

In the radiomics feature extraction approach, all scans were first bias corrected using SPM12 (version 6906, Wellcome Trust Centre for Neuroimaging, University College London) [330] running on MATLAB R2017b (The MathWorks, Inc., Natick, MA). Bias fields are smooth, low-frequency signals corrupting the MRI intensities and are caused by small inhomogeneities in the magnetic field of the MRI scanner. Although they usually do not impede visual inspection, they can influence the extraction of

radiomics features.

Next to bias correction, the image intensities were also normalised following the robust white stripe normalisation algorithm [331]. This is necessary since MRI scans are recorded in arbitrary units resulting in highly varying intensities across different MRIs. Radiomics features are extracted that describe the intensities present in the image and therefore normalisation is required. In white stripe normalisation, the intensities are normalised by subtracting with the mean and dividing by standard deviation of the normal-appearing white matter tissue intensities. This way, healthy white matter has zero mean and unit variance for all patients and images.

After pre-processing of the MRI, the tumour appearance is quantified through the extraction of numerous radiomics features as illustrated in figure 3.10. The manual segmentation labels were used to define five different tumour regions: total abnormal region, tumour core (necrosis+enhancing), enhancing tissue, necrosis and oedema. In every region 207 quantitative features were calculated: 14 histogram, 8 size and shape and 185 texture features. These features were calculated according to the definitions in Aerts el al. [332] and Willaime et al. [333].

Histogram features describe the intensity distribution in each region, i.e. contain information on heterogeneity. Example statistics are mean, median, standard deviation, minimum, maximum, skewness (asymmetry), kurtosis (presence of heavy tails) etc.

Shape and size features include volume, maximal diameter, surface, surface to volume ratio etc.

Texture features further describe tumour heterogeneity and spatial distribution of the different intensities. Different types of texture features are calculated: 138 grey-level co-occurence, 22 grey-level run-length matrix, 12 neighbourhood grey-tone difference matrix and 13 grey-level size-zone matrix features.

A complete description of the radiomics feature extraction process can be found in the work of Stijn Bonte [325].

8.2.3 Feature Extraction: Pre-trained CNN

Instead of extracting hand-engineered features from the segmented tumour volumes, deep features were extracted using a pre-trained convolutional neural network. Hence we employ a transfer learning approach

as explained in section 2.3.3. We expect that features learned by the network to classify natural images in ImageNet also hold predictive value for classification of brain tumours. Through the use of a pre-trained network, there is no need to train a CNN on the limited dataset used in this chapter and we can use a more traditional machine learning classifier as in the radiomics procedure (see next section).

The VGG-11 architecture was used consisting of 8 convolutional and 3 fully connected layers [45]. The architecture is depicted in figure 8.2, after removing the last two fully connected layers. The model, pre-trained on the ImageNet dataset, was loaded from the PyTorch torchvision package. Features were obtained by forward propagating an MRI slice through the network and extracting the 4096-dimensional output of the first fully connected layer. The first layer was chosen under the assumption that earlier layers learn more generally applicable features than layers deeper into the network. Before being propagated through the network, the slices were pre-processed to match the expected input of the pre-trained pytorch models. The image intensities were scaled to a range between zero and one, the slice was resized to a shape of 224 × 224 through bilinear interpolation and finally normalised with mean and standard deviation values provided by PyTorch. Because the model expects RGB images, the MRI slice was provided at the R channel and the B and G channels were set to zero.

Feature extraction and corresponding grading performance was evaluated for four different ways of providing the T1ce scan at the input of the network. The different approaches are illustrated in figure 8.2.

In a first approach, the manual segmentation labels were used to select the slice in the T1ce scan containing the largest tumour contour and crop this slice to the size of the tumour (figure 8.2: method 1). After applying the pre-processing steps explained above, the tumour patch was propagated through the network, thereby obtaining one 4096-dimensional feature vector per patient with a corresponding label indicating LGG or GBM.

For the second method, all tumour slices were propagated through the network after being cropped to the size of the tumour (figure 8.2: method 2). Hence, multiple feature vectors are obtained for each patient and every slice or feature vector was classified into one of three classes: (1) LGG, (2) GBM where only oedema is visible, (3) GBM with contrast enhancement and necrosis. In each slice, either a LGG or a GBM is visible. Additionally, a GBM may in some slices only display oedema

and no contrast enhancement and necrosis. Because these slices may have a similar appearance as LGG slices, this could be confusing for the classifier and therefore a separate class was added for GBM slices only demonstrating oedema.

In the third method, the same slice was selected as in the first approach, but now it was not cropped (figure 8.2: method 3). Hence the entire slice was propagated through the network and again one feature vector is obtained per patient.

To design a system able to classify a T1ce scan without requiring segmentation information, a fourth method was investigated. Here, every slice of the T1ce scan was propagated through the network (figure 8.2: method 4). One entire scan contains 155 slices, so 155 feature vectors were obtained for each patient and a fourth class, besides the three classes of the second method, was added for slices containing no tumour. Using this approach, no segmentation data is required to classify slices from a T1ce sequence of a new patient resulting in a fully automatic CAD system. The method used to aggregate results from multiple feature vectors from one patient is explained in next section.
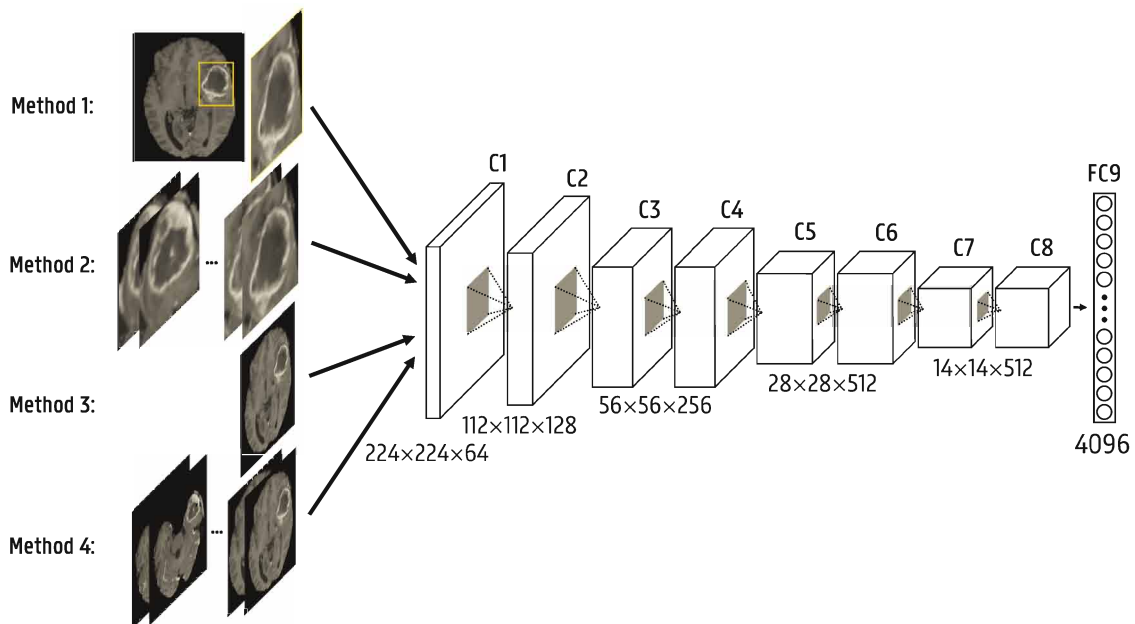
Figure 8.2: Feature extraction with the pre-trained VGG-11 CNN. Method 1: Propagate tumour region of the slice containing the largest tumour contour. Method 2: Propagate tumour region of all tumour slices. Method 3: Propagate entire slice containing the largest tumour contour. Method 4: Propagate all slices

8.2.4   Random Forest classification

After feature extraction, classification was performed with the goal to predict whether a patient has a glioblastoma or lower-grade glioma. The feature vectors were first scaled to unit norm and features showing no variance between different samples were removed.

For classification, the python scikit-learn *RandomForestClassifier* was used with 200 decision trees. All Random Forest models were trained for the binary classification task except for the second and fourth method of feature extraction with the pre-trained CNN. In those cases, the RF model was trained to classify a slice into one of 3, respectively 4 classes as explained in section 8.2.3. For each patient, multiple slices were classified. All predictions were combined by calculating their mean probability and the sum of the probabilities of the two GBM classes was used as the final probability value of having a GBM. The performance of the classifier was evaluated on a separate test set containing 57 (20%) of the 285 glioma cases. The class ratio of 210:75 was equal in both training and test set. To enhance sensitivity and specificity of the model, the probability threshold of classifying a glioma as GBM was optimised through 5-fold cross-validation applied on the training set. The training and evaluation process was repeated 50 times with different random splits in train and test set to estimate average performance and variability of the model.

8.3   Results

For the radiomics feature extraction and for each of the feature extraction methods with the pre-trained CNN, a random forest model was trained and evaluated to asses the predictive value of the resulting feature vectors. The area under the ROC curve, accuracy, sensitivity and specificity scores are reported in table 8.1.

The RF model trained on the radiomics features achieves the highest performance with an average AUC score of 96%. The optimal probability threshold to classify a case as a glioblastoma that balances sensitivity and specificity was 0.72. In random forest classifiers, feature importances can be extracted based on where they are used in the trees and how much they decrease impurity. The most predictive features were volume of

Table 8.1: Mean(std) (%) area under the ROC curve, accuracy, sensitivity and specificity classification scores.

Feature Extraction Method	AUC	Acc.	Sens.	Spec.
Radiomics	96.4(2.6)	89.6(3.8)	89.9(5.4)	88.8(8.6)
CNN: Method 1	92.2(3.9)	83.8(4.6)	83.3(5.2)	85.2(9.6)
CNN: Method 2	93.5(3.0)	86.1(4.3)	85.4(5.4)	88.5(8.1)
CNN: Method 3	86.8(4.6)	79.1(4.9)	78.6(6.4)	80.7(9.6)
CNN: Method 4	91.1(3.6)	82(5.3)	81.5(7.2)	83(9.6)

contrast enhanced tissue and histogram and texture feature extracted form the tumour core region.

With features extracted using a pre-trained CNN, best results were obtained when zooming in on the tumour region and using all tumour slices (CNN, method 2). When using features extracted from the entire slice containing the largest tumour contour (CNN, method 3), performance is lower with an AUC of 87% compared to 92%. However, when predicting glioma grade based on all slices of the T1ce scan (CNN, method 4), performance could be improved to an AUC score of 91%.

Classifying a T1ce scan was possible within 0.3 seconds with *CNN: method 1* and *3*, 12 seconds with *CNN: method 2* and 30 seconds with *CNN: method 4* on a Macbook Pro with 2.8 GHz Intel Core i7 CPU. Propagating all slices through the CNN required most of the computation time.

8.4   Discussion

The results shown in table 8.1 show that the best performance is achieved with the radiomics features, matching or even outperforming state-of-the-art accuracies reported today. This proves that hand-engineered features hold high predictive value to classify brain tumours. Even though there is a growing trend towards the use of deep learning algorithms, more traditional radiomics approaches can still be very valuable, especially when dealing with small datasets. They have the additional benefit of interpretability. Analysis of the feature importances shows that the RF classifier indeed focuses on the tumour core regions and the presence of enhancing tissue which is expected based on existing insights

into non-invasive grading (see section 7.2.1).

The radiomics features were, however, extracted from manually segmented tumour tissues which is time-consuming and introduces subjectivity. A lot of research has been performed towards automatic segmentation algorithms reaching high performances (see section 7.2.2) and manual segmentation could therefore be replaced with automatic delineation. Automatic segmentation is also more objective and reproducible resulting in less variation between radiomics features extracted across different institutions. This can result in more generalisable radiomics pipelines. The difference in performance between using a state-of-the-art automatic segmentation algorithm or manual segmentation remains to be investigated.

In this chapter we only included a brief overview and discussion on the use of radiomics for brain tumour grading to set a baseline for comparison with deep learning approaches. A more complete analysis can be found in: Stijn Bonte. "Artificial intelligence in medical imaging for the diagnosis of primary brain tumours". PhD thesis. Ghent University, 2018, pp. XVIII, 222. ISBN: 9789463551687.

Although performance is slightly lower compared to the radiomics results, accurate grading could be achieved with a pre-trained CNN as feature extractor as well.

With the first method of feature extraction through a CNN, an AUC is achieved of 92% while only requiring a bounding box around the tumour which is considerably less time-consuming than accurate segmentation of the different tissues.

Furthermore, when estimating grade based on all tumour slices, performance could be improved to an AUC of 93.5%. Drawing a 3D bounding box can easily be performed manually but an automatic segmentation algorithm could be used as well. We expect that small variations or segmentation inaccuracies will not have a large influence on the extracted bounding box and thus the classification performance.

Features extracted from the entire slice were less informative but by calculating an ensemble prediction from all slices, accurate grading could still be achieved reaching a performance similar to the first method. Through the use of all slices, much more samples are created resulting in a larger training set which can explain the improved performance. Moreover, by aggregating predictions from the entire MRI, a final prediction can be made based on the entire tumour region instead of just one slice of the tumour. This way, a binary grading system could be designed that

is fast, does not require segmentation or manual input to classify new T1ce sequences and is trained on a very heterogeneous dataset making it robust to variations in imaging protocols.

These results show that a CNN, trained on an entirely different image dataset containing natural images, is able to extract informative features from MRI sequences as well. Their predictive value is lower than radiomics features extracted from manually segmented tumour volumes, but we expect that by fine-tuning the network on brain tumour MRI, results could further be improved. Chapter 9 will focus on gathering more data, allowing to specialise CNNs on brain MRI and open the path towards more accurate and automatic brain tumour characterisation. Best results are achieved when allowing the CNN to focus on the tumour ROI using segmentation labels. When propagating the entire MRI, a lot of information is included that is not necessarily informative for characterising the brain tumour and could lead to overfitting. This is especially important when training a CNN from scratch. In chapter 10, we will therefore train a CNN on the 3D tumour region of interest in order to limit overfitting. To automatically determine the tumour bounding box, a segmentation algorithm is required which will be designed in chapter 9.

The performance of the classifiers could further be improved by using more sophisticated feature selection methods, providing features from additional MRI sequences, ensembles of different models etc. The main goal of this study was, however, to compare the performance between radiomics and pre-trained CNN features and to gain first insights in the potential of CNNs for non-invasive characterisation of brain tumours. We therefore did not optimise every possible design parameter. Nonetheless very high classification performances are already achieved.

8.5 Conclusion

In this chapter, we assessed the predictive value of radiomics features and features extracted using a pre-trained CNN for binary brain tumour grading in a limited data setting. Classification results showed that the best performance is achieved with shape, intensity and texture features extracted from manually segmented tumour volumes. Features from a pre-trained CNN, on the other hand, had a high predictive value as well and allowed to design an accurate, fast, automatic and robust binary

grading system. These results indicate that CNNs hold the potential to develop an accurate, reproducible and fully automatic CAD system. In the next chapters, a large dataset of glioma cases will be collected and CNNs will be trained from scratch to segment glioma and not only predict grade but also IDH and 1p/19q co-deletion status.

9 | Brain tumour segmentation

In previous chapters, we already discussed the importance of automatic segmentation in the diagnosis and management of glioma. On this account, we design a deep learning algorithm in this chapter that delineates the different glioma tissues on routinely acquired MRI. The segmentation network needs to be automatic, accurate, fast and generalisable to data from different institutions.

This work has been presented on the Medical Imaging with Deep Learning (MIDL) 2020 conference [334] and published as a part of an A1 publication: Milan Decuyper et al. "Automated MRI based pipeline for segmentation and prediction of grade, IDH mutation and 1p19q codeletion in glioma". In: *Computerized Medical Imaging and Graphics* 88 (Mar. 2021). ISSN: 08956111. DOI: 10.1016/j.compmedimag.2020.101831.

9.1 Introduction

Delineation of the different brain tumour tissues on MRI is not only a necessary pre-processing step for the radiomics pipeline. The tumour grading results with the pre-trained CNN in previous chapter showed that best performance was achieved when cropping the input MRI to the tumour region of interest. This allowed the network to focus on the tumour appearance. Moreover, reducing the input size of the CNN also reduces memory and computational requirements. Using segmentation as a pre-processing step to detect the tumour and define the 3D bounding box can therefore also help to improve the subsequent classification

189

performance with CNNs.

Next to pre-processing, automatic delineation is also important for surgery planning, volume estimation and assessing tumour progression and treatment response. We can conclude that segmentation plays a crucial role in computer-aided characterisation and clinical management of brain tumours and we therefore design an automatic segmentation algorithm in this chapter.

Existing automatic segmentation algorithms using U-Net architectures already achieve very high performances matching manual segmentations performed by radiologists (see section 7.2.2). These algorithms require that all four MRI sequences (T1, T1ce, T2 and FLAIR) are available. This is not always the case in clinical practice. Often there are some sequences not available because they are not acquired or are not of sufficient quality due to artefacts, motion blurring, noise etc. In chapter 10, we use data acquired from multiple public datasets to train a classification network. Whereas for almost all patients, a good quality T1ce scan is available, for some patients, a good quality T1, T2 or FLAIR MRI is lacking. All four sequences were only available for 60% of the patients. Only 67% of the cases include a T1 scan, 92% a T2 scan and 65% a FLAIR sequence. As we want to use a dataset that is as large as possible and be able to accurately segment glioma even when not all MRI are available, we develop an automatic segmentation algorithm that is robust to these missing modalities.

9.2 Automatic segmentation

9.2.1 The BraTS 2019 dataset

To train the segmentation network, we used the BraTS 2019 training dataset [285, 293, 329]. This dataset contains data from the BraTS 2017 dataset with additional data from 50 patients. Accordingly, data is included from 335 patients (76 glioma WHO grade II, III and 259 glioma grade IV). Routine clinically acquired pre-operative T1, T1ce, T2 and FLAIR MRI are provided from multiple institutions together with manual segmentation maps denoting the GD-enhancing, peritumoural oedema and the necrotic and non-enhancing tumour core regions (see figure 7.5 for an example). The MRI and segmentation maps are provided

in Neuroimaging Informatics Technology Initiative (NIfTI) format[1] with image dimensions of 240 × 240 × 155 voxels.

All MRI were co-registered to the same anatomical template, interpolated to 1 mm^3 voxel sizes and skull-stripped.
Next to the pre-processing steps performed by BraTS, we independently normalised each sequence by subtracting the mean and dividing by the standard deviation. The mean and standard deviation are only calculated based on the brain (non-zero) voxels to limit the influence of the amount of background voxels on the normalisation statistics. Furthermore the MRI are cropped to the brain region.

9.2.2 Architecture

In recent BraTS challenges, U-Nets have shown state-of-the-art performance for brain tumour segmentation (section 7.2.2). A U-Net is an encoder-decoder network that combines semantic and spatial information through the use of skip connections from the encoder to the decoder which allows to segment fine structures very well. We therefore implemented a 3D U-Net similar to the architecture proposed by Isensee et al. [336] as illustrated in figure 9.1. The sizes denoted in figure 9.1 are shown for an input patch of 112 × 112 × 112 voxels.
The network has four input channels (one for each modality), 32 feature maps at the highest resolution, five levels (depths in the U shape) and four output channels (background, necrosis, oedema and enhancing tissue). Every encoding stage consists of two convolutional blocks comprising a convolutional layer with kernel size 3, followed by instance normalisation and leakyReLU activation. Instance instead of batch normalisation was used as the exponential moving averages of mean and variance within small batches (see next section) are unstable. At each encoding level the amount of feature maps is doubled and after each encoding part, the feature map sizes are halved with a max-pooling layer. The decoding part again comprises two convolutional blocks. The number of filters is reduced right (by the second convolutional layer) before upsampling to increase the feature map size. Trilinear upsampling is used instead of transposed convolutions to limit the number of parameters and memory consumption. This allows a suitable number of feature maps while limiting overfitting and not exceeding the GPU memory limit.

---

[1] https://nifti.nimh.nih.gov

After the last decoding stage, the network ends with a convolutional layer with a $1 \times 1 \times 1$ kernel and four output channels followed by softmax activation to convert the values into class probabilities. The final output segmentation map has the same size as the input of the network.
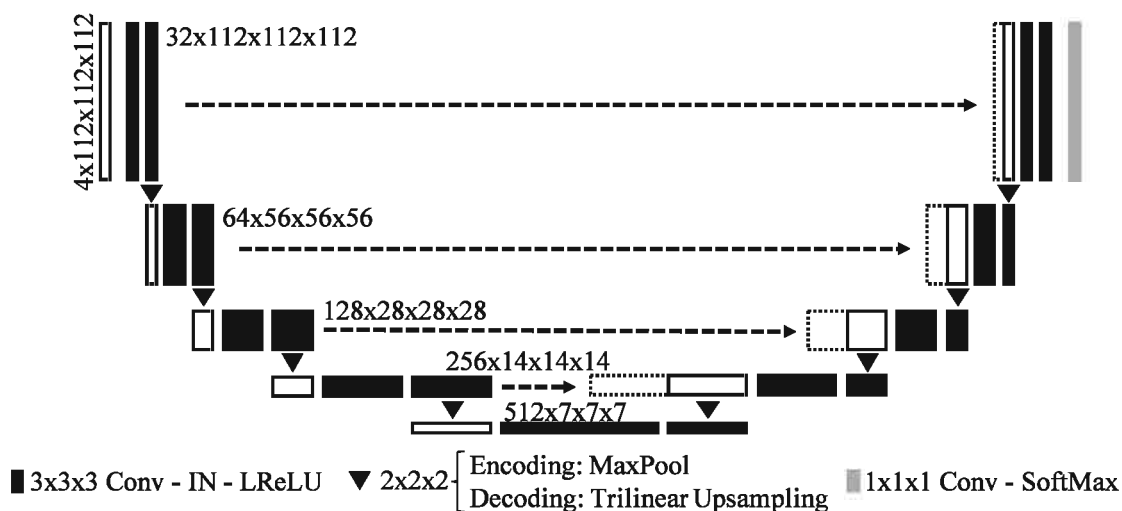
Figure 9.1: Schematic overview of the used U-Net architecture for automatic segmentation of glioma.

9.2.3 Training

The network is trained on patches of $112 \times 112 \times 112$, randomly sampled from the MRI brain volume, and a small batch size of two because of the high memory consumption of 3D convolutions. Patch extraction during training is necessary as the network cannot be trained on the entire MRI volumes due to memory limitations. Moreover, it allows to extract many different data samples from one MRI which helps to reduce overfitting. The input dimensions should be a power of two and at least $2^4 = 16$ as the size is halved four times and to preserve the same dimensions after downsampling and upsampling. After cropping the MRI to the brain region, the dimensions are on average $136 \times 169 \times 138$. A size of 112 was chosen as it is small enough in order to extract many different patches and fit into memory during training. Additionally, it is sufficiently large to contain enough context information on the different tissue intensities in the image.

The network was trained using the Adam optimiser with an initial learning rate $lr_{init} = 10^{-4}$ and an L2 weight decay of 10^{-6}. The learning

rate is halved every time the validation loss did not improve in the last 50 epochs. In case of no improvement for 250 epochs, early stopping is applied. A combination of cross-entropy and multi-class soft Dice loss was used as the optimisation metric. Soft Dice loss uses the probabilities for calculating the Dice overlap score instead of the binary predictions after thresholding. The network was implemented in Python using PyTorch and trained on an 11GB NVIDA GTX 1080 Ti GPU.

Sixty patients were held out for validation (40 GBM and 20 LGG cases) and the network was trained on the remaining 275 patients. To limit overfitting, data augmentations such as flipping and random axial rotations were applied on the fly during training before patch extraction to prevent introduction of boundary effects.

To increase robustness of the segmentation network to missing T1 and T2 or FLAIR modalities, channels dropout was applied to simulate this. Different input channels were randomly set to zero during training with a probability of 50%. We made sure that at least the T1ce and a T2 or FLAIR sequence was available at the input. The T1ce scan is important to accurately segment the contrast enhancing tissue, whereas a T2 sequence more clearly shows oedema.

9.2.4 Evaluation

During evaluation, the entire brain volume can be propagated through the network and no patch extraction is necessary. After cropping to the brain region, zero padding is added to each dimension to reach a size that is a power of two. The network is evaluated on the BraTS 2019 validation dataset containing 125 patients. Dice scores and robust Hausdorff distances are reported as calculated by the online evaluation platform[2]. The Hausdorff distance denotes the maximal surface distance between the predicted and ground truth segmentation surfaces. Robust Hausdorff distance reports the 95% quantile over all surface distances.

The developed segmentation algorithm needs to be generalisable to data from different institutions with highly varying image acquisition protocols. In the next chapter we will train a tumour diagnosis network on public data from TCIA and evaluate it on data that was retrospectively collected at the Ghent University Hospital (GUH). This was done with permission from the local ethics committee, and informed consent

---

[2]https://ipp.cbica.upenn.edu

was waived (Belgian registration number: B670201838395 2018/1500). We will therefore, also qualitatively evaluate wether the network is able to accurately segment brain tumour MRI from our centre as well.

9.3   Results

9.3.1   Quantitative results

Segmentation results on the BraTS 2019 validation data are summarised in table 9.1 and table 9.2. Dice scores and Hausdorff distances are reported for the enhancing tumour (ET), whole tumour (WT) and tumour core (TC) regions and for different available modalities: all four sequences, only T1ce and FLAIR or only T1ce and T2. To illustrate the increased robustness to missing modalities, the results are included with (table 9.1) and without (table 9.2) randomly setting input channels to zero during training.

When training with channel dropout, a mean whole tumour dice score of 90% is achieved which lowers to 89% and 87% when only the T1ce and FLAIR and T1ce and T2 MRI are available respectively. The median WT Dice scores are higher: 92%, 92% and 89% respectively. For the other tumour regions the average Dice scores vary between 74% and 76% for enhancing tissue and 82%-83% for the tumour core. The network achieves a high average WT specificity of 99% and a sensitivity of 92% in case all modalities are available. When only providing T1ce and T2, the specificity and sensitivity measures are 99% and 88%. Without randomly removing channels while training, the difference in performance is larger with mean WT dice scores of 90% based on all sequences, 83% based on T1ce and FLAIR and 61% based on only T1ce and T2.

On the Nvidia 1080 Ti GPU, a patient's MRI can be segmented in less than one second.

Table 9.1: Segmentation results on the BraTS 2019 validation data with randomly setting input channels to zero while training. Metrics were computed by the online evaluation platform.

		Dice Score (%)			Hausdorff distance (mm)		
	Available modalities	ET	WT	TC	ET	WT	TC
Mean	T1, T1ce, T2, FLAIR	75.71	89.81	83.18	5.08	4.99	6.66
	T1ce, FLAIR	74.35	89.37	82.74	4.34	5.12	6.82
	T1ce, T2	74.09	86.98	82.20	5.58	7.15	7.37
Median	T1, T1ce, T2, FLAIR	85.29	92.08	89.59	2.24	3.16	3.74
	T1ce, FLAIR	84.86	92.05	89.75	2.24	3.16	3.32
	T1ce, T2	84.49	89.27	89.56	2.24	4.24	3.46

Table 9.2: Segmentation results on the BraTS 2019 validation data without randomly setting input channels to zero while training. Metrics were computed by the online evaluation platform.

		Dice Score (%)			Hausdorff distance (mm)		
	Available modalities	ET	WT	TC	ET	WT	TC
Mean	T1, T1ce, T2, FLAIR	76.33	90.02	79.68	3.89	5.72	6.97
	T1ce, FLAIR	64.37	82.77	69.33	51.88	15.09	26.13
	T1ce, T2	62.46	60.98	59.86	9.02	23.03	23.27
Median	T1, T1ce, T2, FLAIR	85.84	91.66	88.62	2.12	3.16	3.53
	T1ce, FLAIR	82.27	89.25	83.99	3.61	5.74	7.87
	T1ce, T2	75.83	69.24	71.82	3.46	13.93	13.40

9.3.2 Qualitative results

Figures 9.2 to 9.6 provide a qualitative overview of the segmentation performance. To avoid cherry picking, the cases are selected as best (figure 9.2), 75th percentile (figure 9.3), median (figure 9.4), 25th percentile (figure 9.5) and worst (figure 9.6) based on whole tumour Dice score. These Dice scores are 98%, 95%, 92%, 88% and 54% respectively. As the ground truth segmentation labels are not provided by BraTs for the validation data, these are not included.

In figures 9.2, 9.3 and 9.5 the predicted segmentations are very similar when all sequences or only two sequences are provided and therefore only one segmentation result is shown.
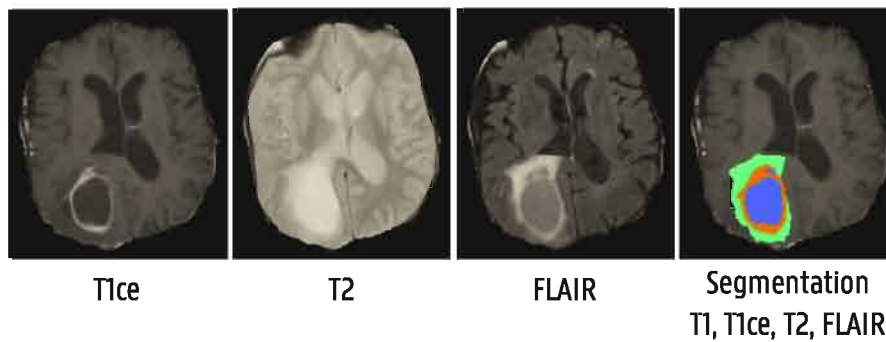
Figure 9.2: Example MRI and segmentation (overlaid on T1ce) of the patient with best whole tumour Dice score (98%). Blue denotes the necrotic and non-enhancing tissue, orange, enhancing tissue and green peritumoural oedema. The predicted segmentations are included when providing all sequences. Results when only providing T1ce and T2 or FLAIR are similar.
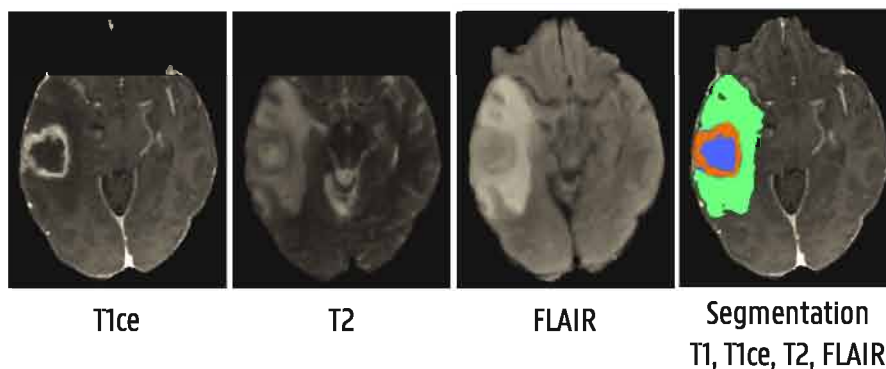
Figure 9.3: Example MRI and segmentation (overlaid on T1ce) of the patient with whole tumour Dice score at 75th quantile (95%). Blue denotes the necrotic and non-enhancing tissue, orange, enhancing tissue and green peritumoural oedema. The predicted segmentations are included when providing all sequences. Results when only providing T1ce and T2 or FLAIR are similar.

The brain tumour depicted in figure 9.4 possibly demonstrates infiltrating oedema around the ventricle on the left which is visible on the FLAIR sequence but much less on the T2 scan. When only providing the T1ce and T2 sequences, this oedema tissue is not detected. When the FLAIR scan is available, the segmentation network does delineate this oedema tissue.
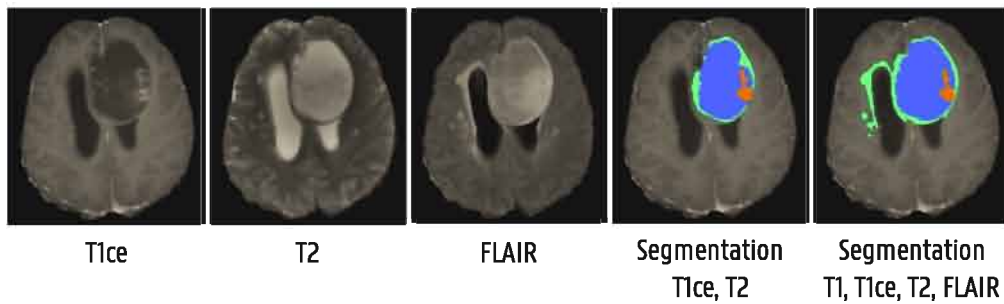
Figure 9.4: Example MRI and segmentation (overlaid on T1ce) of the patient with median whole tumour Dice score (92%). Blue denotes the necrotic and non-enhancing tissue, orange, enhancing tissue and green peritumoural oedema. The predicted segmentations are included when only providing the T1ce and T2 sequences and when providing all sequences. Results when only providing T1ce and FLAIR are similar as when providing all scans
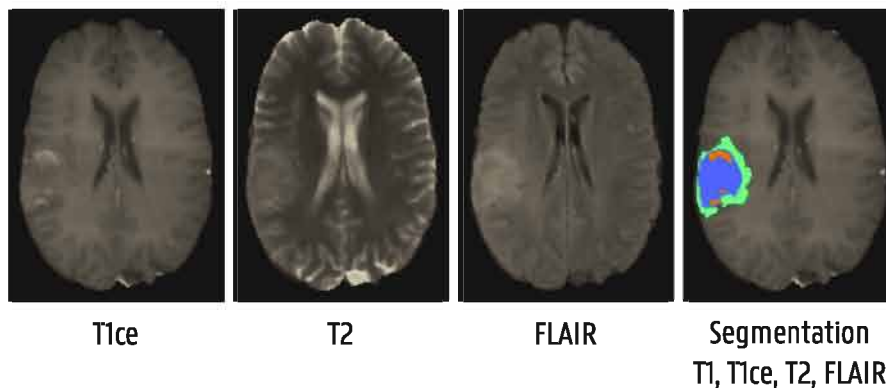
Figure 9.5: Example MRI and segmentation (overlaid on T1ce) of the patient with whole tumour Dice score at 25th quantile (88%). Blue denotes the necrotic and non-enhancing tissue, orange, enhancing tissue and green peritumoural oedema. The predicted segmentations are included when providing all sequences. Results when only providing T1ce and T2 or FLAIR are similar.

A similar observation is made in figure 9.6. This example shows the patient from the validation set with the lowest whole tumour dice score of 32% (when only providing T1ce and T2). Some of the oedema surrounding the segmented tumour core is missed. When providing all four MRI, more oedema is segmented and the WT dice increases to 54%.

| T1ce | T2 | FLAIR | Segmentation T1ce, T2 | Segmentation T1, T1ce, T2, FLAIR |
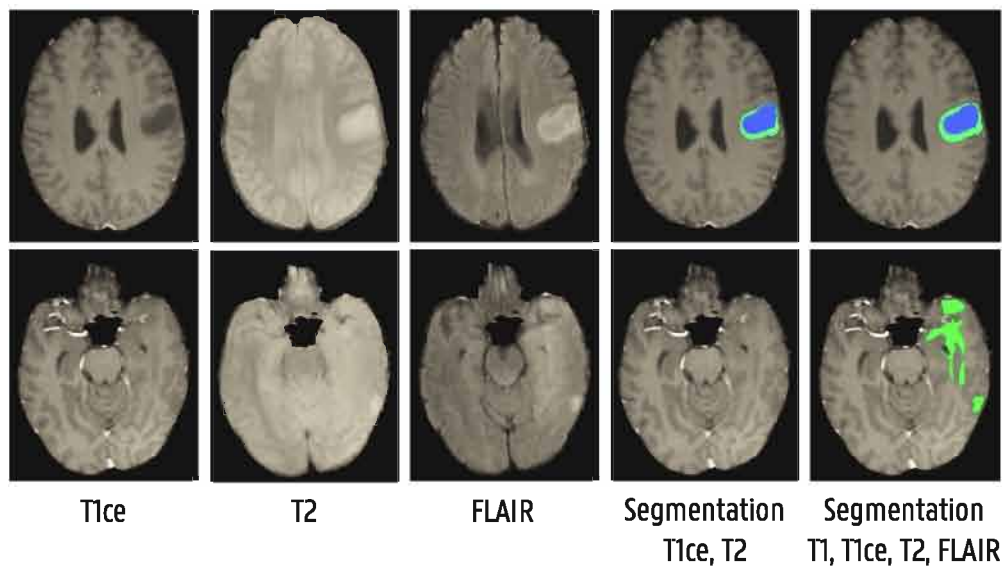
Figure 9.6: Example MRI and segmentation (overlaid on T1ce) for two different slices of the patient with lowest whole tumour dice score (54%). Blue denotes the necrotic and non-enhancing tissue, green indicates the peritumoural oedema. The predicted segmentations are included when only providing the T1ce and T2 sequences and when providing all sequences (for the bottom slice).

9.3.3   Ghent University Hospital data

Figure 9.7 illustrates the segmentation performance on the Ghent University Hospital data. Example MRI and segmentation maps are included from four different patients. Two cases with high grade glioma (first two rows in figure 9.7) and two with lower-grade glioma (bottom two rows in figure 9.7) are randomly selected. In all cases the brain tumour was detected by the segmentation network. No manual delineations are available to compare the obtained segmentation maps with, so we can only qualitatively assess the segmentation results.

9.4   Discussion

The segmentation results on the BraTS 2019 validation set show that very good dice scores are achieved. With an average whole tumour dice score of 90%, our segmentation algorithm matches the performance of state-of-the art algorithms of the BraTS 2019 challenge with the top three winning algorithms obtaining a mean WT dice score of 91%

TIce　　　　FLAIR　　　Segmentation
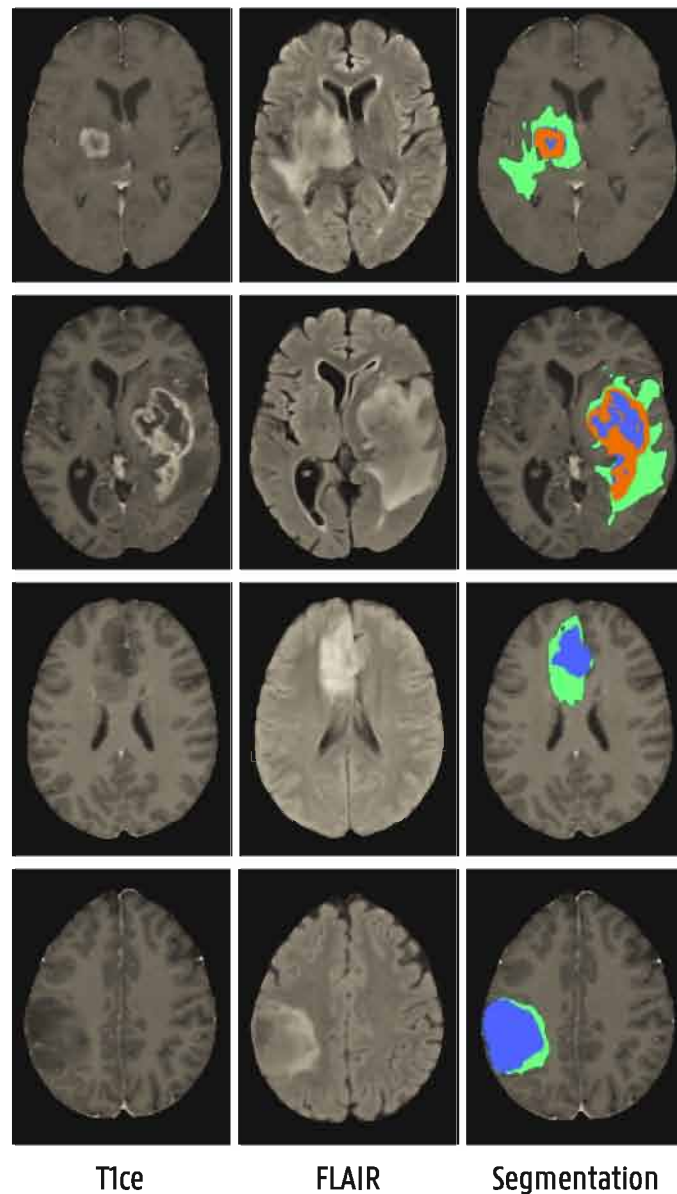
Figure 9.7: Example MRI and segmentation (overlaid on T1ce) of four different patients from the Ghent University hospital dataset. Blue denotes the necrotic and non-enhancing tissue, green indicates the peritumoural oedema.

according to the validation leaderboard [337]. The winning approach of the most recent BraTS 2020 challenge reports average dice scores of 89%, 85% and 82% for whole tumour, tumour core and enhancing tumour respectively [300]. The whole tumour and tumour core results of the network in this work are very similar. The ET Dice score, on the other hand, is lower (76% versus 82%). One of the most challenging

parts in brain tumour segmentation is distinguishing small blood vessels (that are enhanced in the T1ce scan) from enhancing tumour in the tumour core region. This is especially the case for lower-grade glioma patients that have no enhancing tumour. When the network predicts a few enhancing voxels while these are not segmented in the manual labels, an ET Dice score of 0 and a large Hausdorff distance is assigned to that case. So even though the error is not substantial (even for a single false positive) and can easily be interpreted by a clinician, these scores have a significant influence on the average metrics. This can also be observed when comparing the mean with median values in table 9.1 which are significantly higher. In top performing BraTS solutions often a post-processing step is applied where all enhancing tumour voxels are set to the tumour core label when the total number of ET voxels is below a certain threshold [299, 300]. This significantly improves the mean ET dice score. In a clinical scenario, however, it can be of crucial importance to detect small enhancing tumours, and applying this post-processing is not recommended. When comparing the achieved median ET Dice scores (86% versus 85% in this work), the difference is much smaller.

Additionally, the network shows increased robustness to missing modalities when randomly setting the T1 and T2 or FLAIR sequence to zero while training. There is only a small decrease in performance when only providing the T1ce and FLAIR or T1ce and T2 scans compared to all four MRI as input. Without randomly excluding image modalities during training, the performance with only two modalities is much lower. From a WT Dice score of 90% to 83% and 61% respectively. The higher scores when the T2 scan is missing compared when the FLAIR MRI is absent reveals that the FLAIR contains more discriminative information (especially for oedema). Indeed, the abnormal tissue contrast is better on FLAIR than on standard T2-weighted MRI. This can also be noticed in figures 9.4 and 9.6, discussed below.
Robustness to missing T1, T2 and FLAIR sequences is especially useful as not all four MRI modalities are available for all patients in both our public dataset and the Ghent University Hospital dataset that will be used in chapter 10. This way an accurate segmentation and tumour ROI extraction can still be obtained for these patients.
Other works that try to increase robustness to missing modalities use multiple encoders and decoders with correlation representations, feature fusion and attention mechanisms [338, 339]. Still large gaps in Dice scores are reported when a T2 or FLAIR sequence is missing. This shows that

channel dropout is a quite simple, yet effective way to increase robustness to absent input modalities.

Example segmentations shown in figures 9.2 to 9.6 demonstrate that the segmentation quality is high overall, even for the cases with the worst WT Dice scores. The very high specificity and slightly lower sensitivity of the network indicates that segmentation inaccuracies are due to parts of the surrounding oedema that are not detected. This is illustrated by the examples shown in figures 9.4 and 9.6 with the median and lowest WT dice score.
In figure 9.4, some of the FLAIR hyper-intensities around the left ventricle, that could be due to infiltrating oedema, are not segmented when only providing the T1ce and T2 scans. The network does detect this oedema when the FLAIR sequence is added.
A comparable observation is made in figure 9.6. Based only on the T1ce and T2 MRI, the network is able to segment the tumour core but misses the surrounding oedema which is not clearly visible on the T2 sequence. When adding the FLAIR sequence where the oedema is more evident, the network is able to detect more of the oedema tissue.

Segmentation results on the Ghent University Hospital data displayed in figure 9.7 show that the network also works on data from our private institution. In the GBM cases (top two rows) the different tissues being necrosis, surrounding enhancing tissue and peritumoural oedema appear well delineated. For the LGG cases (bottom two rows), which demonstrate no enhancing tumour tissue, the tumour core and oedema are well identified. Small hyper-intensities in the tumour core on the T1ce scan due to blood vessels are correctly not classified as enhancing tissue by the U-Net.

We believe that the obtained performance with the implemented U-Net is sufficiently high to be useful in a clinical setting. Similar to the BraTS top performing solutions, the segmentation performance could also further be improved through the use of ensembles (e.g. by training multiple networks on different random splits). In this work, we wanted to design a segmentation algorithm that is fully automatic, accurate, fast and generalisable to data from different institutions. It has to be taken into account that the segmentation results of the network are compared with manual segmentations. Manual delineations suffer from inter- and intra-reader variability and are also not 100% accurate. It can therefore be debated whether further improving the Dice scores

with a few percentages is clinically relevant. Objectivity and robustness could be more important when analysing brain tumour volumes and progression over time.

In next chapter, we will use the segmentation network to extract the tumour region of interest from the MRI. This will allow the classification network to focus on the relevant tumour region. We expect that the achieved performance is adequate for this task, as small variations between manual and predicted segmentations won't have a strong influence on the tumour ROI. Moreover, the segmentation network also works on the GUH data which will be used in the following chapter.

9.5   Conclusion

In this chapter, we developed an automatic, accurate and fast segmentation deep learning algorithm based on the U-Net architecture. Segmentation requires only one propagation of the entire brain MRI through the network that produces a segmentation map, accurately delineating necrosis, enhancing tissue and peritumoural oedema. Through channel dropout, i.e. randomly excluding input MRI during training, robustness to missing input modalities could be increased. This is useful a clinical setting where often some sequences are not available or of insufficient quality, which is also the case in our retrospectively acquired dataset that will be used to train and evaluate a diagnosis network in the next chapter. Finally, the segmentation quality was assessed on data from our private institution, demonstrating the generalisability of the segmentation network.

10 | Automatic glioma characterisation

In this chapter, we propose a non-invasive fully automatic 3D pipeline to predict clinically relevant markers according to the most recent WHO glioma classification guidelines based on routinely acquired pre-therapy MRI. We collected a large dataset from multiple public databases and an independent dataset from the Ghent University Hospital to test the generalisation performance.

The work in this chapter has been presented on the Medical Imaging with Deep Learning (MIDL) 2020 conference [334] and published as a part of Milan Decuyper et al. "Automated MRI based pipeline for segmentation and prediction of grade, IDH mutation and 1p19q co-deletion in glioma". In: *Computerized Medical Imaging and Graphics* 88 (Mar. 2021). ISSN: 08956111. DOI: 10.1016/j.compmedimag.2020.101831.

10.1 Introduction

In chapter 8 we have discussed the binary brain tumour grading problem and compared performance of hand-engineered radiomics features with features extracted using a pre-trained CNN. Although best results were achieved with the radiomics features, the performance achieved with the pre-trained CNN features was already very close. The used network was pre-trained on ImageNet, containing natural images, and is therefore not optimised to process medical imaging data. We believe that when training a network specifically on medical images, classification performance could even be improved. Training a convolutional neural network from scratch to predict glioma markers from pre-therapy MRI is the topic of this chapter.

Next to tumour grade, the WHO has put increased emphasis on the integration of molecular markers for brain tumour differentiation (see section 7.1.2). Gene expression profiles have shown to be a better predictor of survival and therapy response compared to histopathology alone. As discussed in section 7.1, IDH mutation and 1p/19q co-deletion status play a crucial role in the classification of diffuse glioma (see figure 7.2). Glioma with IDH mutation demonstrate better prognosis and response to temozolomide chemotherapy. Combined loss of chromosome arms 1p and 19q is also linked to more favourable outcomes and response to PCV chemotherapy. Demonstration of both IDH mutation and 1p/19q co-deletion is required for diagnosis of oligodendroglioma.

We can conclude that determination of WHO grade (glioblastoma versus lower-grade glioma), IDH mutation and 1p/19q co-deletion status is necessary for prognosis and optimal therapy planning of diffuse glioma. Genetic markers are currently derived through tissue analysis after biopsy or resection. As already mentioned, these invasive procedures involve risks and are not always possible to perform depending on location of the tumour and clinical condition of the patient (section 7.2.1). Moreover, genetic testing is expensive and time consuming. New genetic alterations, relevant in oncology, are continuously identified and current oncology workflows are not accustomed to incorporate huge amounts of tests [234, 340]. Therefore, non-invasive assessment of clinically relevant genetic markers can aid in characterising glioma and guide initial therapy and surgery planning, especially when extraction of tumour tissue is not possible or genetic testing not available.

In chapter 8, we used manual segmentation labels to extract radiomics features and define the 3D bounding box around the tumour region of interest, similar to many current studies on computer-aided glioma classification (see table 7.1). To improve objectivity and reproducibility, we will use the automatic segmentation network developed in chapter 8 to extract the tumour ROI.
Additionally, existing work is often trained and evaluated on a small dataset acquired from one clinical centre. To train a CNN from scratch, a lot of data is required and for evaluation preferably an independent dataset from a different institution should be used to test the generalisation performance. Collecting a large and well-curated dataset of brain tumour MRI is challenging. Included cases often have unavailable or low quality input MRI modalities or missing ground truth labels describing tumour characteristics. Initiatives such as The Cancer Imaging Archive

try to resolve these issues by hosting a large archive of publicly available medical image datasets.

In this chapter, we collected a large dataset from multiple collections available on TCIA and an independent dataset at the Ghent University Hospital for evaluation of the generalisation performance. In chapter 9 on segmentation we have seen that robustness to missing modalities can be improved by applying channel dropout during training. We will additionally employ multi-task learning to train a CNN from scratch on the entire collected dataset while dealing with missing labels and restricting potential overfitting. The goal is to develop an accurate and automatic pipeline to segment glioma and subsequently predict WHO grade, IDH mutation and 1p/19q co-deletion status based on pre-therapy MRI.

10.2 Dataset

To acquire a large dataset, we gathered data from multiple databases available on The Cancer Imaging Archive. Additionally, data was collected at the Ghent University Hospital for final evaluation of the generalisation performance.

10.2.1 The Cancer Imaging Archive

The Cancer Imaging Archive hosts a large archive of medical images of cancer that are available for public download [51]. We collected data from multiple collections: The Cancer Genome Atlas [266] glioblastoma [341] and lower-grade glioma [342] collections (TCGA-GBM and TCGA-LGG) and the LGG-1p19qDeletion collection [343]. Furthermore, we also included data form the BraTS 2019 dataset which partially overlaps with patients in TCGA-GBM and TCGA-LGG.

Inclusion criteria were: a histologically proven glioma of WHO grade II, III or IV, the availability of at least a pre-operative T1ce MRI together with a T2 and/or FLAIR sequence of sufficient quality and information on WHO grade, IDH mutation and 1p/19q co-deletion status. In total 628 patients were included: 164 patients from TCGA-GBM, 121 from TCGA-LGG, 141 from 1p19qDeletion and 202 from BraTS 2019 (only patients that were not already included in the TCGA collections). For the cases in BraTS all four MRI sequences (T1, T1ce, T2 and FLAIR)

were available. In the TCGA collections, all four MRI were available for 74% of the patients. A T1 scan was included in 86% of the cases, a T2 sequence in 98% and a FLAIR in 83%. The LGG-1p19qDeletion collection only includes a T1ce and T2 sequence. Hence the required robustness of the segmentation network to lacking T1, T2 and FLAIR MRI.

For all patients, WHO grade information was available (337 GBM vs. 291 LGG). IDH mutation status was known for 380 patients (212 mutated vs. 168 wildtype) and 1p/19q co-deletion status for 280 LGG patients (133 co-deleted vs. 147 intact). The 1p19qDeletion collection included biopsy proven 1p/19q status, determined through fluorescence in-situ hybridisation (FISH). Molecular data of patients in the TCGA-GBM and TCGA-LGG collections were obtained from Ceccarelli et al. [344] where they performed molecular analysis through gene sequencing after tumour resection for the majority of the cases in TCGA.

The collected dataset from TCIA contains data from many different centres with imaging systems from different vendors (both 1.5T and 3T) and differing scanning parameters. This results in a very heterogenous dataset with variability in voxel size, resolution, slice gap, contrast etc.

10.2.2 Ghent University Hospital

Additionally, data was retrospectively acquired at the Ghent University Hospital. Permission was granted by the local ethics committee and informed consent was waived (Belgian registration number B670201838395 2018/1500).

Using the same inclusion criteria as for the TCIA data, we collected data from 110 patients with known WHO grade (61 GBM vs. 49 LGG). For 86 patients IDH status was determined (32 IDH mutant vs. 54 IDH wildtype) through immunohistochemistry and for 40 LGG patients (12 co-deleted vs. 28 intact) 1p/19q co-deletion status was known by fluorescence in-situ hybridisation.

Of these 110 patients, 79 included all four pre-therapy MRI sequences. For 15 patients, the T1 scan was missing. Two cases did not include a FLAIR sequence and for 17 cases the T2 was absent.

10.2.3   Pre-processing

The collected scans are pre-processed in a similar way as the BraTS segmentation data. The different pre-processing step were performed fully automatically using SPM12 (version 7219, Wellcome Trust Centre for Neuroimaging, University College London) [330] and MATLAB R2018b (The MathWorks, Inc., Natick, MA).

All scans are first converted from standard Digital Imaging and Communications in Medicine (DICOM) format to NIfTI format.

Subsequently, the T1, T2 and FLAIR scans were co-registered to the T1ce scan as this sequence was always available and typically had the highest resolution.

After co-registration, the scans are spatially normalised to MNI space, interpolated to 1 mm^3 voxel sizes and bias corrected.

Finally, the MRI are skull-stripped such that only the brain region is preserved. This effectively removes information from the MRI that is not relevant for the tumour diagnosis.

10.3   Architecture and training

The pipeline designed in this study consists of a segmentation stage and a subsequent classification stage as illustrated in figure 10.1. The brain tumours are first delineated using the segmentation network from chapter 9. The obtained segmentation map is then used to extract the tumour region of interest from the MRI. Finally, the obtained tumour ROI is classified using a classification network which is explained in this section.

10.3.1   ResNet

Using the segmentation mask, a tumour region of interest is extracted from the MRI and subsequently fed into the classification network as illustrated in figure 10.1. A similar architecture design is used as described in the original ResNet paper [46]. The network is modified to a 3D CNN and several modifications were applied to reduce complexity (number of parameters) and thereby limit the risk of overfitting.

The architecture starts with a convolutional layer with 64 $7 \times 7 \times 7$ filters and stride of two followed by four residual blocks. Each residual
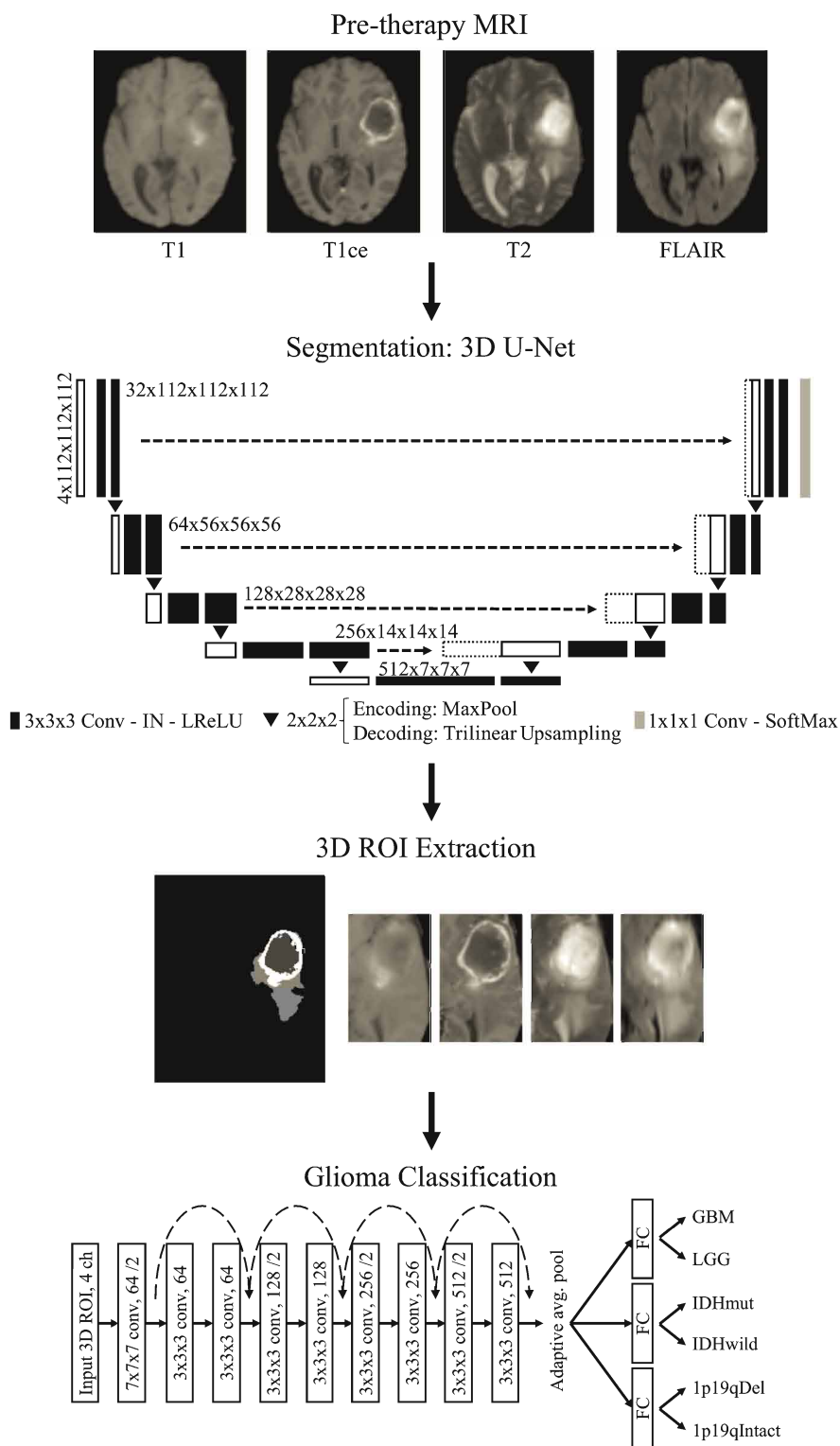
Figure 10.1: Schematic overview of the two-stage pipeline presented in this study. Both the segmentation and classification network architectures are illustrated.

block consists of two convolutional layers with kernels of size $3 \times 3 \times 3$ and a skip connection via addition. The convolutional layers in the first block have 64 filters without downsampling of the feature maps. In the following three residual blocks the number of filters are doubled and downsampling is directly performed by the first convolutional layer that has a stride of two. This results in 512 features maps after the last convolutional layer. To allow identity shortcuts, input and output of a residual block must have the same dimensions. This is not the case in the last three residual blocks where the input is matched to the output dimensions through a $1 \times 1 \times 1$ convolutional layer with stride two. Every convolutional layer is succeeded with instance normalisation and Leaky ReLU activation (negative slope of 0.01). The adaptive average pool layer after the last convolutional layer allows the network to process different ROI input sizes hence no resizing to a fixed shape is required. In the end the network splits into three separate fully connected layers to simultaneously predict WHO grade, IDH mutation and 1p/19q co-deletion.

10.3.2 Training and evaluation

As seen in figure 10.1, the classification network shares the main convolutional part across the three tasks and only splits in the last fully connected layers. This so-called multi-task learning helps the network to learn features that are relevant for multiple tasks, reduces the risk of overfitting and allows a better generalisation (see section 2.3.3). As explained in section 7.2.1, MRI features describing enhancing regions and tumour margins are important to predict grade, IDH and 1p/19q status which shows that these tasks are very much related and that knowledge on one characteristic is informative for the other markers as well. Moreover, not all ground truth labels are available for every patient in our dataset. Multi-task learning allows us to deal with missing labels and train one network on all data instead of training separate networks for each task on a smaller dataset.

The 1p/19q co-deletion classifier (fully-connected layer) is only trained for LGG patients as all GBM patients in the dataset are 1p/19q intact and the 2016 WHO classification system does not include 1p/19q status for GBM cases (see figure 7.2).

The network is trained with AdamW optimisation ($lr_{init} = 10^{-5}$), L2 weight decay of 10^{-2}, a batch size of eight and focal binary cross-

entropy loss. Focal loss weighs the contribution of each sample based on the classification error and thereby reduces the contribution of already correctly classified samples. This is especially useful to deal with class imbalance. The loss is calculated for each task separately on all samples in the batch with known ground truth labels and averaged to a global loss which is backpropagated through the network. If the validation loss did not improve in the last 10 epochs, the learning rate is halved and early stopping occurs after no improvement for 30 epochs. In the last fully connected layer, dropout is applied with a probability of 10% to help reduce potential overfitting. Different hyperparameters of the network were tuned based on the validation set (see below). The network was implemented in Python and the PyTorch deep learning framework and trained on an 11GB NVIDIA GeForce RTX 2080 Ti GPU.

To artificially generate more training samples and further reduce the risk of overfitting, data augmentation was applied during training. The augmentations are visualised in figure 10.2 and include:

- Randomly making the 3D tumour bounding box larger within a range of 10 voxels along each dimension (figure 10.2a). This simulates potential variations in segmented tumour borders and includes more or less surrounding (healthy) tissue

- Random left-right flipping along brain midline (figure 10.2b)

- Random axial rotations with an angle between -10° and 10° (figure 10.2d)

- Random intensity scaling with a factor within a range of 0.8 and 2 (figure 10.2e)

- Elastic transformations which slightly change the shape of the brain (figure 10.2c)

- Randomly setting input channels to zero as was done to train the segmentation network (see section 9.2.3)

These augmentations were randomly applied resulting in a lot of different combinations and accordingly, many additional data samples. Furthermore, they are performed on the MRI and segmentation maps before cropping to the tumour ROI to prevent introduction of boundary effects.

(a) Bounding box size.   (b) Flipping.   (c) Elastic transform.

(d) Axial rotation -10° and 10°.   (e) Intensity scaling.
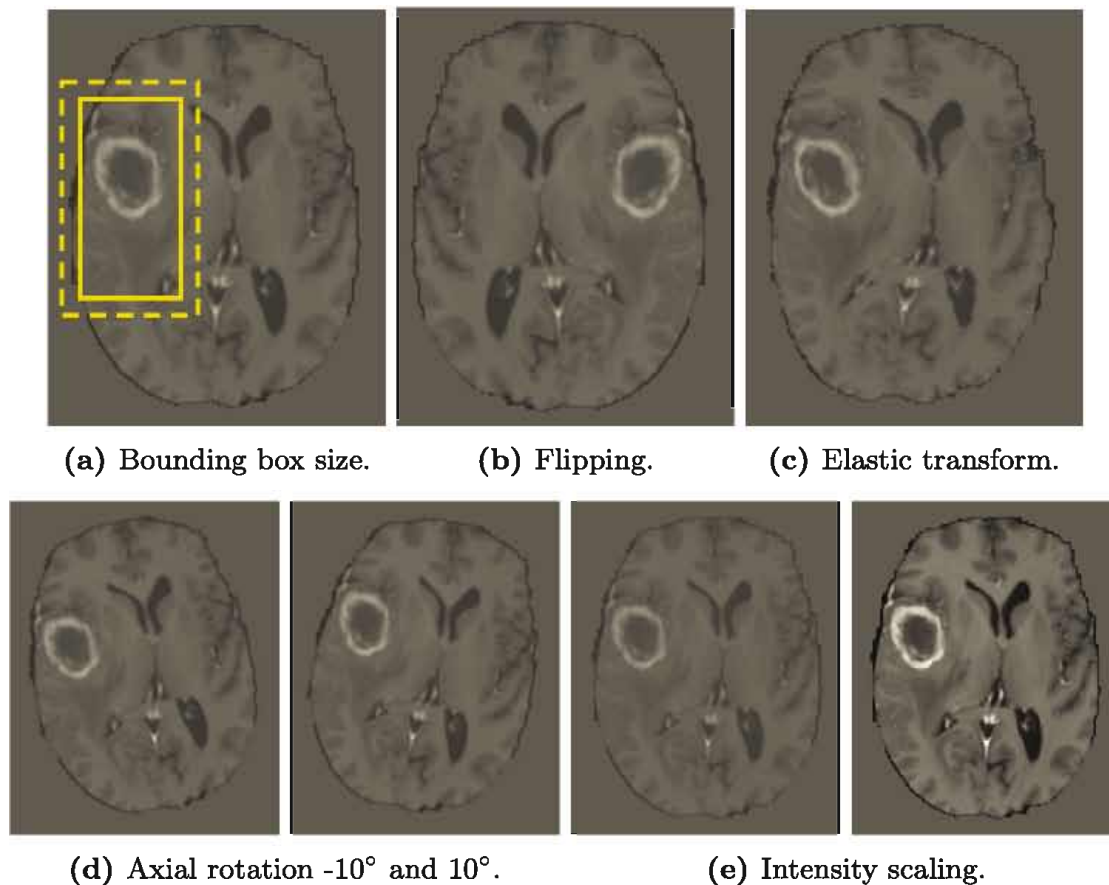
Figure 10.2: Visualisation of different data augmentations applied during training.

The 628 patients are split into a training set of 458 patients (264 GBM vs. 194 LGG, 123 IDH mutant vs. 87 IDH wildtype and 83 1p/19q co-deleted vs. 100 1p/19q intact), a validation set of 70 cases (27 GBM vs. 43 LGG, 41 IDH mutant vs. 29 IDH wildtype and 20 1p/19q co-deleted vs. 23 1p/19q intact) and a test set of 100 (46 GBM vs. 54 LGG, 48 IDH mutant vs. 52 wildtype and 30 1p/19q co-deleted vs. 24 1p/19q intact) patients. For patients in the validation and test set, all ground truth labels were available. Furthermore we made sure that test patients were not used in the training set of the segmentation network (as the BraTS dataset partially overlaps with the TCGA collections) in order to evaluate the system on new cases that both the segmentation and classification stages have never seen before. Data from the Ghent University Hospital was used to evaluate the performance of the classification pipeline on an entirely independent dataset.

10.4   Results

In table 10.1 the results are presented of the multi-task classification network. For each task (WHO grade, IDH mutation and 1p/19q co-deletion status) the AUC, Matthews Correlation Coefficient, accuracy, sensitivity and specificity scores are included. The sensitivity scores indicate the percentage of GBM, IDH mutant and 1p/19q co-deleted cases that are correctly classified as such when using a probability threshold of 0.5. The receiver operating curves are visualised in figure 10.3 for the three tasks and on both evaluation sets.

The results on the unseen TCIA test data show high classification performances with AUC scores of 93% and 94% for grade and IDH status respectively. Predicting 1p/19q co-deletion status for lower-grade glioma is harder but still an AUC of 82% is achieved.
The performance was also evaluated on the completely independent data from the Ghent University Hospital. The resulting AUC scores on the GUH data are 94%, 86% and 87% for grade, IDH and 1p/19q status respectively.

Training the network took around 15 hours and once trained, a tumour ROI could be classified in less than one second on the GPU.

Table 10.1: Classification performance on the TCIA and Ghent University Hospital test data. AUC, Matthews Correlation Coefficient, accuracy, sensitivity and specificity scores are reported for all three tasks: WHO grade, IDH mutation and 1p/19q co-deletion status. A case is classified as Glioblastoma (WHO grade IV), IDH mutant and 1p/19q co-deleted respectively if the predicted probability is higher than 0.5.

Dataset	Task	AUC	MCC	Acc.	Sens.	Spec.
TCIA test data	GBM vs. LGG	93.28	80.26	90.00	93.48	87.04
	IDH mutation	94.03	78.00	89.00	89.58	88.46
	1p/19q co-deletion	82.08	66.16	83.33	86.67	79.17
GUH data	GBM vs. LGG	93.98	79.81	90.00	90.16	89.80
	IDH mutation	86.23	52.92	75.58	84.38	70.37
	1p/19q co-deletion	86.61	40.48	75.00	58.33	82.14

(a) TCIA test data   (b) GUH data
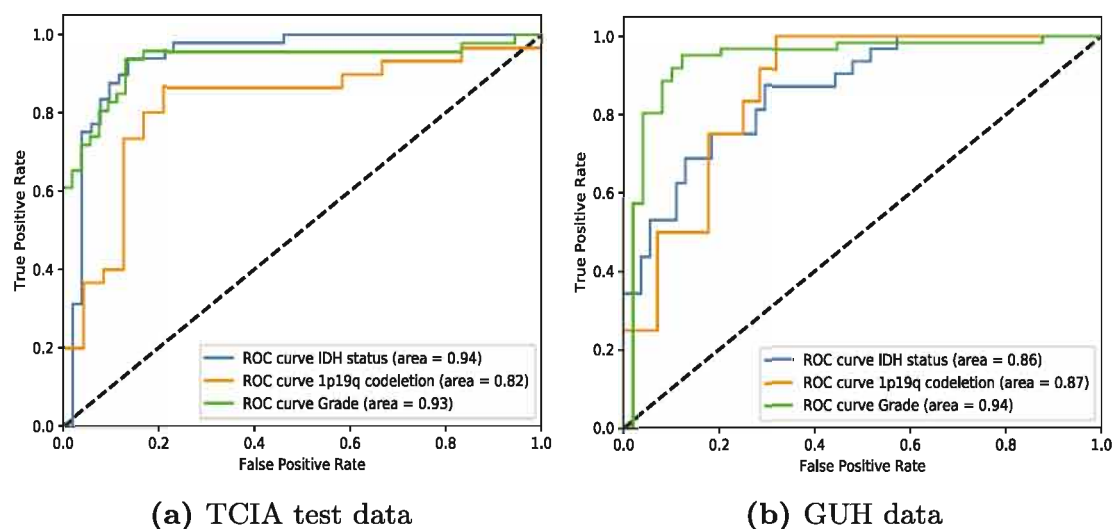
Figure 10.3: Receiver operating characteristic curves for predicting WHO grade, IDH mutation and 1p/19q co-deletion status.

10.4.1   Nearest neighbour visualisation

To probe the network's visual knowledge, we look at the feature activations induced by an input ROI at the last, 512-dimensional hidden layer. For a patient that is classified into a certain class with a high probability, the tumour ROIs of other patients in the dataset were selected that have the smallest Euclidean distance between their feature vectors of length 512, extracted after the average pooling layer. In other words, we visualise the nearest neighbours in feature space. If tumour ROIs produce feature activations within a small Euclidean distance, we can suspect that the higher levels of the network consider these ROIs as similar.

For three types of glioma, a 2D slice of the tumour ROI of a patient in the GUH dataset is shown in figure 10.4 and the four nearest neighbours from the TCIA dataset are presented. The three included types are oligodendroglioma (LGG, IDH mutant and 1p/19q co-deleted, figure 10.4a), astrocytoma (LGG, IDH mutant and 1p/19q intact, figure 10.4b) and glioblastoma (GBM, IDH wildtype and 1p/19q intact, figure 10.4c). Only the T1ce and T2 sequence is shown for the oligodendroglioma case as the nearest neighbours were from the LGG-1p19qDeletion collection which only contains those two sequences.
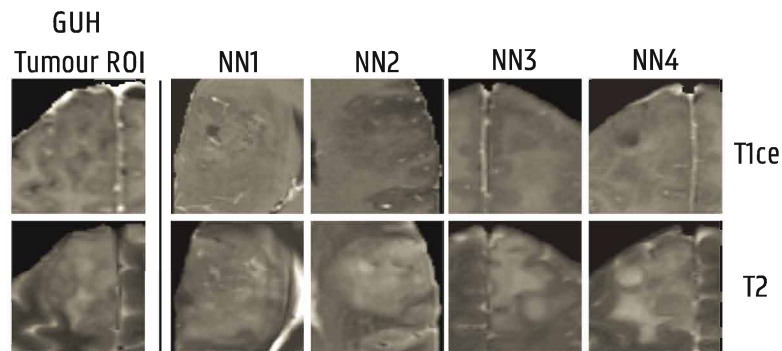
(a) LGG, IDH mutant and 1p/19q co-deleted (oligodendroglioma)
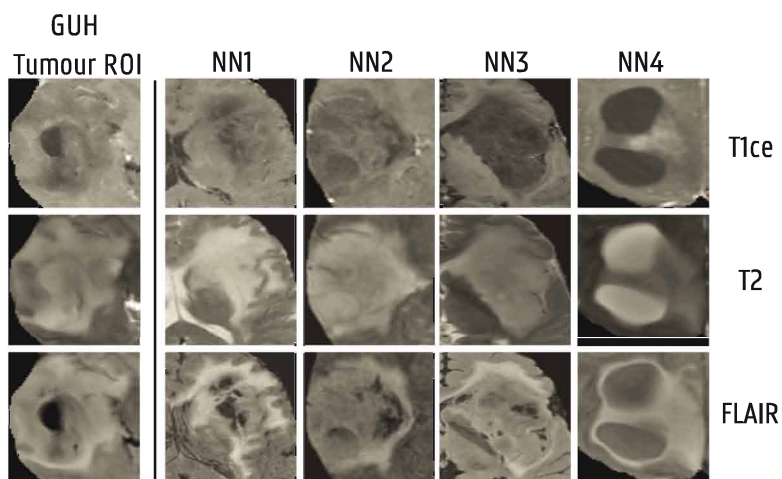
(b) LGG, IDH mutant and 1p/19q intact (astrocytoma)
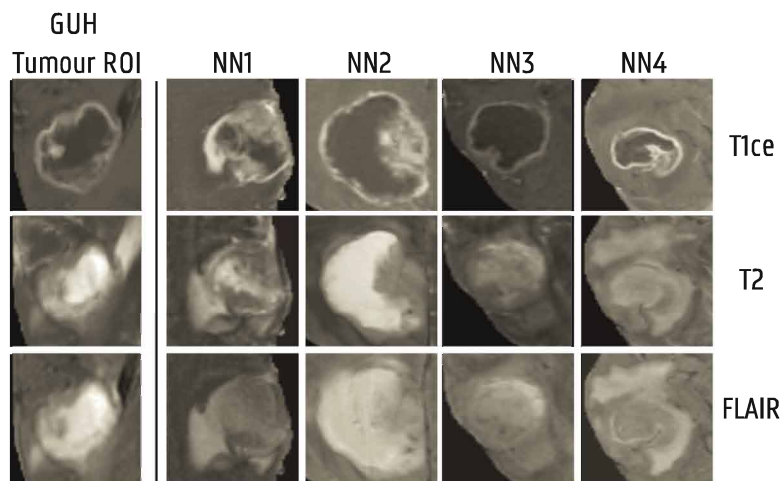
(c) Glioblastoma, IDH wildtype

Figure 10.4: Nearest neighbour visualisation for three cases from the GUH dataset and three types of glioma.

10.5   Discussion

With the segmentation network of chapter 9, a 3D tumour ROI is extracted and used as input to the subsequent 3D CNN that predicts binary tumour grade, IDH mutation and 1p/19q co-deletion status.

For binary grade prediction, very high accuracies of 90% on both the TCIA and GUH test data are achieved. This shows that the network is able to accurately distinguish glioblastoma from lower-grade glioma and generalises well to unseen data from different institutions.

The IDH mutation prediction performance is high on the TCIA test set (AUC of 94%). On the GUH data, the performance is lower with an AUC of 86%. Especially a lower specificity of 70% compared to 88% is observed. This difference in performance might be because immunohisto-chemistry was used to determine IDH status for the GUH data while for the TCGA data IDH status was assessed using gene sequencing [344]. However, a negative IDH status using IHC does not necessarily mean an IDH wildtype tumour and if no sequencing is available the resulting diagnosis suggested by the WHO is astrocytoma, not otherwise specified (NOS) [3]. The GUH database contains 14 IDH wildtype astrocytoma while this diagnosis should be very rare according to the WHO. Hence some IDH mutant astrocytoma might be missed with IHC resulting in more false positives of the model and thus a lower specificity. An additional limitation to the dataset is that there are only two IDH mutant glioblastoma in the TCIA training set making it very unlikely that the network will predict this combination of classes. Therefore, the four GBM with IDH mutation in the GUH database were predicted as GBM, IDH wildtype.

In terms of 1p/19q co-deletion status prediction of lower-grade glioma, a good performance is achieved on both the TCIA and GUH datasets (AUC of 82% and 87% respectively). Although 1p/19q status was known for the GBM cases in the TCGA-GBM collection (all 1p/19q intact), we only included LGG patients as this marker is only considered for those patients according to the WHO guidelines (see figure 7.2). Including the GBM cases would increase the overall prediction accuracy of 1p/19q status but would introduce a large data imbalance and thereby decrease the performance for LGG cases. Results on the GUH dataset show a lower sensitivity compared to the results on the more balanced TCIA test set. In the GUH dataset, 1p/19q status was only available for 40 LGG patients with just 12 1p/19q co-deleted cases which might be too small to

obtain reliable performance estimations. Depending on the classification threshold, the sensitivity can also be optimised. For example, with a threshold of 0.45 the sensitivity on the GUH dataset increased to 75% with the same specificity.

In figure 10.4, we visualised the nearest neighbours in feature space for several test patients from the GUH dataset. For the LGG, 1p/19q co-deleted case illustrated in figure 10.4a we see common characteristics between the nearest neighbours such as ill-defined tumour borders and heterogeneous signal intensity on the T2 sequence. This corresponds to imaging features found in existing literature [7, 8].
A very interesting result is the LGG, IDH mutant and 1p/19q intact case shown in figure 10.4b. In all nearest neighbours, a shared imaging feature is observed. Certain tumour regions show a hyperintense signal on T2 but a hypointense signal on FLAIR. Indeed, this T2-FLAIR mismatch is reported as a highly specific imaging marker for non-enhancing LGG, IDH mutant astrocytoma [345, 346]. Similarly, in figure 10.4c, thick, irregular enhancement with necrosis are common characteristics in the found nearest neighbours which is attributed to GBM, IDH wildtype glioma [6].
Although this is not a statistically sound proof that the network has learned to recognise these features, it can give a first hint to the network's visual knowledge. Further research is still necessary to explain and interpret the network.

In this work, we trained a 3D classification network to make a prediction based on the entire tumour ROI. Current applications of CNNs for brain tumour classification are mostly 2D, taking only a small part of the tumour into account while brain tumours have a very heterogeneous appearance with strong variations between different slices. Furthermore, extracting only the tumour ROI allows the network to focus on this region. It was shown in chapter 8 that this improved the classification performance with a pre-trained CNN. On the other hand, context information on surrounding tissues and location is excluded. Including this information in the input may further improve diagnostic performance when training a CNN from scratch.

The clinical translation potential of the developed pipeline is strong as only routinely acquired MRI are necessary as input and no further human interaction is required. Data pre-processing is minimal, and the entire segmentation and classification pipeline takes less than 5 seconds

on an NVIDIA GeForce RTX 2080 Ti GPU.

10.6 Conclusion

In conclusion, we developed a fully automatic 3D pipeline to segment
glioma and non-invasively predict important (molecular) markers ac-
cording to the WHO classification guidelines with high diagnostic perfor-
mance. Through the use of multi-task learning to handle missing labels,
one classification network could be trained on a large multi-institutional
database. Evaluation on an independent private dataset demonstrated
the generalisability of the algorithm. The non-invasive assessment of
clinically relevant genetic mutations can help to characterise glioma and
thereby guide initial therapy and surgery planning.

11 | Combined segmentation and classification: Y-Net

In this chapter, a network is presented, called Y-Net, that is able to process full brain MRI and concurrently perform glioma segmentation and prediction of important tumour markers. This approach does not require prior segmentation before classification and tumour markers can be predicted based on a complete view of the brain. Different techniques are be presented that allow to train a complex 3D network with limited GPU memory and on a heterogenous dataset with missing labels. Segmentation and classification performance is evaluated on the same datasets as in chapters 9 and 10 and visualisation techniques are employed to interpret the trained network and examine the learned imaging features.

11.1 Introduction

In chapter 10 we employed a two-stage pipeline for automatic glioma segmentation and prediction of WHO grade, IDH status and 1p/19q co-deletion status. Although high segmentation and classification performances are achieved, there are several potential downsides to this approach.

Extraction of the tumour region of interest before the classification stage allows the network to focus on this region but could also remove potentially relevant information on surrounding (healthy) tissues and

tumour location. Several studies have shown significant correlation between IDH mutation and 1p/19q co-deletion status and tumour location [6, 7]. IDH mutant tumours are mostly located in a single lobe such as frontal lobe, temporal lobe or cerebellum. IDH wildtype tumours, on the other hand, often occur in combined lobes such as diencephalon and brain stem. Tumours that are 1p/19q co-deleted most commonly occur in the frontal lobe. It could therefore be beneficial to train a network that predicts tumour markers based on a complete view of the brain. Training such a network is however challenging in terms of memory requirements and potential overfitting to non-relevant information contained in the entire MRI.

Furthermore, possible errors in the segmentation stage could propagate and influence the performance of the classification network. Small regions that are detected as tumour tissue, far from the actual tumour, can result in an ROI that is too large. Conversely when parts of the tumour such as oedema tissue is not delineated, this can lead to an ROI that is much smaller than the true tumour ROI and an incomplete view of the tumour appearance.

In this chapter, we investigate whether it is possible to train one convolutional neural network for simultaneous glioma classification and segmentation of the different tumour tissues. We extend the U-Net architecture with additional classification layers and call this architecture Y-Net. The goal is that Y-Net receives the full pre-therapy brain MRI as input and concurrently produces an output segmentation map and predicts Grade, IDH mutation and 1p/19q co-deletion status.

This approach has the advantage that the entire MRI and therefore all available information is provided as input. No prior segmentation step is required which could influence the subsequent classification.
Moreover, training a network to simultaneously segment and classify glioma can also improve the performance as this takes the multi-task learning strategy applied in chapter 10 even further. Features learned by the U-Net to segment enhancing, necrosis and oedema tissue are relevant for prediction of the different tumour markers as well. Consequently, adding the segmentation task can help the classification part to focus on the relevant tumour regions and act as a regulariser.

Training this network on the brain tumour dataset of chapter 10 is challenging due to missing labels and GPU memory constraints.
Ground truth segmentation labels are only available for patients that are

part of the BraTS dataset. At the same time, many cases in BraTS and other used public collections do not include information on IDH mutation and 1p/19q co-deletion status. Consequently, there are only few patients for which all labels are available. The segmentation network of chapter 9 could be trained with a maximum batch size of two on an 11GB RAM GPU. A batch size of two is too small when training with missing labels as in many training updates, segmentation and classification labels would only be available for one or no patients in the batch. This can lead to very erratic training behaviour. To enable efficient training of the network, labels are necessary for every task of at least a few patients in most training updates.

Furthermore, for the network to learn features describing tumour location and surrounding tissues with respect to classification, this information has to be included during training. In chapter 9, the U-Net is trained on patches, randomly extracted from the brain MRI. Random patch extraction allows to generate many different training samples from the MRI of one patient and limits the risk of overfitting. When only training the network on small patches, however, information on tumour location could be lost. The network will not learn to extract this information even though the entire MRI is provided at test time. Additionally, when extracting random patches, it is possible that this patch does not include any tumour tissue or only a small fragment of the tumour. In that case, the classification part cannot be trained as this would confuse the network. Hence patches need to be extracted that contain enough tumour to perform classification which requires knowledge on tumour location for every patient. Training the network on the entire brain region, on the other hand, further increases memory requirements. Moreover, training without random patch extraction significantly reduces the amount of training samples which can lead to overfitting. Hence for optimal performance, the network should be trained on patches that are sufficiently large to contain all relevant information and enough tumour while being not too large to allow enough differentiation between training samples of one patient.

In section 11.3, the Y-Net architecture and training procedure will be explained with the different techniques that are used to deal with the above challenges. Performance will be evaluated on the same TCIA and UZ test sets as in chapters 9 and 10 and compared with previous results. Finally, we investigate several visualisation techniques that are used to understand and interpret convolutional neural networks (see

section 11.4).

11.2   Data

The same dataset is used as in chapter 10, extended with an additional collection from TCIA: the Ivy Glioblastoma Atlas Project (GAP) collection [51, 347, 348]. This collection contains 35 glioblastoma cases for which pre-therapy MRI of sufficient quality were available. For one patient, IDH status was not known, three GBM cases are IDH mutant and the remaining 31 glioblastomas are IDH wildtype. IDH status was assessed through gene sequencing.

Furthermore, we loosen the requirement of including at least a T2 or FLAIR sequence. Hence, also patients are included for which only T1ce sequence was available. Although a T2 sequence contains highly relevant information for segmentation and classification and performance will be optimal when including those sequences, we also aim to maximise accuracy based on T1ce alone to improve clinical applicability. All data is pre-processed in the same way as explained in section 10.2.3.

This results in a total dataset collected from TCIA of 701 glioma patients. The dataset is split in the same way as in chapter 10 and all additional patients are added to the training dataset. Consequently, the following split is used: a training set of 531 patients (327 GBM vs. 204 LGG, 136 IDH mutant vs. 143 IDH wildtype and 84 1p/19q co-deleted vs. 109 1p/19q intact), a validation set of 70 cases (27 GBM vs. 43 LGG, 41 IDH mutant vs. 29 IDH wildtype and 20 1p/19q co-deleted vs. 23 1p/19q intact) and a test set of 100 patients (46 GBM vs. 54 LGG, 48 IDH mutant vs. 52 wildtype and 30 1p/19q co-deleted vs. 24 1p/19q intact). The same data from the Ghent University Hospital is used for final evaluation of the classification performance on an independent dataset (see section 10.2.2). The segmentation accuracy is evaluated on the BraTS 2019 validation data through the online evaluation platform as in chapter 9.

11.3   Architecture and training

11.3.1 Y-Net architecture

The Y-Net architecture used in this chapter is an extension of the U-Net from section 9.2.2 with classification layers after the last (middle) encoding stage. Figure 11.1 illustrates the architecture with sizes shown for an input patch of $128 \times 128 \times 128$. We only discuss the differences with the architecture in section 9.2.2 here.

The number of feature maps at the highest resolution is reduced to 24 to limit GPU memory usage. Hence the maximum amount of feature maps is equal to 384. In the last encoding stage, the feature maps are reduced to 192 by the second convolutional layer before upsampling. The network then splits into the segmentation (decoding) and classification part. For classification, the 192 feature maps are average pooled with an adaptive average pooling layer and fed to three separate fully connected layers to predict WHO grade, IDH status and 1p/19q co-deletion status.
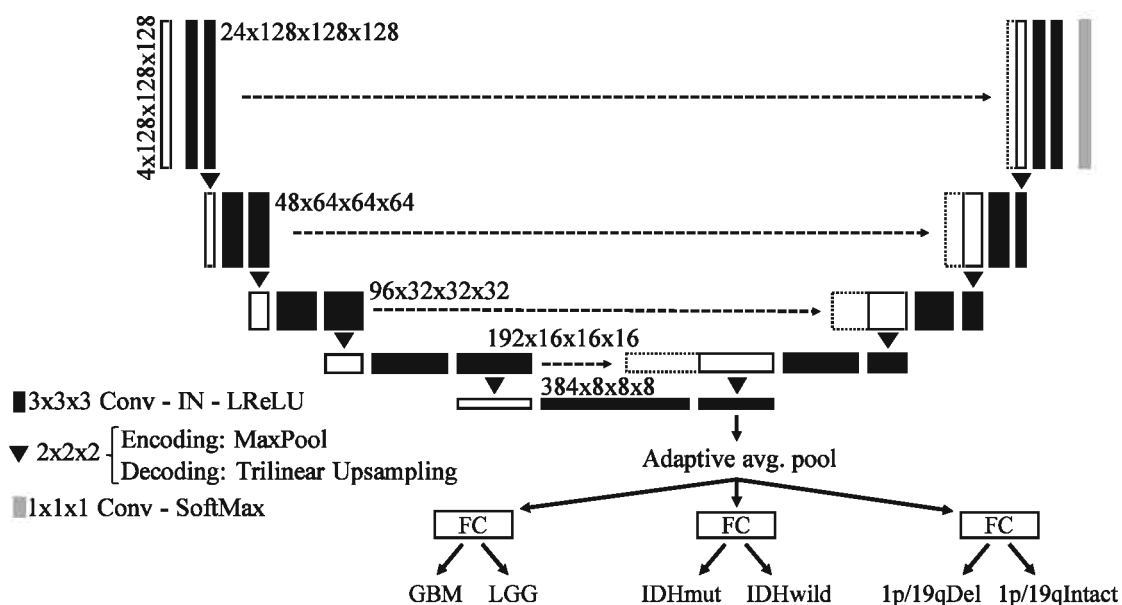
Figure 11.1: Schematic overview of the Y-Net architecture for simultaneous glioma segmentation and classification.

11.3.2 Training procedure

The network is trained with the AdamW optimisation algorithm and L2 weight decay set to 10^{-2}. The initial learning rate is set to 0.00005 for the U-Net part (encoding end decoding branches) of the network. For the fully connected classification layers, the initial learning rate is

set to 0.00001. During initial training runs, it was noticed that the classification tasks converged and started to overfit sooner than the segmentation task. For this reason, the learning rate of the classification layers is reduced. A batch size of six samples is used. Larger batch sizes resulted in out-of-memory errors even after application of mixed precision training and gradient accumulation (see below). Losses are calculated for each sample and available label in the batch. The segmentation loss includes soft Dice and cross-entropy loss. The classification tasks are trained using focal cross-entropy loss. The average loss is calculated over all tasks with available ground truth labels and back-propagated through the network. Learning rate is halved in case of no improvement in overall validation loss for 50 epochs. After no improvement in the last 200 epochs, training is stopped and the optimal epoch is chosen based on the validation loss. One epoch is defined as an iteration over all patients in the training set. In the fully connected classification layer, dropout was applied with a 20% probability of dropping units.

The different hyperparameters of the network and training procedure were tuned based on performance on the validation set. The network was implemented in Python and the PyTorch deep learning framework and trained on an 11GB NVIDIA GeForce RTX 2080 Ti GPU.

In what follows, different techniques are explained that are employed to deal with the challenges mentioned in the introduction of this chapter.

Automatic mixed precision training

Traditionally, deep learning uses single-precision (floating-point 32, FP32) to represent parameters. With the growing complexity of neural networks, there has been an interest in reduced precision or FP16 training. This allows to significantly increase training speed and reduce memory consumption. However, the range of values that can be represented with FP16 is smaller and precision decreases for small numbers. This can result in a lower accuracy of the model. In 2017, researchers at NVIDIA introduced a methodology to train networks using half-precision without losing accuracy, called mixed precision training [349]. They maintain a single-precision copy of the network weights which are converted to FP16 during each training iteration. Forward and backward passes are computed in half-precision and the resulting gradients are first converted back to FP32 precision before multiplication with the learning rate and updating the weights. More details and additional methods that are

applied to improve mixed precision training can be found in Narang et al. [349]. The automatic mixed precision training methodology has been implemented in the PyTorch framework[1] which allows easy conversion of existing models and code to mixed precision.

Gradient accumulation

An additional technique that is used to reduce memory requirements during training is gradient accumulation. Instead of forward propagating all samples in the batch, calculating the average loss and performing back-propagation over all samples, each sample can be forward and back-propagated one at a time. The gradients are accumulated during each back-propagation step and the weights are only updated after all samples in the batch have been processed. As the gradients do not have to be calculated for every sample at once, this significantly reduces memory consumption.

Patch extraction

During training, patches are randomly extracted from the brain volume to increase the number of training samples and thereby limit overfitting. As discussed in the introduction, patches should be large enough to contain enough context information and tumour volume to make a prediction on the different tumour markers. To this end, a training strategy is devised where either the entire brain volume or a random patch is propagated through the network.

The patch dimensions can vary between a size of 128 or 144 voxels, independently for each axis. This introduces additional variability between patches. Taking into account that the average size of the brain region is 136 × 169 × 138, larger patch sizes than 144 are not included. A minimal patch dimension of 128 is chosen to include a large part of the brain volume and enough context information on surrounding healthy tissue and tumour location. In initial training runs, experiments were performed with smaller patches as well and a dimension of at least 128 proved to be optimal. During patch extraction it is verified that the patch contained at least 50% of the total tumour volume. This was assessed through the use of segmentation labels obtained with the segmentation

---

[1] https://pytorch.org/docs/stable/amp.html

network of chapter 9.

With a probability of 20%, no patch is extracted and the entire brain volume is used as a training sample. This way, the network also sees full brain volumes during training.

Data augmentation

To further increase the amount of training samples, data augmentations are applied similar to the augmentations in section 10.3.2. Included augmentations are:

- Random flipping along each axis

- Random axial, sagittal and coronal 360° rotations

- Contrast augmentation

- Elastic transformations

- Channel dropout while always maintaining the T1ce sequence

Each of these augmentations are randomly applied with a probability of 30% resulting in many different combinations of augmentations.

11.4   Interpretation

Several visualisation techniques will be explored to investigate what the network has learned. Deep learning models are often seen as a block box where it is difficult to understand how and why they make certain predictions. We will therefore plot the features that are fed to the classification layers after feature reduction to two dimensions. Furthermore, we will try to visualise which pixels in a certain input MRI have the most influence on the output of the network. Finally, gradient ascent is used to synthesise an input sample which the network strongly attributes to a certain glioma type.

11.4.1   t-SNE visualisation

To further validate whether the features learned by the model are meaningful with respect to the different glioma markers, the features vectors of length 192, extracted after the last encoding stage, can be visualised after dimensionality reduction. The feature vectors are reduced to 2 dimensions for visualisation through t-Distributed Stochastic Neighbour Embedding (t-SNE) [350] using the Python Scikit-learn package [351]. By plotting the obtained feature embeddings as a scatter plot, different patterns and groups can be identified and compared with the ground truth class labels. This way, we can verify whether the different structures and groups in the scatter plot correspond with different tumour subtypes and thus whether the extracted features are meaningful with respect to the different glioma markers. Moreover, by visualising the MRI of tumours positioned close to each other, we can discover common characteristics that can be associated with the corresponding tumour markers.

11.4.2   Saliency maps

A saliency map visualises the influence of every input pixel to the output that is produced by the network [45]. The influence of an input pixel to the output can be described using gradients. An input image is forward propagated through the network and the output logit with the largest value is identified. The gradient is then calculated of this dominant logit with respect to the input image pixels through back-propagation. The saliency map visualises the absolute value of these gradients to highlight which input pixels had the strongest influence. This method is also called gradient visualisation.
Advantages of this technique are the easy calculation and detailed information on what input regions the network focused on. A downside is that these saliency maps can be visually noisy. Therefore, often median filtering is applied.
Other techniques to visualise saliency maps exist as well such as Grad-CAM [352] and guided backprop [353]. A recent study by Adebayo et al. [354] compared different techniques for generating saliency maps and subjected them to sanity checks. They investigated whether randomising the model parameters or data labels influences the obtained saliency maps. Of the tested techniques, gradient visualisation was one of the

few techniques to pass all tests which is why we have opted to use this technique.

11.4.3   Gradient ascent

The gradients of the output with respect to the input can also be used to adapt the input image. Starting from an image containing random noise, this image can be adapted such that the output score of a certain class is maximised. Hence a synthetic image is generated that is representative for the chosen class according to the network [45]. More formally, we start with input image containing randomly generated noise, which is forward propagated through the trained network. The output logit is then back-propagated and the gradients are calculated with respect to the input pixels. Similar to the training procedure a of network, the obtained gradients are now used to update the input pixels while the network weights are fixed. This process is repeated for many iterations until the output class score reaches a certain stable maximum.

For optimisation of the image, the AdamW algorithm is used with a learning rate of 0.1 and L2 weight decay (now of the image pixels) set to 0.01. At each iteration the input image is normalised to zero mean and unit variance. To further regularise the image generation process, we start with a small input image with a size of 32 × 32 × 32 pixels and 4 channels. The size is increased every 50 updates with a scale of 1.2 until an image size of 128 × 128 × 128 pixels is reached. Furthermore, the image is Gaussian blurred every 5 iterations to reduce high frequency patterns. Using this procedure, we will try to synthesise input images that the network attributes to certain glioma types such as glioblastoma, IDH wildtype or lower-grade glioma, IDH mutant and 1p/9q co-deleted.

11.5   Results

11.5.1   Segmentation

Segmentation performance of Y-Net on the BraTS 2019 validation data is summarised in table 11.1. Mean and median Dice scores and robust Hausdorff distance are reported for the different brain tumour regions: enhancing tumour, whole tumour and tumour core. Similar to table 9.1,

the results are also included when a T1 and T2 or FLAIR sequence is not available.

A mean whole tumour Dice score is achieved of 89% based on all four sequences. When only the T1ce and FLAIR or T1ce and T2 scans are provided, the mean Dice scores lowers to 88% and 85% respectively. The obtained median WT Dice scores are 92%, 91% and 88%. The median ET Dice scores are around 86% and TC Dice scores around 89-90%. The segmentations are highly specific (WT specificity of 99%) with a sensitivity of 90% when all sequences are provided. The lowest, 25th percentile, 75th percentile and best WT Dice scores are 42%, 88%, 94% and 98% respectively and are visualised in figure 11.2.

Table 11.1: Segmentation results on the BraTS 2019 validation data. Metrics were computed by the online evaluation platform.

	Available modalities	Dice Score (%)			Hausdorff distance (mm)		
		ET	WT	TC	ET	WT	TC
Mean	T1, T1ce, T2, FLAIR	75.30	89.23	84.10	3.66	6.07	6.51
	T1ce, FLAIR	72.41	88.17	82.95	4.34	6.57	6.98
	T1ce, T2	74.83	85.08	83.23	4.55	8.03	7.86
Median	T1, T1ce, T2, FLAIR	85.61	91.73	90.19	2.00	3.32	3.16
	T1ce, FLAIR	85.53	91.31	89.36	2.12	3.32	3.16
	T1ce, T2	85.94	87.62	89.87	2.00	5.10	2.83

11.5.2 Classification

Table 11.2 presents the classification performance of the Y-Net. In figure 11.3, the receiver operating characteristic curves are illustrated for both the TCIA test set as the GUH test data. A binary threshold probability of 0.5 is used to calculate the MCC, accuracy, sensitivity and specificity scores.

On the TCIA test data, high performances are reported with AUC scores of 98%, 96% and 87% for WHO grade, IDH mutation and 1p/19q co-deletion prediction respectively.
An additional evaluation on the independent dataset acquired at the Ghent University Hospital illustrates the generalisation performance of the network. WHO grade can be predicted with an AUC of 96%, IDH mutation with an AUC of 83% and 1p/19q co-deletion with 90% AUC.

Training the Y-Net network took multiple days, where a lot of CPU

T1ce   **FLAIR**   **Segmentation**
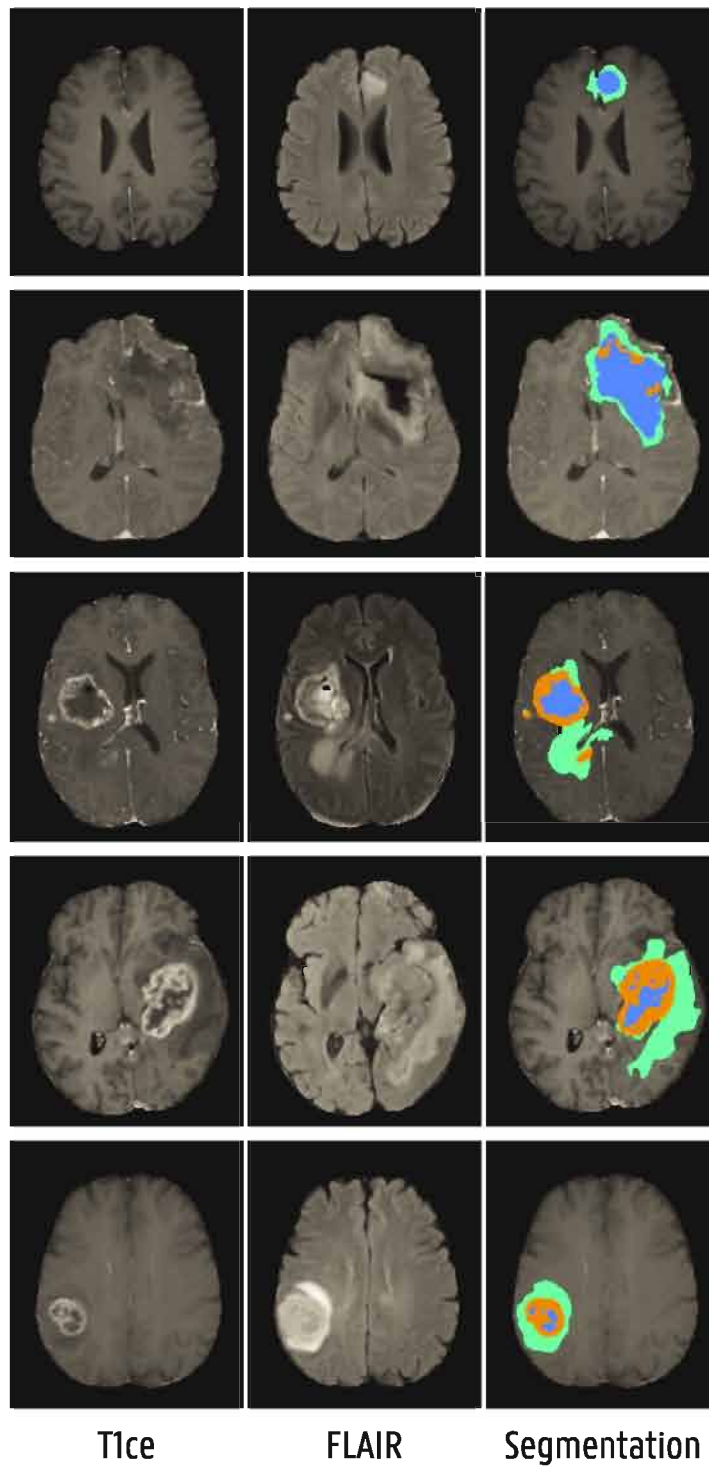
Figure 11.2: Example MRI and segmentations (overlaid on T1ce) of the patients with worst (top), 25th percentile, median, 75th percentile and best (bottom) WT Dice scores. Segmentations are obtained when providing all four MRI sequences.

computation time was spent on on-the-fly data pre-processing and augmentation. Applying random augmentations on-the-fly significantly increases the computation time, but allows to produce the largest variety in training samples. Once trained, however, propagating a patient's MRI through the network only takes a few seconds.

Table 11.2: Classification performance on the TCIA and Ghent University Hospital (GUH) test data. AUC, Matthews Correlation Coefficient (MCC), accuracy, sensitivity and specificity scores are reported for all three tasks: WHO grade, IDH mutation and 1p/19q co-deletion status. A case is classified as Glioblastoma (WHO grade IV), IDH mutant and 1p/19q co-deleted respectively if the predicted probability is higher than 0.5.

Dataset	Task	AUC	MCC	Acc.	Sens.	Spec.
TCIA test data	GBM vs. LGG	98.23	86.45	93.00	97.83	88.89
	IDH mutation	96.23	80.06	90.00	91.67	88.46
	1p/19q co-deletion	86.94	62.50	81.48	83.33	79.17
GUH data	GBM vs. LGG	96.19	83.95	91.82	88.52	95.92
	IDH mutation	83.08	56.17	76.74	87.88	69.81
	1p/19q co-deletion	89.88	71.05	87.50	83.33	89.29

(a) TCIA test data   (b) GUH data
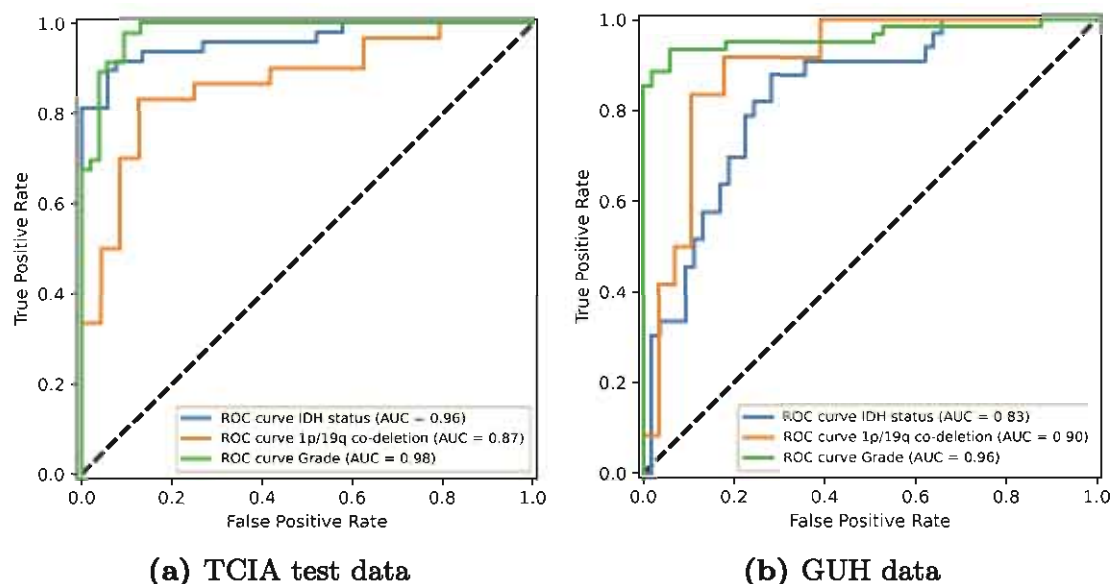
Figure 11.3: Receiver operating characteristic curves for predicting WHO grade, IDH mutation and 1p/19q co-deletion status.

11.5.3   Interpretation

Several visualisation techniques are now investigated to interpret what the classification part of the Y-Net network has learned.

t-SNE visualisation

All data samples from the entire dataset (public data from TCIA and data from GUH) are propagated through the network and the features are extracted that are fed to the final classification layers. These features vectors, with a length of 192, are reduced to 2 feature values with t-SNE and visualised with a scatter plot, colour labelled according to the ground truth labels. Three scatter plots are produced, one for each glioma marker, and presented in figures 11.4 to 11.6. In every plot, examples are included of MRI tumour slices at several locations in the scatter plot to visualise associated features with different clusters.

Figure 11.4 shows the scatter plot for WHO grade. One can clearly identify two clusters (left - right) that strongly correspond with the grade labels GBM (left, red) and LGG (right, green). Lesions in the left cluster show clear enhancing tumour tissue. Two tumours are displayed which are diagnosed as LGG but attributed to the GBM cluster. These examples also present clear enhancing tumour tissue. The right, LGG cluster contains tumours that do not demonstrate enhancing tissue. This is also the case for the two expanded GBM cases that are in the LGG cluster.

For IDH mutation, the clusters are very similar as for WHO grade as shown in figure 11.5. Most of the GBM cases are IDH wildtype and contrast enhancing tissue is an imaging feature for IDH wildtype as well. Within the LGG cluster, one can observe that most LGG, IDH wildtype tumours are positioned towards the left. The expanded examples also demonstrate more enhancing tumour tissue.

In figure 11.6, the scatter points are colour labelled according to 1p/19q co-deletion status for LGG. Two groups can be identified (top - bottom) within the LGG cluster. The cases positioned at the top are mostly 1p/19q co-deleted, whereas the majority of the lesions at the bottom are LGG, 1p/19q intact. Tumour examples in the top cluster are smaller and located in the frontal lobe. At the bottom, the lesions are generally much larger. At the bottom right many tumours demonstrate T2-FLAIR mismatch (see also the example in figure 11.4 in the same region).
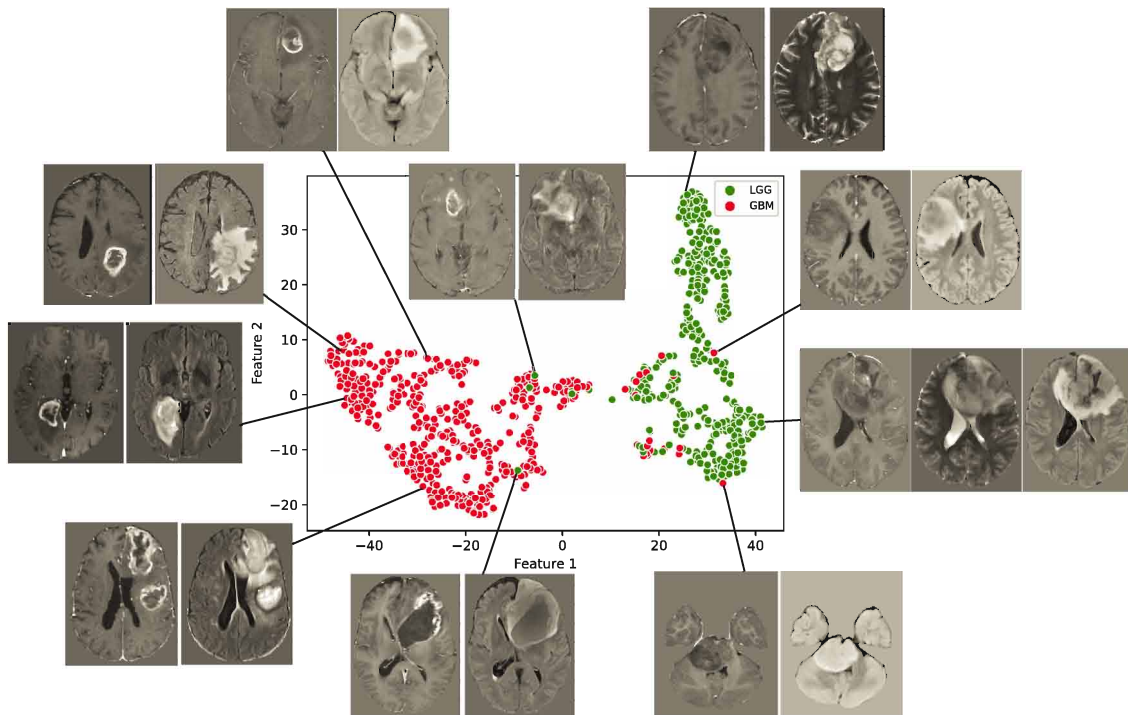
Figure 11.4: Visualisation of feature embeddings using t-SNE with colour labels indicating WHO grade: glioblastoma (red) and lower-grade glioma (green).
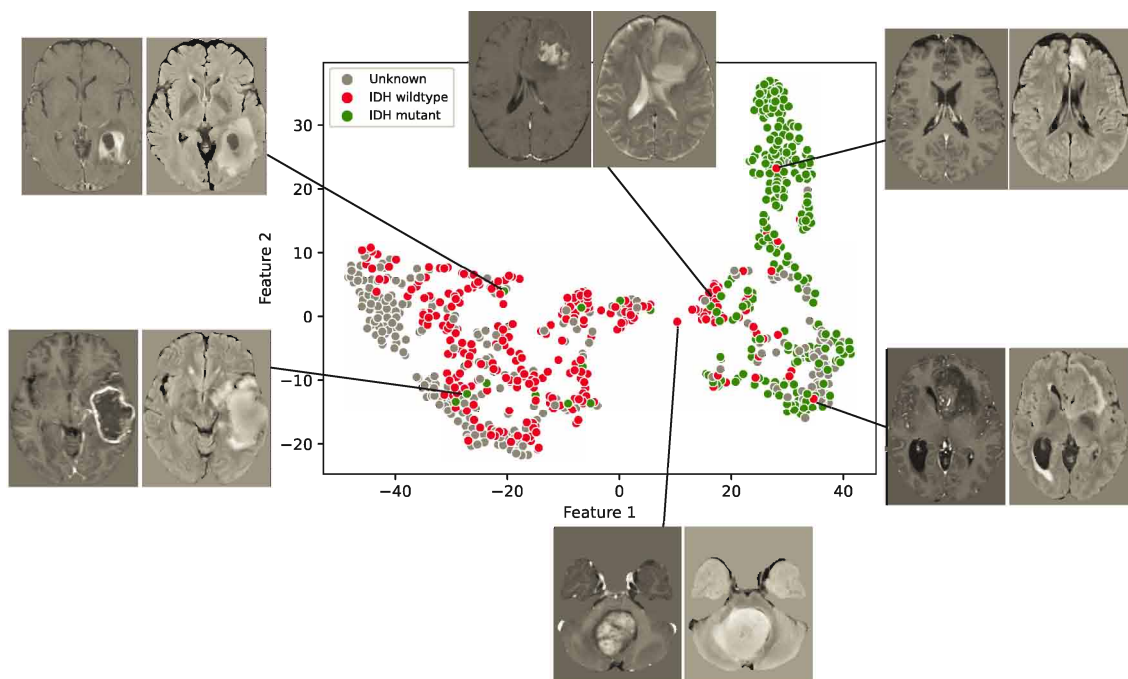
Figure 11.5: Visualisation of feature embeddings using t-SNE with colour labels indicating IDH status: wildtype (red) and mutated (green).
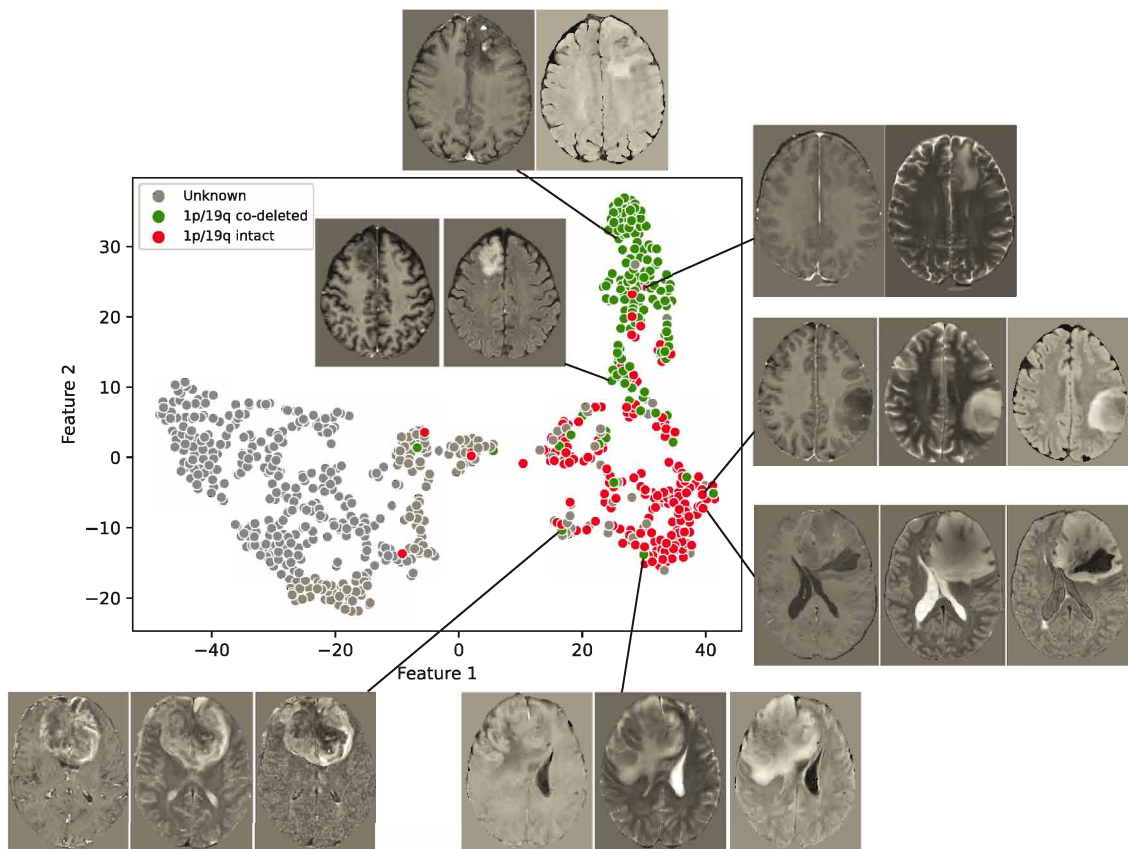
Figure 11.6: Visualisation of feature embeddings using t-SNE with colour labels indicating 1p/19q co-deletion status: intact (red) and deleted (green).

Saliency Maps

Figures 11.7 to 11.9 shows example brain tumour cases from the TCIA test set en GUH data that are correctly classified by the network. Saliency maps are included and overlaid on the T1ce sequence to visualise where the network focused attention to determine the predicted classes. To calculate these saliency maps, the brain tumour MRI are forward propagated through the network. Next, the maximum output logits are identified for each classification task and back-propagated. This way the gradients are computed with respect to the input MRI. The absolute value and maximum over all four sequences is average filtered (with kernel of $3 \times 3 \times 3$) to reduce noise and visualised as a heatmap on the T1ce scan.

In figure 11.7, examples are shown of glioblastoma, IDH wildtype tumours. One can see that the network mainly focuses on the enhancing tumour core region.

<div align="center">T1ce   FLAIR   Saliency</div>
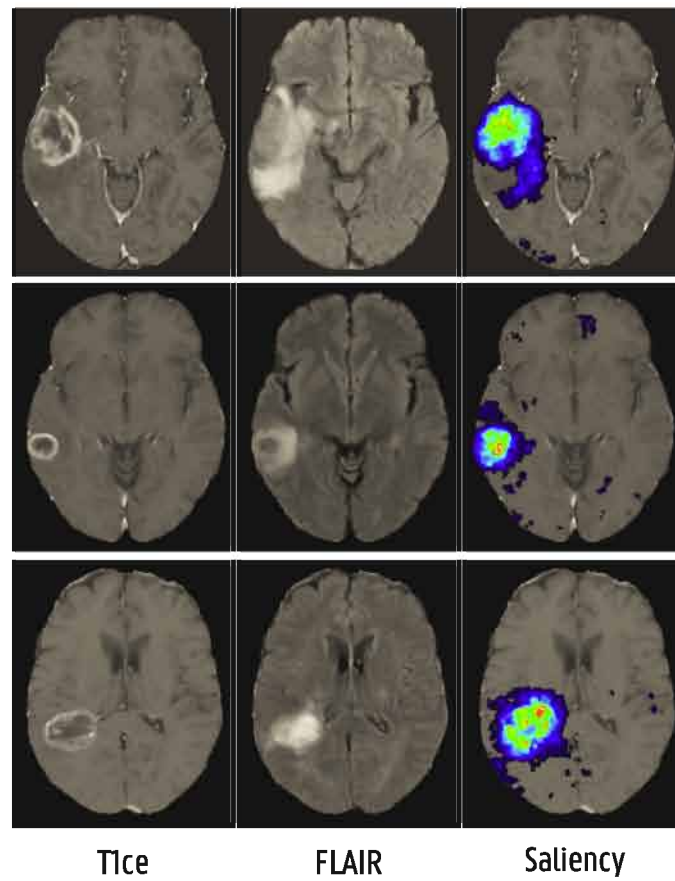
Figure 11.7: Example MRI and saliency maps (overlaid on T1ce) of correctly classified glioblastoma, IDH wildtype tumours.

Examples of LGG, IDH mutated and 1p/19q intact cases are depicted in figure 11.8. Again, the most relevant brain tumour regions are highlighted by the network. For the middle case, some attention is also put slightly below the tumour. The tumours also demonstrate T2-FLAIR intensity mismatch in regions that overlap with main hotspots in the saliency map, although it is difficult to precisely identify the exact tumour regions with the noisy saliency maps.
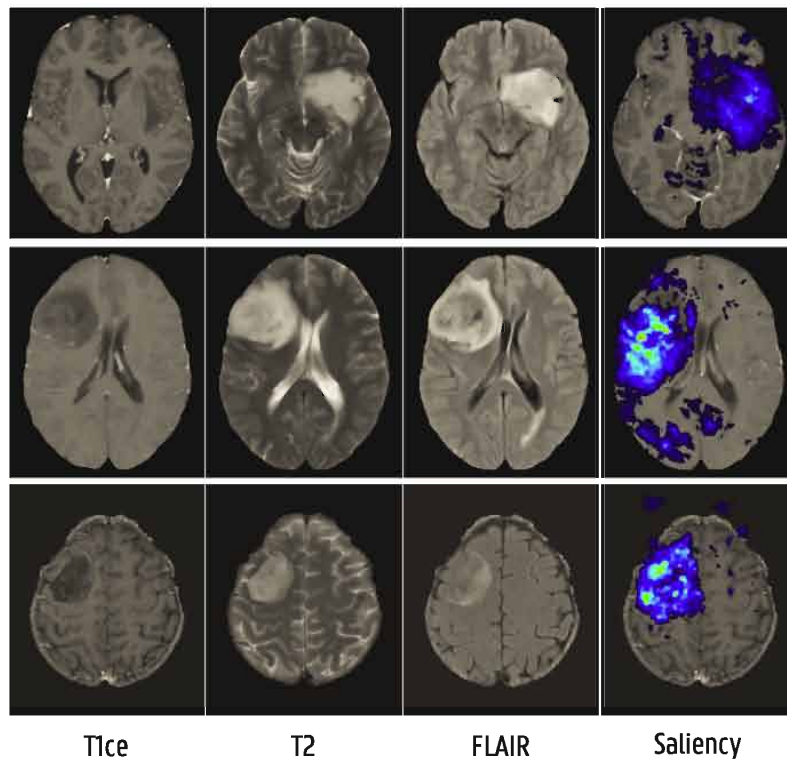
Figure 11.8: Example MRI and saliency maps (overlaid on T1ce) of correctly classified astrocytoma: LGG, IDH mutant, 1p/19q intact.

Figure 11.9 includes examples of correctly classified LGG, IDH mutated, 1p/19q co-deleted brain tumours (oligodendroglioma). The network focuses on the appropriate tumour regions. In the first example, main focus is put on the small contrast enhancing tumour tissue that is visible in the T1ce sequence.

T1ce   FLAIR   Saliency
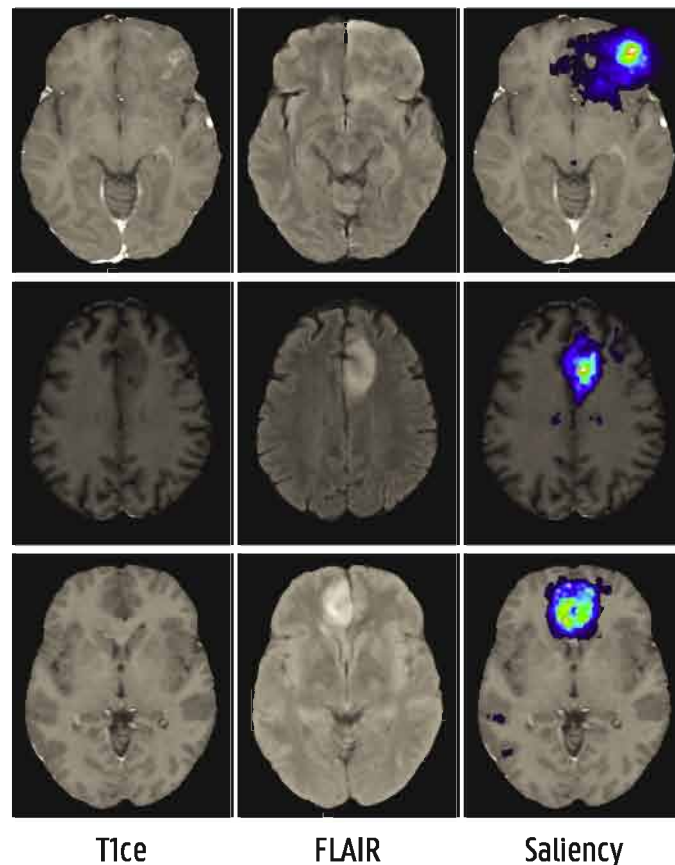
Figure 11.9: Example MRI and saliency maps (overlaid on T1ce) of correctly classified oligodendroglioma: LGG, IDH mutated, 1p/19q co-deleted.

Figures 11.10 to 11.12 illustrate several wrongly classified examples from the TCIA and GUH test datasets. The GBM, IDH wildtype tumours in figure 11.10 are classified as LGG, IDH mutated. In all three cases the network appears to focus on the relevant tumour regions. In the first example, the tumour is hard to identify and little enhancing on T1ce and hyperintensity on FLAIR is observed. The predicted probabilities are 65% and 63% for LGG and IDH mutation respectively. In the second example, main attention is put on the small enhancing lesion. The network does, however, predict this case as an LGG with a probability of 92% and IDH mutated with a probability of 80%. The last example is classified as LGG, IDH mutated with probabilities of 74% and 59%.

Tlce   T2/FLAIR   Saliency
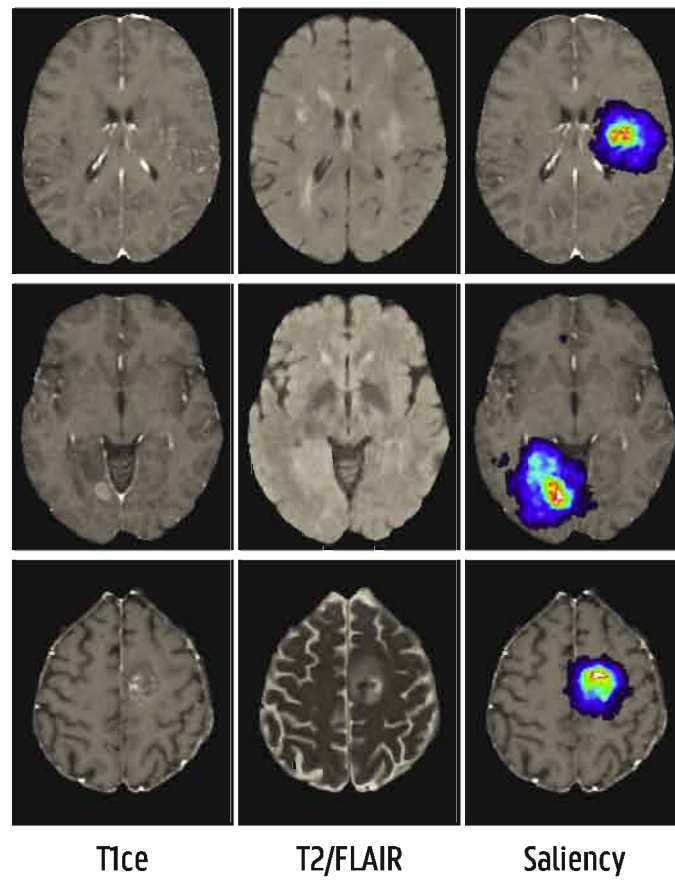
Figure 11.10: Example MRI and saliency maps (overlaid on T1ce) of glioblastoma, IDH wildtype tumours wrongly classified as LGG, IDH mutated.

Example LGG, IDH mutated, 1p/19q intact (astrocytoma) cases that are wrongly classified are depicted in figure 11.11. In the first example, the network focuses on the contrast enhancing tissue that is visible in the T1ce sequence. This tumour is classified as GBM, IDH wildtype, 1p/19q intact with respective probabilities of 90%, 79% and 81%. The middle case is correctly classified as LGG, IDH mutated but incorrectly as 1p/19q co-deleted with a probability of 73%. For the last example, no T2 sequence was available and the network only puts a bit of attention on the tumour region and mainly focuses on a region in the occipital lobe as shown in the additional saliency map slice that is included. This tumour is also classified as LGG, IDH mutated and 1p/19q co-deleted with probabilities: 94%, 88% and 69%.

T1ce   T2   FLAIR   Saliency
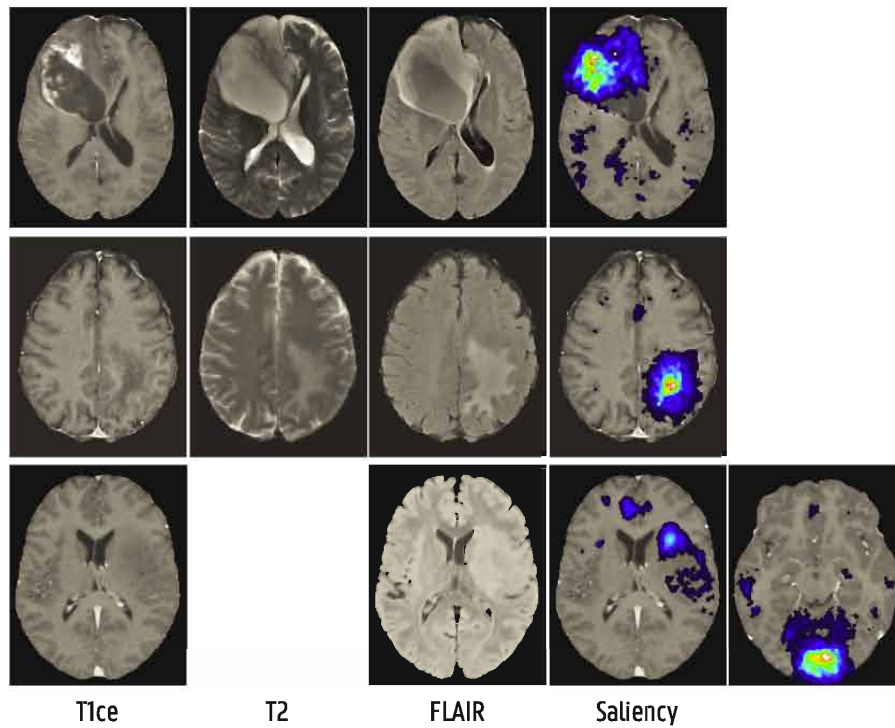
Figure 11.11: Example MRI and saliency maps (overlaid on T1ce) of astrocytoma: LGG, IDH mutated, 1p/19q intact tumours wrongly classified as GBM, IDH wildtype (top) and LGG, IDH mutated, 1p/19q co-deleted (middle and bottom).

Finally, figure 11.12 includes examples of incorrectly classified oligo-dendroglioma (LGG, IDH mutated, 1p/19q co-deleted). In all three cases, the network again focuses on the correct tumour region. The top lesion demonstrates ring enhancement on T1ce and is classified as a glioblastoma, IDH wildtype with respective probabilities of 95% and 92%. The middle case shows some enhancing tissue, but not ring shaped, surrounding a necrotic core. This lesion is predicted as GBM (probability 65%), IDH mutated (probability 56%), 1p/19q intact (probability 76%). For the bottom case, the predicted probabilities for LGG, IDH mutation and 1p/19q co-deletion are 90%, 81% and 28% respectively.

T1ce   T2   FLAIR   Saliency
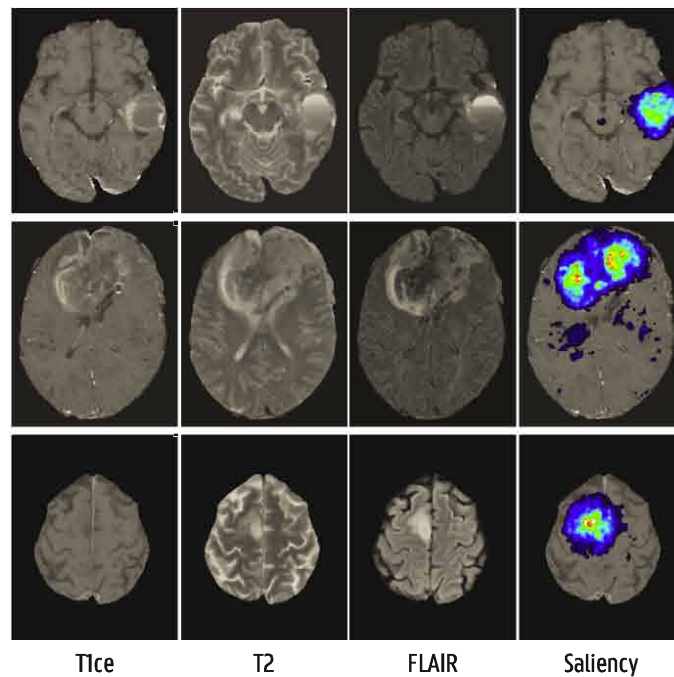
Figure 11.12: Example MRI and saliency maps (overlaid on T1ce) of oligodendroglioma: LGG, IDH mutated, 1p/19q co-deleted tumours wrongly classified as GBM, IDH wildtype (top), GBM, IDH mutated (middle) and LGG, IDH mutated, 1p/19q intact (bottom).

Gradient Ascent

Starting from a random noise 3D matrix, we updated the input with gradient ascent to maximise the output scores towards certain glioma types. This way, imaging features could appear that the network strongly associates with these markers. A synthetic MRI input is generated for a glioblastoma, IDH wildtype, 1p/19q intact brain tumour and shown in figure 11.13. It is interesting to observe that a ring enhancing tumour with necrotic core is generated in the T1ce sequence channel. The tumour core appears surrounded with oedema like tissue that is hypointense on T1 and T1ce and hyperintense on the T2 and FLAIR channels.
Synthetic MRI were also generated for astrocytoma (LGG, IDH mutant, 1p/19q intact) and oligodendroglioma types (LGG; IDH mutated, 1p/19q co-deleted). However, no distinctive features could be clearly identified and are therefore not included.
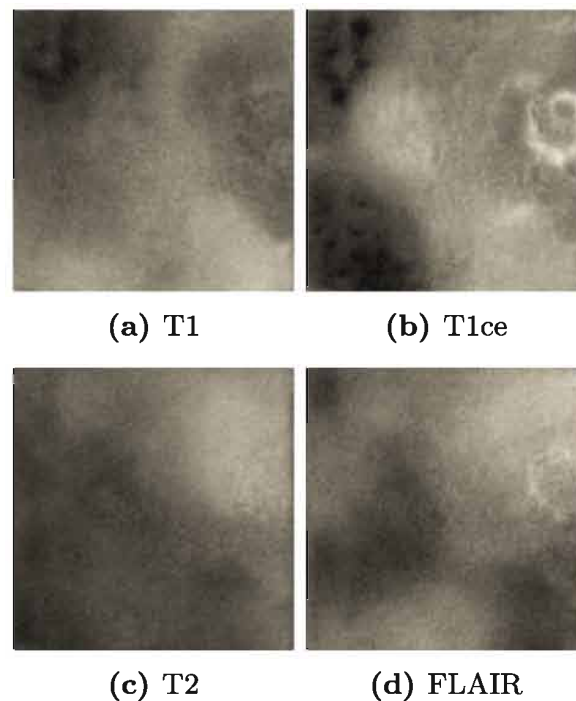
Figure 11.13: Synthesised MRI sequences using gradient ascent for a GBM, IDH wildtype, 1p/19q intact brain tumour.

11.6   Discussion

The segmentation performance reported in section 11.5.1 indicates an overall delineation accuracy that is similar to the performance achieved in chapter 9. Median dice scores are obtained between 85-86%, 88-92% and 89-90% for ET, WT and TC regions, depending on the provided input sequences. This shows that the trained Y-Net is also robust to missing T1 and T2 or FLAIR MRI. The examples in figure 11.2 demonstrating segmentation results with worst, 25th percentile, median, 75th percentile and best dice score also present an overall high delineation accuracy. The tumour with the worst Dice score visually appears well delineated, with potentially some oedema that is missed.

We can conclude that adding the classification tasks and slightly reducing network complexity from 32 to 26 initial feature maps at the highest resolution does not reduce the achieved segmentation performance.

In terms of classification accuracy, the obtained AUC scores appear to be slightly better than in chapter 10. Now, AUC scores are attained on the TCIA test set of 98%, 96% and 87% versus 93%, 94% and 82% in table 10.1 for WHO grade, IDH mutation and 1p/19q co-deletion status

respectively. With a probability threshold of 0.5, a slightly higher sensitivity was obtained for 1p/19q co-deletion with the two-stage pipeline versus the Y-Net approach for the same specificity. Sensitivity and specificity can however be optimised by varying the probability threshold as illustrated in figure 11.3.

On the GUH data, better AUC scores are reached for WHO grade (96% versus 94%) and 1p/19q co-deletion prediction (90% versus 87%) with Y-Net. Especially for 1p/19q co-deletion, a better sensitivity and specificity are achieved. As tumour location is an important feature for 1p/19q co-deletion (more located in the frontal lobe), it is possible the the Y-Net is better able to extract this information from the entire brain MRI. For IDH mutation, AUC is slightly lower (83% versus 86%). On the other hand, sensitivity is higher (88% versus 84%) with only a minor reduction in specificity (difference of 0.56%). The lower specificity compared with TCIA can again be explained by the different method that is used to determine IDH status (see section 10.5). The most difficult cases remain the rare GBM, IDH mutant tumours as only a few of these cases are present in the training dataset. This could be improved by adding additional patients with this tumour type to the dataset.

These results illustrate that high segmentation and classification performances are achieved with the Y-Net architecture and designed training procedure that matches of even outperforms the two-stage approach proposed in chapter 10. Segmentation and classification is done simultaneously on the full brain MRI with only one network. In contrast to most existing approaches (see section 7.2.3), no prior segmentation step is required that could influence the classification performance. The multi-task learning approach applied in chapter 10 is extended with the segmentation task which allowed to train a network from scratch that is able to process entire brain volumes and predict WHO grade, IDH mutation and 1p/19q co-deletion status. Training such a network on the available dataset is challenging due to data heterogeneity, missing labels, missing input modalities and GPU memory constraints. These challenges were tackled by optimising the training procedure through automatic mixed precision training, gradient accumulation, appropriate patch extraction and data augmentations. As a result, high tumour delineation and classification accuracies could be achieved that match state-of-the-art performances reported in existing work (see section 7.2). Moreover, the additional evaluation on an independent dataset acquired at the Ghent University Hospital demonstrated the generalisation capa-

bility of the trained network.

To the author's knowledge no other work has been published that uses one network to perform simultaneous glioma segmentation and prediction of multiple markers. Two other works have been found that use a similar architecture as Y-Net proposed in this chapter. McManigle et al. [355] proposed a Y-Net architecture based on the VGG11 architecture for chest X-ray geometry classification and segmentation of radiographic annotations. A single CNN based on U-Net to segment structures as lung and heart across different modalities and to simultaneously detect the provided input modality was proposed by Harouni et al. [356].

In chapter 10, a first technique (nearest neighbour visualisation) was already applied to gain some insights into the network's visual knowledge. In this chapter, several additional techniques were applied such as t-SNE visualisation, saliency maps and gradient ascent.

Visualisation of the feature embeddings after t-SNE feature reduction, reveals different clusters that strongly correspond with the ground truth tumour labels of grade, IDH mutation and 1p/19q co-deletion. Expanded examples in the glioblastoma cluster demonstrate more (ring-)enhancing tissue with a necrotic core compared to examples in the LGG cluster (see figure 11.4). These are typical features for GBM, IDH wildtype tumours and are also observed in LGG cases that are attributed to the GBM cluster based on the network's extracted features. Conversely, glioblastoma cases that are attributed to the LGG cluster do not show this characteristic.

As contrast enhancement is also a specific feature for IDH wildtype tumours, the IDH clusters strongly overlap with grade. Within the LGG group one can also observe that LGG, IDH wildtype tumours are positioned closer towards the GBM cluster on the left compared to LGG, IDH mutated tumours (see figure 11.5).

Furthermore, two groups can also be distinguished within the LGG cluster corresponding to 1p/19q co-deletion status (see figure 11.6). Expanded examples a the top show small lesions located in the frontal lobe with potentially some slight enhancement. These are indeed characteristic features of 1p/19q co-deleted glioma (see section 7.2.1). At the bottom right, lesions are larger and demonstrate T2-FLAIR mismatch which is highly specific for IDH mutant, 1p/19q intact glioma (see section 10.5).

The above visualisations give additional confidence on the relevance of the features extracted by the network with respect to glioma diagnosis.

Indeed, different cases are grouped that show visually similar characteristics that also correspond with existing knowledge on the correlation between tumour phenotype and genetic markers.

Additionally, saliency maps are produced that visualise which input pixels had the most influence on the network's prediction. Examples are included for correctly classified glioma in figures 11.7 to 11.9 and wrongly classified glioma in figures 11.10 to 11.12. Overall, the network focuses on the relevant tumour regions which gives additional confidence on the network's predictions.

Glioblastoma that are not predicted as such (figure 11.10) show small lesions with little enhancement and not the typical ring-enhancement surrounding a necrotic core which can be a reason why these are classified as LGG, IDH mutated.

In the last example in figure 11.11, illustrating incorrectly classified LGG, IDH mutated, 1p/19q intact cases, the network puts most attention to the wrong region which does not contain tumour tissue. Hence, saliency maps can also be used to verify whether the network indeed looks at the correct region and identify possibly incorrect predictions. The top example demonstrates clear enhancing tumour tissue. The network focused on this region which can explain why this lesion is classified as GBM, IDH wildtype.

The first oligodendroglioma example in figure 11.12 is classified as a GBM, IDH wildtype possibly due to the ring shaped enhancing tissue surrounding a hypo-intense core on T1ce MRI which is more characteristic for glioblastoma, IDH wildtype. In all three examples, the network again focuses on the appropriate tumour regions. The middle is predicted as GBM, IDH mutated, but with low certainty (probabilities close to 0.5). The last example is correctly identified as LGG, IDH mutated but not as 1p/19q co-deleted.

The synthetic MRI of a glioblastoma, IDH wildtype tumour that is generated using gradient ascent (starting from a random noise image) is a neat visualisation that indeed the network learned to associate a ring-enhancing tumour with necrotic core to this glioma type. This imaging characteristic is only clearly visible in the T1ce channel and not in the T1 and T2 or FLAIR channels. Moreover, this pattern is not generated for other LGG, IDH mutant glioma showing that this is not always synthesised and only associated with GBM, IDH wildtype tumours.

Deep learning networks are often seen as a black box that reveal little information on how and why they make certain predictions. The above techniques allow to gain some first insights into the trained deep learning model. This can give additional confidence in the network's predictions and checks to verify where it might go wrong or which predictions could be incorrect (e.g. when focusing on the wrong region). The visualisation results indicate that the network was able to learn several imaging features that are characteristic of certain tumour types and correspond with existing knowledge (see section 7.2.1) such as the presence of contrast enhancing tissue and T2-FLAIR mismatch. To gain more and deeper insights, network visualisation techniques could be applied to visualise and interpret feature maps inside the network. Furthermore, bayesian deep learning techniques could be added to obtain uncertainty information on the classification and segmentation predictions and allow to identify uncertain and potentially incorrectly classified cases that require further attention.

11.7   Conclusion

In this chapter, we developed an architecture, called Y-Net, for simultaneous segmentation and classification of glioma. Based on pre-therapy MRI, the network is able to automatically and accurately delineate different tumour tissues and predict WHO grade, IDH mutation and 1p/19q co-deletion status. This approach is beneficial as it operates on the entire brain MRI and no prior segmentation step is required. Through the use of multi-task learning and techniques to reduce GPU memory consumption, one network could be trained on a large multi-institutional and heterogeneous database containing many cases with missing labels. Furthermore, performance was validated on an entirely independent dataset. Finally, insights into the network's visual knowledge and extracted imaging features were obtained using visualisation techniques such as t-SNE feature embedding, saliency maps and synthesising input patterns that are typical for certain glioma types according to the network. These techniques provide first steps towards opening the black box and interpreting deep neural networks for computer-aided diagnosis.

12 | Conclusions and future perspectives

In this final chapter, we summarise the main conclusions drawn from each part of this thesis. Based on these findings, we formulate some remaining limitations and possibilities for future work and research directions.

12.1 Summary

This work examined the use of artificial intelligence throughout the medical imaging chain. Specifically, two applications were explored situated at the beginning, acquisition, and at the end, analysis, of medical imaging.

As this dissertation is located at a crossroad between artificial intelligence and medical imaging, these research domains were first introduced in chapters 2 and 3.

Chapter 2 presented an overview of artificial intelligence. We explained how AI concepts have been around for decades but are only recently implemented into real applications, driven by the ever increasing computational power and amount of data. The relation between AI and its subfields machine learning and deep learning as techniques to realise intelligent machines was clarified. Furthermore, the basic concepts of machine learning were explained and the elemental ML algorithms, linear and logistic regression, were discussed in more detail. These algorithms and the way their parameters are optimised using gradient descent form the foundation of artificial neural networks. The main challenge of developing machine learning algorithms, being the generalisation performance on new, unseen data was explained together with possible regularisation

techniques to prevent overfitting on training data. Next, the concept of
(deep) artificial neural networks was elucidated, starting from the basic
building block, the artificial neuron as a graphical representation of a
linear regression function. We focused on how the parameters of deep
neural networks can be optimised based on labelled training examples
together with regularisation techniques to improve the generalisation
performance of complex networks. Finally, a special type of neural
networks, called convolutional neural networks, were discussed in detail
as this type of network is most used in medical imaging and computer
vision in general.

Chapter 3 covered the role and state-of-the-art of AI in medical imaging.
The potential and need for AI to cope with the growing amounts of
healthcare data, increase efficiency and enable precision and person-
alised medicine was discussed. The elemental principles were explained
behind the most common medical imaging modalities including X-ray,
ultrasound, computed tomography, magnetic resonance imaging and nu-
clear medicine. Positron emission tomography, the topic of the first AI
application in this work, was covered in more detail with a focus on
detector design and requirements. We explained the advantages of using
a monolithic crystal instead of a pixelated crystal design to stop incoming
gamma rays in terms of sensitivity and spatial, temporal and energy
resolution. Monolithic PET detectors do, however, require lengthy cal-
ibration procedures and complex positioning algorithms to determine
the exact position of interaction inside the crystal. Hence the potential
of AI in PET detector calibration. The remaining part of chapter 3
included an overview of state-of-the-art AI applications throughout the
entire medical imaging chain. Starting with image formation we saw
how AI can improve the quality of the raw acquisition data, advance the
image reconstruction process and further enhance image quality through
post-processing. Furthermore, AI can transform medical image analysis,
helping radiologists meet the rising demand for imaging examinations
and leverage these large amounts of data towards precision medicine.
While challenges remain regarding training data availability, variability
of image quality and interpretability, AI systems already achieve high
performances in segmentation, detection and diagnosis across numerous
anatomical application areas that match or even outperform human
radiologists.

12.1.1 PET detector calibration

In chapter 4, we performed a comprehensive evaluation on the use of neural networks for 3D gamma interaction positioning in a monolithic PET detector using optical simulation data. Spatial resolution was assessed as a function of network complexity (by varying the number of layers and neurons in each layer), amount of training data and training and validation procedure. It was concluded that network complexity should be tuned to the calibration setup and not be too complex to avoid overfitting. Through the use of validation data, acquired at intermediate positions that are not in the training set, potential overfitting on the training grid could be identified. Based on the validation loss, training can be stopped before strong overfitting and thus non-uniform positioning starts to occur. Optimal performance was achieved with a network containing three hidden layers of 256 neurons trained on 1000 events per training grid position. Results showed that a very high spatial resolution was obtained of around 0.50 mm FWHM across the entire detector. Comparison with an established positioning algorithm, called mean nearest neighbour, demonstrated superior performance both in spatial resolution as in computational efficiency.

Two factors that could degrade the positioning accuracy of neural networks are intra-crystal Compton scatter and calibration source beam width. These effects are investigated in chapter 5.
Around 60% of the arriving gamma rays first undergo one or multiple Compton interactions before final photoelectric absorption. Consequently the first Compton interaction position, which is the required position that needs to be estimated, is different from the final interaction position. Estimation of the first interaction position from the measured electronic signal is difficult as often only a small amount of energy is released when Compton scattering. Evaluation of spatial resolution with and without Compton scattered events revealed that Compton scatter has a significant degrading effect on the overall positioning accuracy (mean 3D positioning error of 2.29 mm versus 0.49 mm). However, the positioning error depends on the scatter distance and only a small fraction of events scatters very far (10% more than 8 mm). A network specifically trained to position Compton scattered events did not result in an improvement in performance. We therefore investigated whether networks can identify far scattered events and could help to improve performance. To this end, a network was trained to predict 3D scatter

distance. This network could be used to filter out far scattered events in order to improve spatial resolution with a tradeoff in sensitivity which can be justified in certain applications. Considering the limited practicality of training a scatter prediction network in an experimental setup (no available labels), a different approach was investigated using a Bayesian neural network. This method allows to train one network to predict both the position as the positioning uncertainty related to Compton scatter without requiring additional information on Compton scattering. When filtering out 10% most uncertain events, the mean positioning error could be reduced from 1.54 mm to 1.23 mm.

A calibration source with a certain beam width can introduce differences between the ground truth position label and the actual first interaction position. These errors in the ground truth data could influence the training process of neural networks. Comparison between a network trained on data acquired with a perfectly narrow beam versus a calibration source with a realistic beam width of 0.6 mm showed no significant difference in achieved intrinsic spatial resolution. The beam diameter does, however, influence the measured spatial resolution (0.74 mm versus 0.52 mm FWHM) which should be taken into account when evaluating and comparing spatial resolution of different PET detectors.

Chapters 4 and 5 evaluated the positioning performance of neural networks on simulation data which does not take all possible non-idealities into account that can be present in an experimental setup. Validation of the developed methodology in chapters 4 and 5 on experimental data was performed in chapter 6. Similar to the results on simulation data, high spatial resolutions (around 1 mm FWHM in detector centre) could be achieved with neural networks, superior to the mean nearest neighbour positioning algorithm (1.14 mm FWHM in centre region). Neural networks are trained on individual events and directly learn to infer the interaction position from the measured light distribution. This leads to an improved positioning accuracy of Compton scattered events and less degradation near the detector edges. Moreover, neural networks produce continuous coordinate outputs, not restricted to a discrete calibration grid. Improved spatial resolution of PET detectors with neural networks can help reach the physical limits of PET and a better detection of small tumours. Furthermore, when achieving better spatial resolutions than required, there is room to trade resolution for other parameters e.g.: less readout channels, inexpensive materials with less light output, detector thickness, etc. Lastly, positioning events with the network is fast and

parallelisable, especially when using powerful hardware like GPUs.

12.1.2 Computer-aided primary brain tumour diagnosis

The second part of this work focused on the application of AI in medical image analysis, specifically for primary brain tumour segmentation and diagnosis. Chapter 7 introduced the basic anatomy of the brain which is necessary to understand the different types of primary brain tumours that are defined by the World Health Organisation. We focused on the most common type of PBTs, glioma, and the most recent classification guidelines of the WHO to differentiate tumours based on malignancy (WHO grade) and molecular markers (IDH status and 1p/19q co-deletion). We then further discussed PBT epidemiology, symptoms, diagnosis, survival and different treatment options in relation to these important markers. The importance of non-invasive tumour characterisation based on pre-therapy MRI was described. This allows to avoid biopsy or resection which involve risks and are not always possible to perform. Moreover, early determination of tumour markers can guide initial therapy and surgery planning. After introducing the required background knowledge, an overview was provided of recent literature on glioma segmentation and diagnosis with artificial intelligence techniques.

Primary brain tumour malignancy has strong prognostic and therapeutic implications. Therefore, we investigated the task of non-invasively distinguishing high-grade glioblastoma from lower-grade glioma in chapter 8. The BraTS 2017 dataset consisting of 210 GBM and 75 LGG cases was used for this study. For every patient, four MRI sequences (T1, T1ce, T2 and FLAIR) were provided with manual tumour segmentation labels. Predictive performance was assessed of hand-engineered radiomics features that describe tumour shape, texture and intensity and features extracted using a pre-trained CNN. Moreover, we compared the performance of pre-trained CNN features extracted from different input scales: one or multiple slices and with or without cropping to the tumour ROI. Classification of the features was done using a Random Forest classifier. Best performance was achieved with shape, intensity and texture features extracted from manually segmented tumour volumes (AUC of 96%). Features from a pre-trained CNN, on the other hand, had a high predictive value as well and allowed to design a fast and automatic binary grading system reaching an AUC score of 91%. These results indicate that CNNs hold the potential to develop accurate, reproducible

and fully automatic CAD systems and when training them from scratch to process medical imaging data, performance could possibly even be improved.

In chapter 8, and many other existing studies on brain tumour diagnosis, manually obtained segmentation maps are used to extract tumour features. Manual brain tumour delineation on medical images is, however, time-consuming and can suffer from inter- and intra-observer variability. Existing automatic segmentation algorithms using CNNs and more specifically U-Nets achieve high performances but mostly require all four MRI modalities to be available. This is not always the case in clinical practice and in the brain tumour dataset that is collected in this work. We therefore developed an automatic, accurate and fast segmentation deep learning algorithm based on the U-Net architecture that is robust to missing input modalities. The network was trained using the BraTS 2019 training dataset and evaluated on the BraTS 2019 validation set. Accurate delineation of different tumour regions was achieved with average Dice scores of 90%, 83% and 76% for the total abnormal, tumour core and enhancing tumour regions respectively. Through channel dropout, i.e. randomly excluding input MRI during training, robustness to missing input modalities could significantly be increased. These scores match state-of-the-art results reported in the most recent BraTS challenges and we believe that the obtained performance is sufficiently high to be useful in a clinical setting. It has to be taken into account that the segmentation results of the network are compared with manual segmentations. Manual delineations suffer from inter- and intra-reader variability and thus not 100% accurate. It can therefore be debated whether further improving the Dice scores with a few percentages is clinically relevant. Objectivity and robustness could be more important when analysing brain tumour volumes and progression over time. Qualitative evaluation on independent data acquired at the Ghent University Hospital showed good generalisation performance.

In order to train a brain tumour classification network from scratch in chapter 10, a large dataset of 628 patients was collected from multiple public databases available on The Cancer Imaging Archive. To be included in the dataset, at least a pre-operative T1ce MRI together with a T2 and/or FLAIR sequence of sufficient quality was required together with information on WHO grade, IDH mutation and 1p/19q codeletion status. The segmentation algorithm from chapter 9 was applied to this data to extract the 3D tumour region of interest from every

MRI sequence. Subsequently a classification 3D CNN was trained to not only predict tumour grade but also the important molecular markers: IDH mutation and 1p/19q co-deletion status. One network was trained to simultaneously predict these three markers based on the 3D tumour ROI extracted from the four MRI sequences. This was possible through the use of multi-task learning which also allowed to deal with missing ground truth labels in the dataset and reduce the risk of overfitting. On a test dataset of 100 patients, not used during training, the network achieved AUC scores of 93% for WHO grade, 94% for IDH mutation and 82% for 1p/19q co-deletion prediction. We additionally evaluated the classification performance on an entirely independent dataset of 110 patients retrospectively acquired at the Ghent University Hospital. On this dataset, AUC scores were reported of 94%, 86% and 87% for the three tasks respectively.

The two-stage approach proposed in chapter 10 (segmentation followed by classification) can have some downsides as the classification network only operates on the tumour region of interest which excludes potentially relevant information on location and surrounding tissue. Moreover, possible errors in the prior segmentation step could also influence the subsequent classification performance. As an alternative, a network that performs simultaneous segmentation and classification based on the full brain MRI was explored in chapter 11. The U-Net architecture from chapter 9 was extended with a classification branch and called Y-Net. Through the use of multi-task learning, techniques to reduce GPU memory consumption and appropriate patch extraction, one network could be trained on the large multi-institutional and heterogeneous database containing many cases with missing labels.
A similar segmentation performance was achieved with average Dice scores of 89%, 84% and 75% for the whole tumour, tumour core and enhancing tumour regions respectively. In terms of classification performance, WHO grade could be predicted with 98%, IDH mutation with 96% and 1p/19q co-deletion with 87% AUC on the TCIA test dataset. On the independent GUH test data, the AUC scores were 96%, 83% and 90%. Overall, a slightly better performance was achieved as in chapter 10 which is possibly because Y-Net is able to process the full input MRI instead of only the tumour ROI and the addition of the segmentation task could provide additional regularisation to the training process.
Finally, insights into the network's visual knowledge and extracted imaging features were obtained using several visualisation techniques. The

feature embeddings of the network were plotted for every brain tumour case in the dataset after t-SNE feature reduction. This revealed different clusters of brain tumour cases with similar imaging characteristics that strongly corresponded with the ground truth labels on WHO grade, IDH mutation and 1p/19q co-deletion. Glioblastoma, IDH wildtype tumours that show ring-enhancing tumour tissue with a necrotic core were grouped together. Within the lower-grade glioma cluster, different groups could be identified depending on IDH mutation (IDH wildtype closer to the GBM cluster) and 1p/19q co-deletion. Small lesions located in the frontal lobe were seen as typical for LGG, IDH mutant and 1p/19q co-deleted tumours. On the other hand, LGG, IDH mutant and 1p/19q intact cases that demonstrated larger lesions with T2-FLAIR mismatch were grouped at the other side of the LGG cluster. These are indeed known imaging features that are correlated with these tumour markers. Saliency maps, that visualise where the network places the most attention in the input MRI to make a certain prediction, showed that the network indeed looks at the relevant tumour regions. This allows an additional check to gain confidence in the network's predictions.

Lastly, a synthetic input was generated that maximises the output scores for a glioblastoma, IDH wildtype tumour. Starting from random noise, a ring-enhancing tumour pattern appeared with a hypo-intense core in the T1ce channel and surrounding hyper-intense tissue on the T2 channels. This indicates that the network learned to attribute these features to this tumour type.

12.2 Future directions

12.2.1 PET detector calibration

To bring the use of neural networks for gamma ray positioning into practice, their adoption in a complete PET scanner setup should first be investigated. In this work, a network was trained and evaluated on one detector. A complete PET scanner contains many detectors and the question remains whether a network trained for one detector is applicable to the other detectors as well. Although all detectors in the scanner share the same design, small differences in crystal inhomogeneities, surface finish, connection with the SiPM array, variable SiPM gains, electronic noise etc., could lead to differing measured light distributions

and therefore reduced positioning performance. Acquiring calibration data and training a separate network for every detector would be very time-consuming. It should therefore be investigated how much spatial resolution degrades when a network is applied to a different detector. To reduce calibration time for other detectors, a network trained for one detector could be fine-tuned on data from the new detectors. This way, only a few events, possibly acquired at less calibration positions, would be required instead of a full calibration when training a network from scratch. An other, more efficient approach could be to train a network on a combined calibration dataset of many detectors. Overall spatial resolution could slightly decrease as the network has to cope with variations across all these detectors, but positioning performance would be more stable and similar for every detector. Moreover, only one network needs to be trained that can be applied to all detectors in the scanner and even to the detectors in other PET scanners with the same design. This could significantly reduce the calibration time.

In this context, it could also be examined whether a network trained on simulation data can be used for an experimental PET detector setup with the same design and geometry. Using a network trained on simulation data would eliminate the need to acquire experimental data. Moreover, estimation of the depth-of-interaction could potentially improve as the exact DOI information is available in simulation but not in experimental data. Training and evaluation of algorithms to predict DOI remains challenging in an experimental setting. A preliminary evaluation of 2D resolution, not included in this book, with the simulation and experimental data used in this work, showed a significant reduction in spatial resolution, especially at the edges. But overall, the positioning performance was still acceptable, particularly in the detector centre, with limited bias but a broader spread (larger FWHM). To improve performance, additional noise and non-idealities could be incorporated in the simulation setup to more closely match a realistic setting and make the network more robust to these variations. Furthermore, the network could be fine-tuned on a small amount of experimental data and it could be investigated how much additional data is required.

To further improve spatial resolution, we investigated whether (far) Compton scattered events, associated with the worst positioning performance, could be filtered out, thereby sacrificing some sensitivity. A scatter distance prediction network was implemented which worked well on simulation data. Application of this methodology on experimental

data was, however, more challenging as no ground truth scatter distance information is available. The Bayesian neural network approach does not require additional labels but mostly filtered out events in the corner as the background filtering applied to the training data already filtered many scattered events. Without scattered data in the training set, the network is not able to learn uncertainty related to Compton scatter. The bayesian deep learning approach to extract uncertainty measures related to scatter distance could further be examined by using different background filtering techniques that remove no or less scattered events or by using a different crystal without background activity such as BGO. An additional approach could be to use a scatter distance prediction or bayesian network trained on simulation data.

The required rate and used hardware to process events with neural networks should also be further examined. Events can be positioned very fast on powerful GPUs which would require all events (SiPM signals) to be transferred from the detectors to a central processing unit of the PET scanner, equipped with a GPU. This can be feasible in small (pre-clinical) PET scanners with a limited amount of detectors but could become unattainable in large (total body) PET systems. Very high bandwidth and storage capacity would be required if no realtime positioning is possible. The neural network positioning could also be implemented on detector level on an FPGA. The network complexity might have to be reduced to fit the memory and processing speed of the FPGA with possibly a small reduction is spatial resolution.

Finally, next to positioning, the use of neural networks could also be investigated to estimate timing information and improve time-of-flight resolution. Digitised detector waveforms would have to be processed instead of total SiPM energies and convolutional neural network architectures could be more optimal [86]. Accurate TOF estimation could further improve PET resolution and image quality.

12.2.2 Computer-aided primary brain tumour diagnosis

First of all, before clinical application, the performance of the brain tumour segmentation and diagnosis networks would further have to be clinically validated on a larger cohort of patients. The networks in this dissertation were already validated on an independent dataset of 110 patients, but for 1p/19q co-deletion for example, only 12 co-deleted patients

were included which might be too small to obtain reliable performance estimates.

This work should also not be viewed as a standalone tool that would replace radiologists. Supervision and final decision of expert clinicians that can take the full patient context into account remains of vital importance. The tools developed in this work should be viewed in light of computer-aided brain tumour characterisation. The models provide segmentations and probabilities for several tumour markers that can be interpreted by the radiologist in order to determine the initial diagnosis, prognosis and therapy planning. An additional interesting study is to assess how well the models can aid the radiologist and how they affect diagnosis, prognosis and therapy planning.

To advance clinical applicability, one could also include the prediction of more molecular markers. There are many more markers that hold clinical significance and implications on optimal therapy such as MGMT promoter methylation, PTEN, ATRX and TP53 mutations, EGFR amplification etc. [3].

Furthermore, the models developed in this work are trained to segment and predict important markers for glioma. This requires a prior diagnosis of the brain lesion as glioma which is not always straightforward. To maximise clinical relevance, the CAD tools should also be extended for additional primary brain tumour types and even other brain lesions. Examples of other important primary brain tumours are meningioma, medulloblastoma, CNS lymphoma, ependymoma and pituitary tumours. Additional brain lesions could be metastatic tumours, infections, multiple sclerosis lesions, traumatic lesions and hemorrhagic lesions.

To further improve diagnostic performance and add the prediction of additional markers and lesion types, a much larger curated dataset would have to be collected, preferably from many different centres. Moreover, the dataset could be expanded with other functional imaging modalities such as PET and diffusion and perfusion MRI. Existing literature has already shown the added value of these imaging methods for several brain tumour classification tasks [357].

An additional possible research track is to improve the techniques that are used in this work.

The pre-processing steps applied to the brain MRI were performed with SPM12. Although the pre-processing is fully automatic, the efficiency and accuracy of different steps as co-registration and skull-stripping could also be further improved by using deep learning algorithms.

To deal with missing input modalities, image translation techniques (see section 3.3.3) could be used to generate the missing scans from the available sequences. For example, generative adversarial networks could be trained that translate the available T2 sequence to a FLAIR MRI and vice versa [358]. Using the the translated image instead of setting the corresponding channel to zero could further improve performance.

With more data and GPUs with more memory, increasingly complex neural networks can be trained that are able to predict numerous characteristics and lesion types. Other network architecture and layers can be explored such as Dense U-Nets, attention layers, multi-scale networks, strided convolutions etc.

When applying multi-task learning, the global loss is calculated as an average of the individual losses from the different tasks. Some tasks are, however, potentially easier than others and converge faster. It could therefore be beneficial to assign different weights to each task and use a weighted average to calculate the global loss metric that is back-propagated. Manually tuning these weights would be cumbersome and the different task weights could be automatically learned by the network using homoscedastic uncertainty estimation [359, 360].

In this work, a first step was made into interpreting the features that were learned by the CNNs through several visualisation techniques. However, more research is required towards interpretability of deep neural networks to fully understand how and why they make certain predictions and their reliability. Bayesian deep learning techniques were also not yet implemented in the brain tumour characterisation networks of this work. Extraction of uncertainty estimates is important to inform the clinician when a model is certain on its prediction or when it is merely guessing, e.g. when feeding a brain lesion type that the network has never seen before.

12.3   Integration of AI in radiology

This section highlights several key remaining challenges to the integration of AI in the radiological workflow and its impact on radiologists. For other works containing more thorough discussions on practical, ethical and legal aspects of AI in healthcare, we refer the reader to Ranschaert et al. [361], Allen et al. [362], Geis et al. [363], and Gerke et al. [364].

The current AI systems that are developed for numerous applications

in radiology are examples of 'narrow AI'. Algorithms are trained for one specific well-defined task, often in image interpretation, and are able to reach performances that match or are even better than humans. Since the radiologist's work is mainly known for image interpretation tasks, which in some cases can be considered as narrow and well-defined, the recent successes of AI algorithms outperforming radiologists has led to the misconception that AI will replace the entire radiological profession. The radiological work, however, goes beyond just image interpretation, including discussion with a multidisciplinary team of physicians, selection of the right imaging methods, performing image-guided invasive procedures, integration with other data from the electronic health record, taking the full patient context into account, interaction with patients, quality control and education etc. Performing these tasks automatically will not be possible in the short term and would require 'general AI' which is still far from being a reality. Current 'narrow AI' systems can replace certain image reading tasks and serve as tools to help radiologists in improving the quality and efficiency of image reading. This can provide extra time for other essential tasks, although the growing workload and demand for imaging examinations could limit the freed time and AI could be indispensable.

There remain, however, several practical, ethical and legal challenges that need to be overcome to enable integration of AI into the radiological workflow.
To motivate routine use of AI applications by radiologists, smooth integration of AI into the existing PACS interface is required. The necessary infrastructure should be available that enables big data handling and facilitates seamless interaction between all hardware, software and data services (PACS and electronic health records). Furthermore, exchange of health data between different sites should be facilitated while safeguarding data security and patient privacy. In this context, blockchain technology has been proposed to allow decentralised health data sharing [365]. Exchanging data is necessary to build large curated and high-quality datasets. Deep learning algorithms require large amounts of data for training and testing which should be properly managed. Not only the amount, but also the quality of the data and the ground truth labels is primordial. Moreover, training dataset should be representative and free of unintended bias against subsets of individuals related to ethnic, gender, socioeconomic status etc.
Next to ethical questions related to patient data safety and privacy, there

are also ethical issues concerning algorithm safety and transparency. There are ongoing discussions on regulatory approval, standards of care and liability of AI systems, not only in healthcare but also in many other contexts such as autonomous vehicles. In the context of narrow AI tools, most likely radiologists will remain accountable. AI tools should be rigorously tested and evaluated through clinical validation studies to obtain regulatory approval and human supervision and monitoring of performance in the clinical workflow remains vital. Additionally, the algorithms should be sufficiently transparent in how and why certain decisions are made to optimally help radiologists in making the ultimate decision.

Solving the above practical, ethical and legal question requires collaboration between all involved parties including developers, physicians and governmental agencies. Especially radiological and medical societies should play a leading role to define clear use cases of AI, develop and manage large and high quality datasets and defining policies for the development and usage of AI applications while guarding patient privacy and ethical principles. To cope and take part in this changing radiological landscape, a basic knowledge of AI should be incorporated into the radiologist's training curriculum. Radiologists should be trained to safely use AI tools to improve their radiological workflow while maintaining critical thinking and not become overly reliant on automated decisions.

12.4 Conclusion

In this dissertation, we have shown that artificial intelligence can be applied to and advance the entire medical imaging chain. We have demonstrated that neural networks can improve the image acquisition process on detector level which eventually results in better image quality and affects the entire remaining imaging pipeline. Furthermore, an image analysis application was researched on primary brain tumour characterisation resulting in non-invasive and accurate brain tumour segmentation and diagnosis tools. Although challenges remain regarding standardised datasets and understanding of AI, both by experts and the general public, we can conclude that AI will have a profound impact on radiology. It will improve efficiency, perform routine tasks and enable personalised and precision medicine, thereby liberating time of radiologists to focus on aspects of the medical profession that can never be automated such

as empathy, care, patient and family support, expertise and integration of full clinical and emotional context.

Bibliography

[1] The Royal College of Radiologists. *Clinical radiology UK workforce census 2019 report.* 2020. URL: https://www.rcr.ac.uk/clinical-radiology/service-delivery/rcr-radiology-workforce-census.

[2] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (December 2012 2017), pp. 60–88. ISSN: 13618423. DOI: 10.1016/j.media.2017.07.005.

[3] David N. Louis et al. "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary". In: *Acta Neuropathologica* 131 (6 June 2016), pp. 803–820. ISSN: 0001-6322. DOI: 10.1007/s00401-016-1545-1.

[4] Meng Law et al. "Glioma grading: sensitivity, specificity, and predictive values of perfusion MR imaging and proton MR spectroscopic imaging compared with conventional MR imaging". In: *AJNR American journal of neuroradiology* 24 (10 2003), pp. 1989–1998. ISSN: 0195-6108. DOI: 10.3174/ajnr.a3686.

[5] J A Carrillo et al. "Relationship between Tumor Enhancement, Edema, IDH1 Mutational Status, MGMT Promoter Methylation, and Survival in Glioblastoma". In: *American Journal of Neuroradiology* 33 (7 2012), pp. 1349–1355. DOI: 10.3174/ajnr.A2950.

[6] Songtao Qi et al. "Isocitrate dehydrogenase mutation is associated with tumor location and magnetic resonance imaging characteristics in astrocytic neoplasms." In: *Oncology letters* 7 (6 June 2014), pp. 1895–1902. ISSN: 1792-1074. DOI: 10.3892/ol.2014.2013.

[7] D R Johnson et al. "Genetically Defined Oligodendroglioma Is Characterized by Indistinct Tumor Borders at MRI." In: *AJNR. American journal of neuroradiology* 38 (4 Apr. 2017), pp. 678–684. ISSN: 1936-959X. DOI: 10.3174/ajnr.A5070.

[8] Yukihiko Sonoda et al. "Association between molecular alterations and tumor location and MRI characteristics in anaplastic gliomas". In: *Brain Tumor Pathology* 32 (2 Apr. 2015), pp. 99–104. ISSN: 1433-7398. DOI: 10.1007/s10014-014-0211-3.

[9] S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* 4th ed. Pearson series in artificial intelligence. Pearson, 2020. ISBN: 9780134610993.

[10] Ethem Alpaydin. *Introduction to machine learning.* Third edit. The MIT Press, 2014, p. 613. ISBN: 9780262028189.

[11] Christopher M Bishop. *Pattern Recognition and Machine Learning.* Springer-Verlag New York, 2006. ISBN: 9788578110796. DOI: 10.1017/CBO9781107415324.004.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, 2016. URL: http://www.deeplearningbook.org/.

[13] John McCarthy et al. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955". In: *AI Magazine* 27 (4 2006), p. 12. DOI: 10.1609/aimag.v27i4.1904.

[14] Alan M. Turing. "Computing machinery and intelligence". In: *Mind* LIX (236 1950), pp. 433–460. DOI: 10.1093/mind/LIX.236.433.

[15] *Partnership on AI.* 2016. URL: https://www.partnershiponai.org/about/.

[16] A. L. Samuel. "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM Journal of Research and Development* 3 (3 July 1959). ISSN: 0018-8646. DOI: 10.1147/rd.33.0210.

[17] Leo Breiman. "Random Forests". In: *Machine Learning* 45 (2001), pp. 5–32. DOI: https://doi.org/10.1023/A:1010933404324.

[18] Warren S Mcculloch and Walter Pitts. "A logical calculus of
      the ideas immanent in nervous activity. 1943." In: *Bulletin of
      Mathematical Biology* 52 (2 1990), pp. 99–115. DOI: 10.1007/
      BF02459570.

[19] F. Rosenblatt. "The perceptron: A probabilistic model for infor-
      mation storage and organization in the brain." In: *Psychological
      Review* 65 (6 1958). ISSN: 1939-1471. DOI: 10.1037/h0042519.

[20] G. Cybenko. "Approximation by superpositions of a sigmoidal
      function". In: *Mathematics of Control, Signals, and Systems* 2 (4
      Dec. 1989). ISSN: 0932-4194. DOI: 10.1007/BF02551274.

[21] Kurt Hornik, Maxwell Stinchcombe, and Halbert White.
      "Multilayer feedforward networks are universal approximators".
      In: *Neural Networks* 2 (5 1989), pp. 359–366. ISSN: 0893-6080.
      DOI: https://doi.org/10.1016/0893-6080(89)90020-8.

[22] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Deep Sparse
      Rectifier Neural Networks". In: *Proceedings of Machine Learning
      Research* (June 2011), pp. 315–323. URL: http://proceedings.
      mlr.press/v15/glorot11a.html.

[23] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. "Rectifier
      Nonlinearities Improve Neural Network Acoustic Models". In:
      *ICML Workshop on Deep Learning for Audio, Speech, and
      Language Processing* 30 (2013).

[24] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams.
      "Learning representations by back-propagating errors". In: *Nature*
      323 (6088 Oct. 1986). ISSN: 0028-0836. DOI: 10.1038/323533a0.

[25] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for
      Stochastic Optimization". In: *3rd International Conference for
      Learning Representations* (Dec. 2014). URL: http://arxiv.org/
      abs/1412.6980.

[26] Sergey Ioffe and Christian Szegedy. "Batch Normalization:
      Accelerating Deep Network Training by Reducing Internal
      Covariate Shift". In: *Proceedings of the 32nd International
      Conference on Machine Learning* (2015), pp. 448–456. URL: http:
      //proceedings.mlr.press/v37/ioffe15.html.

[27] Geoffrey E. Hinton et al. "Improving neural networks by
      preventing co-adaptation of feature detectors". In: *arXiv* (July
      2012). URL: http://arxiv.org/abs/1207.0580.

[28]   Nitish Srivastava et al. "Dropout: A Simple Way to Prevent
       Neural Networks from Overfitting". In: *Journal of Machine
       Learning Research* 15 (2014), pp. 1929–1958. URL: http://jmlr.
       org/papers/v15/srivastava14a.html.

[29]   Rich Caruana. "Multitask Learning". In: *Machine Learning* 28
       (1 1997), pp. 41–75. ISSN: 08856125. DOI: 10 . 1023 / A :
       1007379606734.

[30]   Zoubin Ghahramani. "Probabilistic machine learning and artifi-
       cial intelligence". In: *Nature* 521 (7553 May 2015), pp. 452–459.
       ISSN: 0028-0836. DOI: 10.1038/nature14541.

[31]   Yarin Gal. "Uncertainty in Deep Learning". University of
       Cambridge, 2016. URL: http://mlg.eng.cam.ac.uk/yarin/
       blog_2248.html#chapter_1.

[32]   Alex Kendall and Yarin Gal. "What Uncertainties Do We
       Need in Bayesian Deep Learning for Computer Vision?" In:
       Advances in Neural Information Processing Systems (NIPS).
       Curran Associates, Inc., 2017. URL: http://papers.nips.cc/
       paper/7141-what-uncertainties-do-we-need-in-bayesian-
       deep-learning-for-computer-vision.pdf.

[33]   David J. C. MacKay. "A Practical Bayesian Framework for
       Backpropagation Networks". In: *Neural Computation* 4 (3 May
       1992), pp. 448–472. ISSN: 0899-7667. DOI: 10.1162/neco.1992.
       4.3.448.

[34]   Radford M. Neal. "Bayesian learning for neural networks".
       Univeristy of Toronto, 1995.

[35]   Alex Graves. "Practical Variational Inference for Neural Net-
       works". In: Advances in Neural Information Processing Systems
       (NIPS). Ed. by J Shawe-Taylor et al. Vol. 24. Curran Associates,
       Inc., 2011. URL: https://proceedings.neurips.cc/paper/
       2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf.

[36]   Diederik P Kingma and Max Welling. "Auto-Encoding Varia-
       tional Bayes". In: *arXiv preprint* (Dec. 2013). URL: http://
       arxiv.org/abs/1312.6114.

[37] "Weight Uncertainty in Neural Network". In: Proceedings of the 32nd International Conference on Machine Learning. Ed. by Francis Bach and David Blei. Vol. 37. PMLR, May 2015, pp. 1613–1622. URL: http://proceedings.mlr.press/v37/blundell15.html.

[38] Yarin Gal and Zoubin Ghahramani. "Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference". In: arXiv preprint (June 2015). URL: http://arxiv.org/abs/1506.02158.

[39] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: Proceedings of the 33rd International Conference on Machine Learning. Ed. by Kilian Q. Balcan Maria FlorinaWeinberger. PMLR, June 2016, pp. 1050–1059. URL: http://proceedings.mlr.press/v48/gal16.pdf.

[40] Yann Lecun. "Generalization and network design strategies". In: Connectionism in perspective 19 (1989), pp. 143–155.

[41] Kunihiko Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: Biological Cybernetics 36 (4 Apr. 1980). ISSN: 0340-1200. DOI: 10.1007/BF00344251.

[42] Yann LeCun et al. "Gradient-Based Learning Applied to Document Recognition". In: Proceedings of the IEEE 86 (11 1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[43] "ImageNet Large Scale Visual Recognition Challenge". In: International Journal of Computer Vision 115 (3 Dec. 2015), pp. 211–252. ISSN: 15731405. DOI: 10.1007/s11263-015-0816-y.

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: Advances in neural information processing systems 25 (2012), pp. 1097–1105.

[45] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: arXiv preprint (2014), pp. 1–14. URL: http://arxiv.org/abs/1409.1556.

[46] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-December (Dec. 2016), pp. 770–778. ISSN: 10636919. DOI: 10.1109/CVPR.2016.90.

[47] Gao Huang et al. "Densely Connected Convolutional Networks". In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, July 2017, pp. 2261–2269. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.243.

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9351 (2015), pp. 234–241. ISSN: 16113349. DOI: 10.1007/978-3-319-24574-4_28.

[49] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3431–3440.

[50] Frederik Maes et al. "The Role of Medical Image Computing and Machine Learning in Healthcare". In: *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*. Ed. by Erik R. Ranschaert, Sergey Morozov, and Paul R. Algra. Springer International Publishing, 2019. DOI: 10.1007/978-3-319-94878-2_2.

[51] Kenneth Clark et al. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository". In: *Journal of Digital Imaging* 26 (6 Dec. 2013), pp. 1045–1057. ISSN: 0897-1889. DOI: 10.1007/s10278-013-9622-7.

[52] Paul Suetens. *Fundamentals of Medical Imaging*. Cambridge University Press, 2009. ISBN: 9780511596803. DOI: 10.1017/CBO9780511596803.

[53] P. C. LAUTERBUR. "Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance". In: *Nature* 242 (5394 Mar. 1973), pp. 190–191. ISSN: 0028-0836. DOI: 10.1038/242190a0.

[54] David E. Kuhl and Roy Q. Edwards. "Image Separation Radioisotope Scanning". In: *Radiology* 80 (4 Apr. 1963), pp. 653–662. ISSN: 0033-8419. DOI: 10.1148/80.4.653.

[55] Simon R. Cherry, James A. Sorenson, and Michael E. Phelps. *Physics in nuclear medicine.* Elsevier/Saunders, 2012, p. 523. ISBN: 9781416051985.

[56] Shan Tong, Adam M Alessio, and Paul E Kinahan. "Image reconstruction for PET/CT scanners: past achievements and future challenges." In: *Imaging in medicine* 2 (5 Oct. 2010), pp. 529–545. ISSN: 1755-5191. DOI: 10.2217/iim.10.49.

[57] Kuang Gong, Simon R Cherry, and Jinyi Qi. "On the assessment of spatial resolution of PET systems with iterative image reconstruction". In: *Physics in Medicine and Biology* 61 (5 Mar. 2016), N193–N202. ISSN: 0031-9155. DOI: 10.1088/0031-9155/61/5/N193.

[58] Eric Berg and Simon R. Cherry. "Innovations in Instrumentation for Positron Emission Tomography". In: *Seminars in Nuclear Medicine* 48 (4 July 2018), pp. 311–331. ISSN: 0001-2998. DOI: 10.1053/J.SEMNUCLMED.2018.02.006.

[59] Robert S. Miyaoka et al. "Resolution properties of a prototype continuous miniature crystal element (cMiCE) scanner". In: *IEEE Transactions on Nuclear Science* 58 (5 PART 1 2011), pp. 2244–2249. ISSN: 00189499. DOI: 10.1109/TNS.2011.2165296.

[60] Giacomo Borghi, Valerio Tabacchini, and Dennis R Schaart. "Towards monolithic scintillator based TOF-PET systems: practical methods for detector calibration and operation". In: *Physics in Medicine and Biology* 61 (13 July 2016), pp. 4904–4928. ISSN: 0031-9155. DOI: 10.1088/0031-9155/61/13/4904.

[61] Andrea Gonzalez-Montoro et al. "Detector block performance based on a monolithic LYSO crystal using a novel signal multiplexing method". In: *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 912 (February 2018), pp. 372–377. ISSN: 01689002. DOI: 10.1016/j.nima.2017.10.098.

[62] Andrea Gonzalez-Montoro et al. "Validation of photon collimation techniques for monolithic PET detector calibration". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* (Dec. 2020), pp. 1–1. ISSN: 2469-7311. DOI: 10.1109/trpms.2020.3043397.

[63] Marnix C Maas et al. "Monolithic scintillator PET detectors with intrinsic depth-of-interaction correction". In: *Physics in Medicine and Biology* 54 (7 Apr. 2009), pp. 1893–1908. ISSN: 0031-9155. DOI: 10.1088/0031-9155/54/7/003.

[64] Valerio Tabacchini, Giacomo Borghi, and Dennis R. Schaart. "Time-based position estimation in monolithic scintillator detectors". In: *Physics in Medicine and Biology* 60 (14 July 2015), pp. 5513–5525. ISSN: 13616560. DOI: 10.1088/0031-9155/60/14/5513.

[65] M. Pizzichemi et al. "A new method for depth of interaction determination in PET detectors". In: *Physics in Medicine and Biology* 61 (12 June 2016), pp. 4679–4698. ISSN: 13616560. DOI: 10.1088/0031-9155/61/12/4679.

[66] Antonio J. Gonález et al. "A PET Design Based on SiPM and Monolithic LYSO Crystals: Performance Evaluation". In: *IEEE Transactions on Nuclear Science* 63 (5 2016), pp. 2471–2477. ISSN: 00189499. DOI: 10.1109/TNS.2016.2522179.

[67] Srilalan Krishnamoorthy et al. "Performance evaluation of the MOLECUBES β-CUBE a high spatial resolution and high sensitivity small animal PET scanner utilizing monolithic LYSO scintillation detectors". In: *Physics in Medicine and Biology* 63 (15 July 2018), p. 155013. ISSN: 1361-6560. DOI: 10.1088/1361-6560/aacec3.

[68] Giacomo Borghi et al. "Experimental Validation of an Efficient Fan-Beam Calibration Procedure for k-Nearest Neighbor Position Estimation in Monolithic Scintillator Detectors". In: *IEEE Transactions on Nuclear Science* 62 (1 Feb. 2015), pp. 57–67. ISSN: 0018-9499. DOI: 10.1109/TNS.2014.2375557.

[69] Florian Muller et al. "Gradient Tree Boosting-Based Positioning Method for Monolithic Scintillator Crystals in Positron Emission Tomography". In: *IEEE Transactions on Radiation and Plasma*

Medical Sciences 2 (5 Sept. 2018), pp. 411–421. ISSN: 2469-7311.
DOI: `10.1109/TRPMS.2018.2837738`.

[70] James S. Duncan, Michael F. Insana, and Nicholas Ayache.
"Biomedical Imaging and Analysis in the Age of Big Data and
Deep Learning". In: *Proceedings of the IEEE* 108 (1 Jan. 2020),
pp. 3–10. ISSN: 15582256. DOI: `10.1109/JPROC.2019.2956422`.

[71] Ruud J. G. van Sloun, Regev Cohen, and Yonina C. Eldar. "Deep
Learning in Ultrasound Imaging". In: *Proceedings of the IEEE* 108
(1 Jan. 2020), pp. 11–29. ISSN: 0018-9219. DOI: `10.1109/JPROC.`
`2019.2932116`.

[72] Kuang Gong et al. "Machine Learning in PET: From Photon
Detection to Quantitative Image Reconstruction". In: *Proceedings
of the IEEE* 108 (1 Jan. 2020), pp. 51–68. ISSN: 0018-9219. DOI:
`10.1109/JPROC.2019.2936809`.

[73] Akshay S. Chaudhari et al. "Prospective Deployment of Deep
Learning in MRI : A Framework for Important Considerations,
Challenges, and Recommendations for Best Practices". In:
Journal of Magnetic Resonance Imaging (Aug. 2020), jmri.27331.
ISSN: 1053-1807. DOI: `10.1002/jmri.27331`.

[74] S. Kevin Zhou et al. "A review of deep learning in medical
imaging: Image traits, technology trends, case studies with
progress highlights, and future promises". In: *Proceeding of the
IEEE* (Aug. 2021). ISSN: 23318422. DOI: `10.1109/JPROC.2021.`
`3054390`.

[75] Artem Zatcepin et al. "Improving depth-of-interaction resolution
in pixellated PET detectors using neural networks". In: *Physics in
Medicine and Biology* 65 (17 Aug. 2020), p. 175017. ISSN: 1361-
6560. DOI: `10.1088/1361-6560/ab9efc`.

[76] Robert S. Miyaoka et al. "Calibration Procedure for a Continuous
Miniature Crystal Element (cMiCE) Detector". In: *IEEE
Transactions on Nuclear Science* 57 (3 June 2010), pp. 1023–1028.
ISSN: 0018-9499. DOI: `10.1109/TNS.2010.2043261`.

[77] Samuel Españ a et al. "DigiPET: sub-millimeter spatial resolution
small-animal PET imaging using thin monolithic scintillators".
In: *Physics in Medicine and Biology* 59 (13 July 2014), pp. 3405–
3420. ISSN: 0031-9155. DOI: `10.1088/0031-9155/59/13/3405`.

[78]   L. A. Pierce et al. "Characterization of highly multiplexed monolithic PET / gamma camera detector modules". In: *Physics in Medicine and Biology* 63 (7 Mar. 2018). ISSN: 13616560. DOI: 10.1088/1361-6560/aab380.

[79]   H. T. van Dam et al. "Improved Nearest Neighbor Methods for Gamma Photon Interaction Position Determination in Monolithic Scintillator PET Detectors". In: *IEEE Transactions on Nuclear Science* 58 (5 Oct. 2011), pp. 2139–2147. ISSN: 0018-9499. DOI: 10.1109/TNS.2011.2150762.

[80]   Mariele Stockhoff, Roel Van Holen, and Stefaan Vandenberghe. "Optical simulation study on the spatial resolution of a thick monolithic PET detector". In: *Physics in Medicine and Biology* 64 (19 Sept. 2019), p. 195003. ISSN: 1361-6560. DOI: 10.1088/1361-6560/ab3b83.

[81]   Florian Muller et al. "A Novel DOI Positioning Algorithm for Monolithic Scintillator Crystals in PET Based on Gradient Tree Boosting". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 3 (4 July 2019), pp. 465–474. ISSN: 2469-7311. DOI: 10.1109/TRPMS.2018.2884320.

[82]   Peter Bruyndonckx et al. "Neural network-based position estimators for PET detectors using monolithic LSO blocks". In: *IEEE Transactions on Nuclear Science* 51 (5 II Oct. 2004), pp. 2520–2525. ISSN: 00189499. DOI: 10.1109/TNS.2004.835782.

[83]   Y Wang et al. "3D position estimation using an artificial neural network for a continuous scintillator PET detector". In: *Physics in Medicine and Biology* 58 (5 Mar. 2013), pp. 1375–1390. ISSN: 0031-9155. DOI: 10.1088/0031-9155/58/5/1375.

[84]   A Iborra et al. "Ensemble of neural networks for 3D position estimation in monolithic PET detectors". In: *Physics in Medicine and Biology* 64 (19 Oct. 2019), p. 195010. ISSN: 1361-6560. DOI: 10.1088/1361-6560/ab3b86.

[85]   J. M. Monzo et al. "Digital Signal Processing Techniques to Improve Time Resolution in Positron Emission Tomography". In: *IEEE Transactions on Nuclear Science* 58 (4 Aug. 2011), pp. 1613–1620. ISSN: 0018-9499. DOI: 10.1109/TNS.2011.2140382.

[86]   Eric Berg and Simon R Cherry. "Using convolutional neural net-
       works to estimate time-of-flight from PET detector waveforms".
       In: *Physics in Medicine and Biology* 63 (2 Jan. 2018), 02LT01.
       ISSN: 1361-6560. DOI: 10.1088/1361-6560/aa9dc5.

[87]   Ge Wang et al. *Machine Learning for Tomographic Imaging*. IOP
       Publishing, Dec. 2019. ISBN: 978-0-7503-2216-4. DOI: 10.1088/
       978-0-7503-2216-4.

[88]   Hai-Miao Zhang and Bin Dong. "A Review on Deep Learning
       in Medical Image Reconstruction". In: *Journal of the Operations
       Research Society of China* 8 (2 June 2020), pp. 311–340. ISSN:
       2194-668X. DOI: 10.1007/s40305-019-00287-4.

[89]   Andrew J. Reader et al. "Deep Learning for PET Image
       Reconstruction". In: *IEEE Transactions on Radiation and Plasma
       Medical Sciences* 5 (1 Jan. 2021), pp. 1–25. ISSN: 2469-7311. DOI:
       10.1109/TRPMS.2020.3014786.

[90]   Bo Zhu et al. "Image reconstruction by domain-transform
       manifold learning". In: *Nature* 555 (7697 Mar. 2018), pp. 487–492.
       ISSN: 0028-0836. DOI: 10.1038/nature25988.

[91]   Ida Häggström et al. "DeepPET: A deep encoder-decoder
       network for directly solving the PET image reconstruction inverse
       problem". In: *Medical Image Analysis* 54 (May 2019), pp. 253–262.
       ISSN: 1361-8415. DOI: 10.1016/J.MEDIA.2019.03.013.

[92]   Zhanli Hu et al. "DPIR-Net: Direct PET Image Reconstruction
       Based on the Wasserstein Generative Adversarial Network". In:
       IEEE Transactions on Radiation and Plasma Medical Sciences 5
       (1 Jan. 2021), pp. 35–43. ISSN: 2469-7311. DOI: 10.1109/TRPMS.
       2020.2995717.

[93]   Xin Yi, Ekta Walia, and Paul Babyn. "Generative adversarial
       network in medical imaging: A review". In: *Medical Image
       Analysis* 58 (Dec. 2019), p. 101552. ISSN: 1361-8415. DOI: 10.
       1016/J.MEDIA.2019.101552.

[94]   Tobias Wurfl et al. "Deep Learning Computed Tomography:
       Learning Projection-Domain Weights From Image Domain in
       Limited Angle Problems". In: *IEEE Transactions on Medical
       Imaging* 37 (6 June 2018), pp. 1454–1463. ISSN: 0278-0062. DOI:
       10.1109/TMI.2018.2833499.

[95] Karol Gregor and Yann Lecun. "Learning Fast Approximations of Sparse Coding". In: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10. Omnipress, Madison, WI, USA, 2010, pp. 399–406. URL: http://yann.lecun.org/exdb/publis/pdf/gregor-icml-10.pdf.

[96] Taejoon Eo et al. "KIKI-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images". In: Magnetic Resonance in Medicine 80 (5 Nov. 2018), pp. 2188–2201. ISSN: 0740-3194. DOI: 10.1002/mrm.27201.

[97] Jian Zhang and Bernard Ghanem. "ISTA-Net: Interpretable Optimization-Inspired Deep Network for Image Compressive Sensing". In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, June 2018, pp. 1828–1837. ISBN: 978-1-5386-6420-9. DOI: 10.1109/CVPR.2018.00196.

[98] Hemant K. Aggarwal, Merry P. Mani, and Mathews Jacob. "MoDL: Model-Based Deep Learning Architecture for Inverse Problems". In: IEEE Transactions on Medical Imaging 38 (2 Feb. 2019), pp. 394–405. ISSN: 0278-0062. DOI: 10.1109/TMI.2018.2865356.

[99] Chen Qin et al. "Convolutional Recurrent Neural Networks for Dynamic MR Image Reconstruction". In: IEEE Transactions on Medical Imaging 38 (1 Jan. 2019), pp. 280–290. ISSN: 0278-0062. DOI: 10.1109/TMI.2018.2863670.

[100] Dong Liang et al. "Deep MRI Reconstruction: Unrolled Optimization Algorithms Meet Neural Networks". In: arXiv preprint (July 2019). URL: http://arxiv.org/abs/1907.11711.

[101] Yan Yang et al. "ADMM-CSNet: A Deep Learning Approach for Image Compressive Sensing". In: IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (3 Mar. 2020), pp. 521–538. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2018.2883941.

[102] Dufan Wu, Kyungsang Kim, and Quanzheng Li. "Computationally efficient deep neural network for computed tomography image reconstruction". In: Medical Physics 46 (11 Nov. 2019), pp. 4763–4776. ISSN: 0094-2405. DOI: 10.1002/mp.13627.

[103] Haimiao Zhang et al. "MetaInv-Net: Meta Inversion Network for Sparse View CT Image Reconstruction". In: *IEEE Transactions on Medical Imaging* 40 (2 Feb. 2021), pp. 621–634. ISSN: 0278-0062. DOI: 10.1109/TMI.2020.3033541.

[104] Kuang Gong et al. "MAPEM-Net: an unrolled neural network for Fully 3D PET image reconstruction". In: 15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine. Ed. by Samuel Matej and Scott D. Metzler. Vol. 11072. SPIE, May 2019, p. 102. ISBN: 9781510628373. DOI: 10.1117/12.2534904.

[105] Vishal Monga, Yuelong Li, and Yonina C. Eldar. "Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing". In: *arXiv preprint* (Dec. 2019). URL: http://arxiv.org/abs/1912.10557.

[106] Yan Wang et al. "3D conditional generative adversarial networks for high-quality PET image estimation at low dose". In: *NeuroImage* 174 (July 2018), pp. 550–562. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2018.03.045.

[107] Jelmer M. Wolterink et al. "Generative Adversarial Networks for Noise Reduction in Low-Dose CT". In: *IEEE Transactions on Medical Imaging* 36 (12 Dec. 2017), pp. 2536–2545. ISSN: 0278-0062. DOI: 10.1109/TMI.2017.2708987.

[108] Hu Chen et al. "Low-dose CT via convolutional neural network." In: *Biomedical optics express* 8 (2 Feb. 2017), pp. 679–694. ISSN: 2156-7085. DOI: 10.1364/BOE.8.000679.

[109] Qingsong Yang et al. "Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss". In: *IEEE Transactions on Medical Imaging* 37 (6 June 2018), pp. 1348–1357. ISSN: 0278-0062. DOI: 10.1109/TMI.2018.2827462.

[110] Kevin T. Chen et al. "Ultra Low-Dose ^{18}F-Florbetaben Amyloid PET Imaging Using Deep Learning with Multi-Contrast MRI Inputs". In: *Radiology* 290 (3 Mar. 2019), pp. 649–656. DOI: 10.1148/radiol.2018180940.

[111]   Sydney Kaplan and Yang-Ming Zhu. "Full-Dose PET Image Estimation from Low-Dose PET Image Using Deep Learning: a Pilot Study". In: *Journal of Digital Imaging* 32 (5 Oct. 2019), pp. 773–778. ISSN: 0897-1889. DOI: 10.1007/s10278-018-0150-3.

[112]   Katia Katsari et al. "Artificial intelligence for reduced dose 18F-FDG PET examinations: a real-world deployment through a standardized framework and business case assessment". In: *EJNMMI Physics* 8 (1 Dec. 2021), 25–undefined. ISSN: 21977364. DOI: 10.1186/s40658-021-00374-7.

[113]   Huanyu Luo et al. "Deep learning based methods may minimize GBCA dosage in brain MRI". In: *European Radiology* (Mar. 2021). ISSN: 0938-7994. DOI: 10.1007/s00330-021-07848-3.

[114]   Dongwook Lee et al. "Deep Residual Learning for Accelerated MRI Using Magnitude and Phase Networks". In: *IEEE Transactions on Biomedical Engineering* 65 (9 Sept. 2018), pp. 1985–1995. ISSN: 0018-9294. DOI: 10.1109/TBME.2018.2821699.

[115]   Chang Min Hyun et al. "Deep learning for undersampled MRI reconstruction". In: *Physics in Medicine and Biology* 63 (13 June 2018), p. 135007. ISSN: 1361-6560. DOI: 10.1088/1361-6560/aac71a.

[116]   Yuhua Chen et al. "Efficient and Accurate MRI Super-Resolution Using a Generative Adversarial Network and 3D Multi-level Densely Connected Network". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018. Lecture Notes in Computer Science*. Vol. 11070. Springer, Cham, Sept. 2018, pp. 91–99. DOI: 10.1007/978-3-030-00928-1_11.

[117]   Akshay S. Chaudhari et al. "Super-resolution musculoskeletal MRI using deep learning". In: *Magnetic Resonance in Medicine* 80 (5 Nov. 2018), pp. 2139–2154. ISSN: 0740-3194. DOI: 10.1002/mrm.27178.

[118]   Chenyu You et al. "CT Super-Resolution GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE)". In: *IEEE Transactions on Medical Imaging* 39 (1 Jan. 2020), pp. 188–203. ISSN: 0278-0062. DOI: 10.1109/TMI.2019.2922960.

[119] Tzu-An Song et al. "Super-Resolution PET Imaging Using Convolutional Neural Networks". In: *IEEE Transactions on Computational Imaging* 6 (2020), pp. 518–528. ISSN: 2333-9403. DOI: 10.1109/TCI.2020.2964229.

[120] Kuang Gong et al. "PET Image Denoising Using a Deep Neural Network Through Fine Tuning". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 3 (2 Mar. 2019), pp. 153–161. ISSN: 2469-7311. DOI: 10.1109/TRPMS.2018.2877644.

[121] Maosong Ran et al. "Denoising of 3D magnetic resonance images using a residual encoder-decoder Wasserstein generative adversarial network". In: *Medical Image Analysis* 55 (July 2019), pp. 165–180. ISSN: 1361-8415. DOI: 10.1016/J.MEDIA.2019.05.001.

[122] Martijn M A Dietze et al. "Accelerated SPECT image reconstruction with FBP and an image enhancement convolutional neural network." In: *EJNMMI physics* 6 (1 July 2019), p. 14. ISSN: 2197-7364. DOI: 10.1186/s40658-019-0252-0.

[123] Chih-Chieh Liu and Jinyi Qi. "Higher SNR PET image prediction using a deep learning model and MRI image". In: *Physics in Medicine and Biology* 64 (11 May 2019), p. 115004. ISSN: 1361-6560. DOI: 10.1088/1361-6560/ab0dc0.

[124] Jianan Cui et al. "PET image denoising using unsupervised deep learning". In: *European Journal of Nuclear Medicine and Molecular Imaging* 46 (13 Dec. 2019), pp. 2780–2789. ISSN: 1619-7070. DOI: 10.1007/s00259-019-04468-4.

[125] Julian Krebs et al. "Robust Non-rigid Registration Through Agent-Based Action Learning". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017. Lecture Notes in Computer Science*. Vol. 10433. Springer, Cham, Sept. 2017, pp. 344–352. DOI: 10.1007/978-3-319-66182-7_40.

[126] Guha Balakrishnan et al. "VoxelMorph: A Learning Framework for Deformable Medical Image Registration". In: *IEEE Transactions on Medical Imaging* 38 (8 Aug. 2019), pp. 1788–1800. ISSN: 0278-0062. DOI: 10.1109/TMI.2019.2897538.

[127] Grant Haskins et al. "Learning deep similarity metric for 3D MR-TRUS image registration". In: *International Journal of Computer Assisted Radiology and Surgery* 14 (3 Mar. 2019), pp. 417–425. ISSN: 1861-6410. DOI: 10.1007/s11548-018-1875-7.

[128] Seyed Sadegh Mohseni Salehi et al. "Real-Time Deep Pose Estimation With Geodesic Loss for Image-to-Template Rigid Registration". In: *IEEE Transactions on Medical Imaging* 38 (2 Feb. 2019), pp. 470–481. ISSN: 0278-0062. DOI: 10.1109/TMI.2018.2866442.

[129] Grant Haskins, Uwe Kruger, and Pingkun Yan. "Deep learning in medical image registration: a survey". In: *Machine Vision and Applications* 31 (1-2 Feb. 2020), p. 8. ISSN: 0932-8092. DOI: 10.1007/s00138-020-01060-x.

[130] Yabo Fu et al. "Deep learning in medical image registration: a review". In: *Physics in Medicine and Biology* 65 (20 Oct. 2020), 20TR01. ISSN: 1361-6560. DOI: 10.1088/1361-6560/ab843e.

[131] Jelmer M. Wolterink et al. "Deep MR to CT Synthesis Using Unpaired Data". In: *Simulation and Synthesis in Medical Imaging. SASHIMI 2017. Lecture Notes in Computer Science.* Vol. 10557. Springer, Cham, 2017, pp. 14–23. DOI: 10.1007/978-3-319-68127-6_2.

[132] Xiao Han. "MR-based synthetic CT generation using a deep convolutional neural network method". In: *Medical Physics* 44 (4 Apr. 2017), pp. 1408–1419. ISSN: 00942405. DOI: 10.1002/mp.12155.

[133] Dong Nie et al. "Medical Image Synthesis with Context-Aware Generative Adversarial Networks". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017. Lecture Notes in Computer Science.* Vol. 10435. Springer, Cham, Sept. 2017, pp. 417–425. DOI: 10.1007/978-3-319-66179-7_48.

[134] Dong Nie et al. "Medical Image Synthesis with Deep Convolutional Adversarial Networks". In: *IEEE Transactions on Biomedical Engineering* 65 (12 Dec. 2018), pp. 2720–2730. ISSN: 0018-9294. DOI: 10.1109/TBME.2018.2814538.

[135]   Salman UH. Dar et al. "Image Synthesis in Multi-Contrast MRI
        With Conditional Generative Adversarial Networks". In: *IEEE
        Transactions on Medical Imaging* 38 (10 Oct. 2019), pp. 2375–
        2388. ISSN: 0278-0062. DOI: 10.1109/TMI.2019.2901750.

[136]   Karim Armanious et al. "MedGAN: Medical image translation
        using GANs". In: *Computerized Medical Imaging and Graphics*
        79 (Jan. 2020), p. 101684. ISSN: 0895-6111. DOI: 10.1016/J.
        COMPMEDIMAG.2019.101684.

[137]   Lennart B.O. Jans et al. "MRI-based Synthetic CT in the
        Detection of Structural Lesions in Patients with Suspected
        Sacroiliitis: Comparison with MRI". In: *Radiology* 298 (2 Feb.
        2021), pp. 343–349. ISSN: 15271315. DOI: 10.1148/radiol.
        2020201537.

[138]   Ian J Goodfellow et al. "Generative Adversarial Nets". In:
        Proceedings of the 27th International Conference on Neural
        Information Processing Systems - Volume 2. MIT Press, 2014,
        pp. 2672–2680.

[139]   Phillip Isola et al. "Image-to-Image Translation with Conditional
        Adversarial Networks". In: 2017 IEEE Conference on Computer
        Vision and Pattern Recognition (CVPR). IEEE, July 2017,
        pp. 5967–5976. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.
        2017.632.

[140]   Jun-Yan Zhu et al. "Unpaired Image-to-Image Translation Using
        Cycle-Consistent Adversarial Networks". In: IEEE, Oct. 2017,
        pp. 2242–2251. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.
        2017.244.

[141]   Maciej A. Mazurowski et al. "Deep learning in radiology: An
        overview of the concepts and a survey of the state of the art
        with focus on MRI". In: *Journal of Magnetic Resonance Imaging*
        49 (4 Apr. 2019), pp. 939–954. ISSN: 10531807. DOI: 10.1002/
        jmri.26534.

[142]   Erik R Ranschaert, Sergey Morozov, and Paul R Algra. *Artificial
        Intelligence in Medical Imaging*. Ed. by Erikk R. Ranschaert,
        Sergey Morozov, and Paul R. Algra. Vol. 1. Springer, Cham, 2019.
        ISBN: 978-3-319-94878-2. DOI: https://doi.org/10.1007/978-
        3-319-94878-2.

[143] Daniel Rueckert and Julia A. Schnabel. "Model-Based and Data-Driven Strategies in Medical Image Computing". In: *Proceedings of the IEEE* 108 (1 Jan. 2020), pp. 110–124. ISSN: 15582256. DOI: 10.1109/JPROC.2019.2943836.

[144] A. Ibrahim et al. "Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework". In: *Methods* 188 (Apr. 2021), pp. 20–29. ISSN: 1046-2023. DOI: 10.1016/J.YMETH.2020.05.022.

[145] Virendra Kumar et al. "Radiomics: the process and the challenges". In: *Magnetic Resonance Imaging* 30 (9 Nov. 2012). Good overview of radiomics process, pp. 1234–1248. ISSN: 0730725X. DOI: 10.1016/j.mri.2012.06.010.

[146] Mohammad Hesam Hesamian et al. "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges". In: *Journal of Digital Imaging* 32 (4 Aug. 2019), pp. 582–596. ISSN: 1618727X. DOI: 10.1007/s10278-019-00227-x.

[147] Nima Tajbakhsh et al. "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation". In: *Medical Image Analysis* 63 (July 2020), p. 101693. ISSN: 1361-8415. DOI: 10.1016/J.MEDIA.2020.101693.

[148] Saeid Asgari Taghanaki et al. "Deep semantic segmentation of natural and medical images: a review". In: *Artificial Intelligence Review* 54 (June 2021), pp. 137–178. ISSN: 0269-2821. DOI: 10.1007/s10462-020-09854-1.

[149] Fausto Milletari, Nassir Navab, and Seyed Ahmad Ahmadi. "V-Net: Fully convolutional neural networks for volumetric medical image segmentation". In: Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016. Institute of Electrical and Electronics Engineers Inc., Dec. 2016, pp. 565–571. ISBN: 9781509054077. DOI: 10.1109/3DV.2016.79.

[150] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18 (2 Feb. 2021), pp. 203–211. ISSN: 15487105. DOI: 10.1038/s41592-020-01008-z.

[151] Amber L. Simpson et al. "A large annotated medical image
       dataset for the development and evaluation of segmentation
       algorithms". In: *arXiv* (Feb. 2019). URL: http://arxiv.org/
       abs/1902.09063.

[152] Bo Liu et al. "Evolving the pulmonary nodules diagnosis from
       classical approaches to deep learning-aided decision support: three
       decades' development course and future prospect". In: *Journal
       of Cancer Research and Clinical Oncology* 146 (1 Jan. 2020),
       pp. 153–185. ISSN: 14321335. DOI: 10.1007/s00432-019-03098-
       5.

[153] Arnaud Arindra Adiyoso Setio et al. "Validation, comparison,
       and combination of algorithms for automatic detection of
       pulmonary nodules in computed tomography images: The
       LUNA16 challenge". In: *Medical Image Analysis* 42 (Dec. 2017),
       pp. 1–13. ISSN: 13618423. DOI: 10.1016/j.media.2017.06.015.

[154] SG Armato III et al. "Data From LIDC-IDRI". In: *The Cancer
       Imaging Archive* (2015). DOI: http://doi.org/10.7937/K9/
       TCIA.2015.LO9QL9SX.

[155] Samuel G. Armato et al. "The Lung Image Database Consortium
       (LIDC) and Image Database Resource Initiative (IDRI): A
       Completed Reference Database of Lung Nodules on CT Scans".
       In: *Medical Physics* 38 (2 Jan. 2011). ISSN: 00942405. DOI: 10.
       1118/1.3528204.

[156] Fangzhou Liao et al. "Evaluate the Malignancy of Pulmonary
       Nodules Using the 3-D Deep Leaky Noisy-OR Network". In: *IEEE
       Transactions on Neural Networks and Learning Systems* 30 (11
       Nov. 2019), pp. 3484–3495. ISSN: 21622388. DOI: 10.1109/TNNLS.
       2019.2892409.

[157] Diego Ardila et al. "End-to-end lung cancer screening with
       three-dimensional deep learning on low-dose chest computed
       tomography". In: *Nature Medicine* 25 (6 June 2019), pp. 954–961.
       ISSN: 1546170X. DOI: 10.1038/s41591-019-0447-x.

[158] Joao Carreira and Andrew Zisserman. "Quo Vadis, action
       recognition? A new model and the kinetics dataset". In:
       Proceedings - 30th IEEE Conference on Computer Vision
       and Pattern Recognition, CVPR 2017. Vol. 2017-January.
       Institute of Electrical and Electronics Engineers Inc., Nov. 2017,

pp. 4724–4733. ISBN: 9781538604571. DOI: 10.1109/CVPR.2017.502.

[159] The National Lung Screening Trial Research Team. "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening". In: *New England Journal of Medicine* 365 (5 Aug. 2011). ISSN: 0028-4793. DOI: 10.1056/NEJMoa1102873.

[160] Edwin J. R. van Beek and John T. Murchison. "Artificial Intelligence and Computer-Assisted Evaluation of Chest Pathology". In: *Artificial Intelligence in Medical Imaging*. Ed. by Erik R. Ranschaert, Sergey Morozov, and Paul R. Algra. Springer International Publishing, 2019. DOI: 10.1007/978-3-319-94878-2_12.

[161] Keelin Murphy et al. "COVID-19 on chest radiographs: A multireader evaluation of an artificial intelligence system". In: *Radiology* 296 (3 Sept. 2020), E166–E172. ISSN: 15271315. DOI: 10.1148/radiol.2020201874.

[162] Dinggang Shen et al. "Guest Editorial: Special Issue on Imaging-Based Diagnosis of COVID-19". In: *IEEE Transactions on Medical Imaging* 39 (8 Aug. 2020), pp. 2569–2571. ISSN: 1558254X. DOI: 10.1109/TMI.2020.3008025.

[163] Guillaume Chassagnon et al. "AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia". In: *Medical Image Analysis* 67 (Jan. 2021), 101860–undefined. ISSN: 13618423. DOI: 10.1016/j.media.2020.101860.

[164] Kai Gao et al. "Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images". In: *Medical Image Analysis* 67 (Jan. 2021), 101836–undefined. ISSN: 13618423. DOI: 10.1016/j.media.2020.101836.

[165] Nikolas Lessmann et al. "Automated assessment of COVID-19 reporting and data system and chest CT severity scores in patients suspected of having COVID-19 using artificial intelligence". In: *Radiology* 298 (1 2021), E18–E28. ISSN: 15271315. DOI: 10.1148/RADIOL.2020202439.

[166] Feng Shi et al. "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19". In: *IEEE Reviews in Biomedical Engineering* 14 (2021), pp. 4–15. ISSN: 19411189. DOI: 10.1109/RBME.2020.2987975.

[167] Hayit Greenspan et al. "Position paper on COVID-19 imaging and AI: From the clinical needs and technological challenges to initial AI solutions at the lab and national level towards a new era for AI in healthcare". In: *Medical Image Analysis* 66 (Dec. 2020). ISSN: 13618423. DOI: 10.1016/j.media.2020.101800.

[168] Sofie Tilborghs et al. "Comparative study of deep learning methods for the automatic segmentation of lung, lesion and lesion type in CT scans of COVID-19 patients". In: *ArXiv.org* (2020). URL: https://arxiv.org/abs/2007.15546.

[169] Mathias Prokop et al. "CO-RADS: A Categorical CT Assessment Scheme for Patients Suspected of Having COVID-19-Definition and Evaluation". In: *Radiology* 296 (2 Aug. 2020), E97–E104. ISSN: 15271315. DOI: 10.1148/radiol.2020201473.

[170] Weiyi Xie et al. "Relational Modeling for Robust and Efficient Pulmonary Lobe Segmentation in CT Scans". In: *IEEE Transactions on Medical Imaging* 39 (8 Aug. 2020), pp. 2664–2675. ISSN: 1558254X. DOI: 10.1109/TMI.2020.2995108.

[171] Hugh Harvey et al. "Deep Learning in Breast Cancer Screening". In: *Artificial Intelligence in Medical Imaging*. Ed. by Erik R. Ranschaert, Sergey Morozov, and Paul R. Algra. Springer International Publishing, 2019. DOI: 10.1007/978-3-319-94878-2_14.

[172] Hugh Harvey et al. "The Role of Deep Learning in Breast Screening". In: *Current Breast Cancer Reports* 11 (1 Mar. 2019), pp. 17–22. ISSN: 19434596. DOI: 10.1007/s12609-019-0301-7.

[173] Rebecca Sawyer Lee et al. "Curated Breast Imaging Subset of DDSM [Dataset]". In: *The Cancer Imaging Archive* (2016).

[174] Rebecca Sawyer Lee et al. "A curated mammography data set for use in computer-aided detection and diagnosis research". In: *Scientific Data* 4 (Dec. 2017), 170177–undefined. ISSN: 20524463. DOI: 10.1038/sdata.2017.177.

[175] Thomas Schaffter et al. "Evaluation of Combined Artificial
       Intelligence and Radiologist Assessment to Interpret Screening
       Mammograms". In: *JAMA network open* 3 (3 Mar. 2020),
       e200265. ISSN: 25743805. DOI: 10.1001/jamanetworkopen.2020.
       0265.

[176] Thijs Kooi et al. "Large scale deep learning for computer aided
       detection of mammographic lesions". In: *Medical Image Analysis*
       35 (Jan. 2017), pp. 303–312. ISSN: 13618423. DOI: 10.1016/j.
       media.2016.07.007.

[177] Scott Mayer McKinney et al. "International evaluation of an AI
       system for breast cancer screening". In: *Nature* 577 (7788 Jan.
       2020), pp. 89–94. ISSN: 14764687. DOI: 10.1038/s41586-019-
       1799-6.

[178] Tsung-Yi Lin et al. "Focal Loss for Dense Object Detection". In:
       IEEE Transactions on Pattern Analysis and Machine Intelligence
       42 (2 Feb. 2020). ISSN: 0162-8828. DOI: 10.1109/TPAMI.2018.
       2858826.

[179] Mark Sandler et al. "MobileNetV2: Inverted Residuals and Linear
       Bottlenecks". In: *2018 IEEE/CVF Conference on Computer
       Vision and Pattern Recognition*. IEEE, June 2018. ISBN: 978-1-
       5386-6420-9. DOI: 10.1109/CVPR.2018.00474.

[180] Benjamin Haibe-Kains et al. "Transparency and reproducibility in
       artificial intelligence". In: *Nature* 586 (7829 Oct. 2020), E14–E16.
       ISSN: 0028-0836. DOI: 10.1038/s41586-020-2766-y.

[181] Johan Verjans et al. "Cardiovascular Diseases". In: *Artificial
       Intelligence in Medical Imaging*. Ed. by Ranschaert E. and
       Morozov S.Algra P. Springer International Publishing, 2019. DOI:
       10.1007/978-3-319-94878-2_13.

[182] Carlos Martin-Isla et al. "Image-Based Cardiac Diagnosis With
       Machine Learning: A Review". In: *Frontiers in Cardiovascular
       Medicine* 7 (Jan. 2020). ISSN: 2297055X. DOI: 10.3389/fcvm.
       2020.00001.

[183] Riemer H. J. A. Slart et al. "Position paper of the EACVI and
       EANM on artificial intelligence applications in multimodality
       cardiovascular imaging using SPECT/CT, PET/CT, and cardiac
       CT". In: *European Journal of Nuclear Medicine and Molecular*

Imaging 48 (5 May 2021), pp. 1399–1413. ISSN: 1619-7070. DOI: 10.1007/s00259-021-05341-z.

[184] Chen Chen et al. "Deep learning for cardiac image segmentation: A review". In: *Frontiers in Cardiovascular Medicine* 7 (Nov. 2019). ISSN: 23318422. DOI: 10.3389/fcvm.2020.00025.

[185] Chengqin Ye et al. "Multi-depth fusion network for whole-heart CT image segmentation". In: *IEEE Access* 7 (2019), pp. 23421–23429. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2899635.

[186] Kevinminh Ta et al. "A Semi-supervised Joint Network for Simultaneous Left Ventricular Motion Tracking and Segmentation in 4D Echocardiography". In: *Medical Image Computing and Computer Assisted Intervention MICCAI 2020. MICCAI 2020. (Lecture Notes in Computer Science)*. Vol. 12266. Springer, 2020. DOI: 10.1007/978-3-030-59725-2_45.

[187] Nripesh Parajuli et al. "Flow network tracking for spatiotemporal and periodic point matching: Applied to cardiac motion analysis". In: *Medical Image Analysis* 55 (July 2019), pp. 116–135. ISSN: 13618423. DOI: 10.1016/j.media.2019.04.007.

[188] Suyu Dong et al. "VoxelAtlasGAN: 3D Left Ventricle Segmentation on Echocardiography with Atlas Guided Generation and Voxel-to-Voxel Discrimination". In: *Medical Image Computing and Computer Assisted Intervention MICCAI 2018. MICCAI 2018. (Lecture Notes in Computer Science)*. Ed. by Schnabel J. Frangi A. et al. Vol. 11073. Springer, 2018. DOI: 10.1007/978-3-030-00937-3_71.

[189] Ozan Oktay et al. "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation". In: *IEEE Transactions on Medical Imaging* 37 (2 Feb. 2018), pp. 384–395. ISSN: 1558254X. DOI: 10.1109/TMI.2017.2743464.

[190] Fabian Isensee et al. "Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features". In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (Lecture Notes in Computer Science)*. Vol. 10663. Springer International Publishing, 2018. DOI: 10.1007/978-3-319-75541-0_13.

[191] Qiao Zheng, Hervé Delingette, and Nicholas Ayache. "Explainable
        cardiac pathology classification on cine MRI with motion
        characterization by semi-supervised learning of apparent flow".
        In: *Medical Image Analysis* 56 (Aug. 2019), pp. 80–95. ISSN:
        13618423. DOI: 10.1016/j.media.2019.06.001.

[192] Zhaohan Xiong et al. "Fully Automatic Left Atrium Segmentation
        From Late Gadolinium Enhanced Magnetic Resonance Imaging
        Using a Dual Fully Convolutional Neural Network". In: *IEEE
        Transactions on Medical Imaging* 38 (2 Feb. 2019), pp. 515–524.
        ISSN: 1558254X. DOI: 10.1109/TMI.2018.2866845.

[193] Edward Ferdian et al. "Fully automated myocardial strain
        estimation from CMR Tagged images using a deep learning
        framework in the UK biobank". In: *Radiology: Cardiothoracic
        Imaging* 2 (1 Apr. 2020), e190032–undefined. ISSN: 23318422. DOI:
        10.1148/ryct.2020190032.

[194] Olivier Bernard et al. "Deep Learning Techniques for Automatic
        MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is
        the Problem Solved?" In: *IEEE Transactions on Medical Imaging*
        37 (11 Nov. 2018), pp. 2514–2525. ISSN: 1558254X. DOI: 10.1109/
        TMI.2018.2837502.

[195] Lien-Hsin Hu et al. "Machine learning predicts per-vessel
        early coronary revascularization after fast myocardial perfusion
        SPECT: results from multicentre REFINE SPECT registry". In:
        European Heart Journal - Cardiovascular Imaging 21 (5 May
        2020), pp. 549–559. ISSN: 2047-2404. DOI: 10.1093/ehjci/
        jez177.

[196] Julian Betancur et al. "Deep Learning for Prediction of
        Obstructive Disease From Fast Myocardial Perfusion SPECT: A
        Multicenter Study". In: *JACC: Cardiovascular Imaging* 11 (11
        Nov. 2018), pp. 1654–1663. ISSN: 18767591. DOI: 10.1016/j.
        jcmg.2018.01.020.

[197] Ke Yan et al. "DeepLesion: automated mining of large-scale lesion
        annotations and universal lesion detection with deep learning". In:
        Journal of Medical Imaging 5 (3 July 2018), p. 1. ISSN: 2329-4302.
        DOI: 10.1117/1.jmi.5.3.036501.

[198] Juan J. Cerrolaza et al. "Computational anatomy for multi-organ analysis in medical imaging: A review". In: *Medical Image Analysis* 56 (Aug. 2019), pp. 44–67. ISSN: 13618423. DOI: 10.1016/j.media.2019.04.002.

[199] Arshia Rehman and Fiaz Gul Khan. "A deep learning based review on abdominal images". In: *Multimedia Tools and Applications* (2020). ISSN: 15737721. DOI: 10.1007/s11042-020-09592-0.

[200] Pranav Rajpurkar et al. "AppendiXNet: Deep Learning for Diagnosis of Appendicitis from A Small Dataset of CT Exams Using Video Pretraining". In: *Scientific Reports* 10 (1 Dec. 2020). ISSN: 20452322. DOI: 10.1038/s41598-020-61055-6.

[201] Chin-Chi Kuo et al. "Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning". In: *npj Digital Medicine* 2 (1 Dec. 2019). ISSN: 2398-6352. DOI: 10.1038/s41746-019-0104-2.

[202] Andrei Iantsen et al. "Convolutional neural networks for PET functional volume fully automatic segmentation: development and validation in a multi-center setting". In: *European Journal of Nuclear Medicine and Molecular Imaging* (Mar. 2021), pp. 1–13. ISSN: 1619-7070. DOI: 10.1007/s00259-021-05244-z.

[203] Marta Ferreira et al. "[18F]FDG PET radiomics to predict disease-free survival in cervical cancer: a multi-scanner/center study with external validation". In: *European Journal of Nuclear Medicine and Molecular Imaging* (Mar. 2021), pp. 1–12. ISSN: 1619-7070. DOI: 10.1007/s00259-021-05303-5.

[204] Nathaniel Swinburne and Andrei Holodny. "Neurological Diseases". In: ed. by Ranschaert E. and Morozov S.Algra P. Springer International Publishing, 2019. DOI: 10.1007/978-3-319-94878-2_15.

[205] Vishnu M. Bashyam et al. "MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide". In: *Brain* 143 (7 July 2020), pp. 2312–2324. ISSN: 14602156. DOI: 10.1093/brain/awaa160.

[206] Han Peng et al. "Accurate brain age prediction with lightweight deep neural networks". In: *Medical Image Analysis* 68 (Feb. 2021), p. 101871. ISSN: 1361-8415. DOI: 10.1016/J.MEDIA.2020.101871.

[207] Shuo Han et al. "Automatic cerebellum anatomical parcellation using U-Net with locally constrained optimization". In: *NeuroImage* 218 (Sept. 2020), 116819–undefined. ISSN: 10959572. DOI: `10.1016/j.neuroimage.2020.116819`.

[208] Benjamin Thyreau and Yasuyuki Taki. "Learning a cortical parcellation of the brain robust to the MRI segmentation with convolutional neural networks". In: *Medical Image Analysis* 61 (Apr. 2020), p. 101639. ISSN: 1361-8415. DOI: `10.1016/J.MEDIA.2020.101639`.

[209] Mr Amir Ebrahimighahnavieh, Suhuai Luo, and Raymond Chiong. "Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review". In: *Computer Methods and Programs in Biomedicine* 187 (Apr. 2020), 105242–undefined. ISSN: 18727565. DOI: `10.1016/j.cmpb.2019.105242`.

[210] Razvan V. Marinescu et al. "The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge: Results after 1 Year Follow-up". In: *arXiv* (Feb. 2020). URL: `http://arxiv.org/abs/2002.03419`.

[211] Weizheng Yan et al. "Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data". In: *EBioMedicine* 47 (Sept. 2019), pp. 543–552. ISSN: 23523964. DOI: `10.1016/j.ebiom.2019.08.023`.

[212] Jihoon Oh et al. "Identifying Schizophrenia Using Structural MRI With a Deep Learning Algorithm". In: *Frontiers in Psychiatry* 11 (Feb. 2020), p. 16. ISSN: 1664-0640. DOI: `10.3389/fpsyt.2020.00016`.

[213] Hai Ye et al. "Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network". In: *European Radiology* 29 (11 Nov. 2019), pp. 6191–6201. ISSN: 0938-7994. DOI: `10.1007/s00330-019-06163-2`.

[214] Adam E. Flanders et al. "Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge". In: *Radiology: Artificial Intelligence* 2 (3 May 2020), e190211. ISSN: 2638-6100. DOI: `10.1148/ryai.2020190211`.

[215] Zhao Shi et al. "A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images". In: *Nature Communications* 11 (1 Dec. 2020), p. 6090. ISSN: 2041-1723. DOI: 10.1038/s41467-020-19527-w.

[216] Bio Joo et al. "A deep learning algorithm may automate intracranial aneurysm detection on MR angiography with high diagnostic performance". In: *European Radiology* 30 (11 Nov. 2020), pp. 5785–5793. ISSN: 0938-7994. DOI: 10.1007/s00330-020-06966-8.

[217] Jiehua Yang et al. "Deep Learning for Detecting Cerebral Aneurysms with CT Angiography". In: *Radiology* 298 (1 Jan. 2021), pp. 155–163. ISSN: 0033-8419. DOI: 10.1148/radiol.2020192154.

[218] Yiming Ding et al. "A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using ¹⁸ F-FDG PET of the Brain". In: *Radiology* 290 (2 Feb. 2019), pp. 456–464. ISSN: 0033-8419. DOI: 10.1148/radiol.2018180958.

[219] Susanne G. Mueller et al. "The Alzheimer's Disease Neuroimaging Initiative". In: *Neuroimaging Clinics of North America* 15 (4 Nov. 2005), pp. 869–877. ISSN: 1052-5149. DOI: 10.1016/J.NIC.2005.09.008.

[220] Seung Kwan Kang, Hongyoon Choi, and Jae Sung Lee. "Translating amyloid PET of different radiotracers by a deep generative model for interchangeability". In: *NeuroImage* 232 (May 2021), p. 117890. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2021.117890.

[221] Haohui Liu et al. "Improved amyloid burden quantification with nonspecific estimates using deep learning". In: *European Journal of Nuclear Medicine and Molecular Imaging* 48 (6 June 2021), pp. 1842–1853. ISSN: 1619-7070. DOI: 10.1007/s00259-020-05131-z.

[222] Jerry W. Froelich and Ali Salavati. "Artificial Intelligence in PET/CT Is about to Make Whole-Body Tumor Burden Measurements a Clinical Reality". In: *Radiology* 294 (2 Feb. 2020), pp. 453–454. ISSN: 0033-8419. DOI: 10.1148/radiol.2019192425.

[223] Ludovic Sibille et al. "¹⁸ F-FDG PET/CT Uptake Classification in Lymphoma and Lung Cancer by Using Deep Convolutional Neural Networks". In: *Radiology* 294 (2 Feb. 2020), pp. 445–452. ISSN: 0033-8419. DOI: 10.1148/radiol.2019191114.

[224] Pierre Pinochet et al. "Evaluation of an Automatic Classification Algorithm Using Convolutional Neural Networks in Oncological Positron Emission Tomography". In: *Frontiers in Medicine* 8 (Feb. 2021), p. 117. ISSN: 2296-858X. DOI: 10.3389/fmed.2021.628179.

[225] Kaustav Bera et al. "Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology". In: *Nature Reviews Clinical Oncology* 16 (11 Nov. 2019), pp. 703–715. ISSN: 1759-4774. DOI: 10.1038/s41571-019-0252-y.

[226] Shujian Deng et al. "Deep learning in digital pathology image analysis: a survey". In: *Frontiers of Medicine* 14 (4 Aug. 2020), pp. 470–487. ISSN: 2095-0217. DOI: 10.1007/s11684-020-0782-9.

[227] Jun Xu et al. "Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images". In: *IEEE Transactions on Medical Imaging* 35 (1 Jan. 2016), pp. 119–130. ISSN: 0278-0062. DOI: 10.1109/TMI.2015.2458702.

[228] Andrew Janowczyk and Anant Madabhushi. "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases". In: *Journal of Pathology Informatics* 7 (1 2016), p. 29. ISSN: 2153-3539. DOI: 10.4103/2153-3539.186902.

[229] Angel Cruz-Roa et al. "High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection". In: *PLOS ONE* 13 (5 May 2018). Ed. by Yuanquan Wang, e0196828. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0196828.

[230] Gabriele Campanella et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". In: *Nature Medicine* 25 (8 Aug. 2019), pp. 1301–1309. ISSN: 1078-8956. DOI: 10.1038/s41591-019-0508-1.

[231] Nicolas Coudray et al. "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning". In: *Nature Medicine* 24 (10 Oct. 2018), pp. 1559–1567. ISSN: 1078-8956. DOI: 10.1038/s41591-018-0177-5.

[232] Wouter Bulten et al. "Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study." In: *The Lancet. Oncology* 21 (2 Feb. 2020), pp. 233–241. ISSN: 1474-5488. DOI: 10.1016/S1470-2045(19)30739-9.

[233] Jakob Nikolas Kather et al. "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer". In: *Nature Medicine* 25 (7 July 2019), pp. 1054–1056. ISSN: 1078-8956. DOI: 10.1038/s41591-019-0462-y.

[234] Jakob Nikolas Kather et al. "Pan-cancer image-based detection of clinically actionable genetic alterations". In: *Nature Cancer* 1 (8 Aug. 2020), pp. 789–799. ISSN: 2662-1347. DOI: 10.1038/s43018-020-0087-6.

[235] Geert Litjens et al. "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis". In: *Scientific Reports* 6 (1 Sept. 2016), p. 26286. ISSN: 2045-2322. DOI: 10.1038/srep26286.

[236] Wouter Bulten et al. "Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard". In: *Scientific Reports* 9 (1 Dec. 2019), p. 864. ISSN: 2045-2322. DOI: 10.1038/s41598-018-37257-4.

[237] Peter Ström et al. "Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study." In: *The Lancet. Oncology* 21 (2 Feb. 2020), pp. 222–232. ISSN: 1474-5488. DOI: 10.1016/S1470-2045(19)30738-7.

[238] Milan Decuyper et al. "Artificial neural networks for positioning of gamma interactions in monolithic PET detectors". In: *Physics in Medicine and Biology* 66 (7 Mar. 2021), p. 075001. ISSN: 0031-9155. DOI: 10.1088/1361-6560/abebfc.

[239] Radoslaw Marcinkowski et al. "Sub-millimetre DOI detector based on monolithic LYSO and digital SiPM for a dedicated small-animal PET system". In: *Physics in Medicine and Biology* 61 (5

Mar. 2016), pp. 2196–2212. ISSN: 0031-9155. DOI: 10.1088/0031-9155/61/5/2196.

[240]   William W Moses. "Fundamental Limits of Spatial Resolution in PET." In: *Nuclear instruments and methods in physics research. Section A, Accelerators, spectrometers, detectors and associated equipment* 648 Supplement 1 (Aug. 2011), S236–S240. ISSN: 0168-9002. DOI: 10.1016/j.nima.2010.11.092.

[241]   Stefaan Vandenberghe et al. "PET20.0: a cost efficient, 2mm spatial resolution Total Body PET with point sensitivity up to 22% and adaptive axial FOV of maximum 2.00m". In: Annual Congress of the European Association of Nuclear Medicine. Vol. 44. 2017, S305.

[242]   Andrea González-Montoro et al. "Novel method to measure the intrinsic spatial resolution in PET detectors based on monolithic crystals". In: *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 920 (Mar. 2019), pp. 58–67. ISSN: 01689002. DOI: 10.1016/j.nima.2018.12.056.

[243]   Emilie Roncali, Mariele Stockhoff, and Simon R. Cherry. "An integrated model of scintillator-reflector properties for advanced simulations of optical transport". In: *Physics in Medicine and Biology* 62 (12 2017), pp. 4811–4830. ISSN: 13616560. DOI: 10.1088/1361-6560/aa6ca5.

[244]   Mariele Stockhoff et al. "Advanced optical simulation of scintillation detectors in GATE V8.0: First implementation of a reflectance model based on measured data". In: *Physics in Medicine and Biology* 62 (12 2017), pp. L1–L8. ISSN: 13616560. DOI: 10.1088/1361-6560/aa7007.

[245]   Adam Paszke et al. "PyTorch: An imperative style, high-performance deep learning library". In: Advances in Neural Information Processing Systems. Vol. 32. 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[246]   Mikiko Ito, Seong Jong Hong, and Jae Sung Lee. "Positron emission tomography (PET) detectors with depth-of- interaction

(DOI) capability". In: *Biomedical Engineering Letters* 1 (2 2011), pp. 70–81. ISSN: 20939868. DOI: 10.1007/s13534-011-0019-6.

[247] Andrea Gonzalez-Montoro et al. "Performance Study of a Large Monolithic LYSO PET Detector With Accurate Photon DOI Using Retroreflector Layers". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 1 (3 2017), pp. 229–237. ISSN: 2469-7311. DOI: 10.1109/trpms.2017.2692819.

[248] O. Klein and Y. Nishina. "Über die Streuung von Strahlung durch freie Elektronen nach der neuen relativistischen Quantendynamik von Dirac". In: *Zeitschrift für Physik* 52 (11-12 Nov. 1929), pp. 853–868. ISSN: 0044-3328. DOI: 10.1007/BF01366453.

[249] Dscraggs. *Klein-Nishina distribution.* 2009. URL: https : / / commons . wikimedia . org / wiki / File : Klein - Nishina _ distribution.png.

[250] Sang-June Park, W Leslie Rogers, and Neal H Clinthorne. "Design of a very high-resolution small animal PET scanner using a silicon scatter detector insert". In: *Physics in Medicine and Biology* 52 (15 Aug. 2007), pp. 4653–4677. ISSN: 0031-9155. DOI: 10.1088/0031-9155/52/15/019.

[251] M Rafecas et al. "Inter-crystal scatter in a dual layer, high resolution LSO-APD positron emission tomograph". In: *Physics in Medicine and Biology* 48 (7 Apr. 2003), pp. 821–848. ISSN: 0031-9155. DOI: 10.1088/0031-9155/48/7/302.

[252] Eiji Yoshida et al. "Inter-crystal scatter identification for a depth-sensitive detector using support vector machine for small animal positron emission tomography". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 571 (1-2 Feb. 2007), pp. 243–246. ISSN: 0168-9002. DOI: 10.1016/J.NIMA.2006. 10.073.

[253] Jean-Baptiste Michaud et al. "Sensitivity Increase Through a Neural Network Method for LOR Recovery of ICS Triple Coincidences in High-Resolution Pixelated- Detectors PET Scanners". In: *IEEE Transactions on Nuclear Science* 62 (1 Feb. 2015), pp. 82–94. ISSN: 0018-9499. DOI: 10.1109/TNS.2014. 2372788.

[254] Xiaoli Li et al. "Use of Cramer-Rao Lower Bound for Performance Evaluation of Different Monolithic Crystal PET Detector Designs." In: *IEEE transactions on nuclear science* 59 (1 2012), pp. 3–12. ISSN: 0018-9499. DOI: 10.1109/TNS.2011.2165968.

[255] Mariele Stockhoff, Roel Van Holen, and Stefaan Vandenberghe. "Identifying potential sources of resolution degradation in monolithic scintillators: simulations vs experiments". In: 2020 IEEE Nuclear Science Symposium and Medical Imaging Conferene (NSS-MIC). 2020.

[256] Mariele Stockhoff et al. "High-resolution monolithic LYSO detector with 6-layer depth-of-interaction for clinical PET". In: *Physics in Medicine and Biology* 66 (15 Aug. 2021), p. 155014. ISSN: 0031-9155. DOI: 10.1088/1361-6560/ac1459.

[257] Rita Carter et al. *The human brain book: An illustrated guide to its structure, function and disorders.* Dorling Kindersley Limited, 2009, p. 256. ISBN: 075666215X.

[258] National Cancer Institute. *Adult Central Nervous System Tumors Treatment (PDQ): Patient Version.* 2020. URL: https://www.cancer.gov/types/brain/patient/adult-brain-treatment-pdq (visited on 06/10/2021).

[259] BioNinja. *Brain Matter.* URL: https://ib.bioninja.com.au/options/option-a-neurobiology-and/a2-the-human-brain/brain-matter.html (visited on 06/10/2021).

[260] Holly Fischer. *Glial cell types.* 2013. URL: https://en.wikipedia.org/wiki/Glia#/media/File:Glial_Cell_Types.png (visited on 06/10/2021).

[261] David N. Louis et al. "The 2007 WHO Classification of Tumours of the Central Nervous System". In: *Acta Neuropathologica* 114 (2 July 2007), pp. 97–109. ISSN: 0001-6322. DOI: 10.1007/s00401-007-0243-4.

[262] C. Giannini et al. "Oligodendrogliomas: Reproducibility and Prognostic Value of Histologic Diagnosis and Grading". In: *Journal of Neuropathology and Experimental Neurology* 60 (3 Mar. 2001), pp. 248–262. ISSN: 0022-3069. DOI: 10.1093/jnen/60.3.248.

[263] Martin J van den Bent. "Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective." In: *Acta neuropathologica* 120 (3 Sept. 2010), pp. 297–304. ISSN: 1432-0533. DOI: 10.1007/s00401-010-0725-7.

[264] Catherine L Nutt et al. "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification." In: *Cancer research* 63 (7 Apr. 2003), pp. 1602–1607. ISSN: 0008-5472. URL: http://www.ncbi.nlm.nih.gov/pubmed/12670911.

[265] Lonneke A.M. Gravendeel et al. "Intrinsic Gene Expression Profiles of Gliomas Are a Better Predictor of Survival than Histology". In: *Cancer Research* 69 (23 Dec. 2009), pp. 9065–9072. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-09-2307.

[266] The Cancer Genome Atlas Research Network. "Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas". In: *New England Journal of Medicine* 372 (26 June 2015), pp. 2481–2498. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1402121.

[267] Quinn T Ostrom et al. "CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2011-2015". In: *Neuro-Oncology* 20 (suppl4 Oct. 2018), pp. iv1–iv86. ISSN: 1522-8517. DOI: 10.1093/neuonc/noy131.

[268] Sue Han et al. "IDH mutation in glioma: molecular mechanisms and potential therapeutic targets". In: *British Journal of Cancer* 122 (11 May 2020), pp. 1580–1589. ISSN: 0007-0920. DOI: 10.1038/s41416-020-0814-x.

[269] Jeanette E. Eckel-Passow et al. "Glioma Groups Based on 1p/19q, IDH , and TERT Promoter Mutations in Tumors". In: *New England Journal of Medicine* 372 (26 June 2015), pp. 2499–2508. DOI: 10.1056/NEJMoa1407279.

[270] Hai Yan et al. "IDH1 and IDH2 Mutations in Gliomas". In: *New England Journal of Medicine* 360 (8 Feb. 2009), pp. 765–773. DOI: 10.1056/NEJMoa0808710.

[271] David E. Reuss et al. "IDH mutant diffuse and anaplastic astrocytomas have similar age at presentation and little difference in survival: a grading problem for WHO". In: *Acta Neuropathologica* 129 (6 June 2015), pp. 867–873. ISSN: 0001-6322. DOI: 10.1007/s00401-015-1438-8.

[272] Martin J. van den Bent et al. "Interlaboratory comparison of IDH mutation detection". In: *Journal of Neuro-Oncology* 112 (2 Apr. 2013), pp. 173–178. ISSN: 0167-594X. DOI: 10.1007/s11060-013-1056-z.

[273] Michael Weller et al. "European Association for Neuro-Oncology (EANO) guideline on the diagnosis and treatment of adult astrocytic and oligodendroglial gliomas". In: *The Lancet Oncology* 18 (6 June 2017), e315–e329. ISSN: 1470-2045. DOI: 10.1016/S1470-2045(17)30194-8.

[274] Paula de Robles et al. "The worldwide incidence and prevalence of primary brain tumors: a systematic review and meta-analysis". In: *Neuro-Oncology* 17 (6 June 2015), pp. 776–783. ISSN: 1523-5866. DOI: 10.1093/neuonc/nou283.

[275] Jill S. Barnholtz-Sloan, Quinn T. Ostrom, and David Cote. "Epidemiology of Brain Tumors". In: *Neurologic Clinics* 36 (3 Aug. 2018), pp. 395–419. ISSN: 0733-8619. DOI: 10.1016/J.NCL.2018.04.001.

[276] Katharine A. McNeill. "Epidemiology of Brain Tumors". In: *Neurologic Clinics* 34 (4 Nov. 2016), pp. 981–998. ISSN: 0733-8619. DOI: 10.1016/J.NCL.2016.06.014.

[277] Sarah Lapointe, Arie Perry, and Nicholas A Butowski. "Primary brain tumours in adults". In: *The Lancet* 392 (10145 Aug. 2018), pp. 432–446. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(18)30990-5.

[278] Michael Weller et al. "EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood". In: *Nature Reviews Clinical Oncology* 18 (3 Mar. 2021), pp. 170–186. ISSN: 1759-4774. DOI: 10.1038/s41571-020-00447-z.

[279] Maarten M J Wijnenga et al. "Does early resection of presumed low-grade glioma improve survival? A clinical perspective." In: *Journal of neuro-oncology* 133 (1 May 2017), pp. 137–146. ISSN: 1573-7373. DOI: 10.1007/s11060-017-2418-8.

[280] R J Jackson et al. "Limitations of stereotactic biopsy in the initial management of gliomas." In: *Neuro-oncology* 3 (3 2001), pp. 193–200. ISSN: 1522-8517. DOI: 10.1093/neuonc/3.3.193.

[281] Graeme Woodworth et al. "Accuracy of frameless and frame-based image-guided stereotactic brain biopsy in the diagnosis of glioma: comparison of biopsy and open resection specimen". In: *Neurological Research* 27 (4 June 2005), pp. 358–362. ISSN: 0161-6412. DOI: 10.1179/016164105X40057.

[282] Nathalie L. Jansen et al. "MRI-suspected low-grade glioma: is there a need to perform dynamic FET PET?" In: *European Journal of Nuclear Medicine and Molecular Imaging* 39 (6 June 2012), pp. 1021–1029. ISSN: 1619-7070. DOI: 10.1007/s00259-012-2109-9.

[283] L Khalid et al. "Imaging Characteristics of Oligodendrogliomas That Predict Grade". In: *American Journal of Neuroradiology* 33 (5 2012), pp. 852–857. DOI: 10.3174/ajnr.A2895.

[284] Johan Pallud et al. "Prognostic significance of imaging contrast enhancement for WHO grade II gliomas". In: *Neuro-Oncology* 11 (2 Apr. 2009), pp. 176–182. ISSN: 1523-5866. DOI: 10.1215/15228517-2008-066.

[285] Bjoern H. Menze et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34 (10 Oct. 2015), pp. 1993–2024. ISSN: 0278-0062. DOI: 10.1109/TMI.2014.2377694.

[286] Marcel Prastawa et al. "A brain tumor segmentation framework based on outlier detection". In: *Medical Image Analysis* 8 (3 Sept. 2004), pp. 275–283. ISSN: 1361-8415. DOI: 10.1016/J.MEDIA.2004.06.007.

[287] Miri Erihov et al. "A Cross Saliency Approach to Asymmetry-Based Tumor Detection". In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Springer, Cham, 2015, pp. 636–643. DOI: 10.1007/978-3-319-24574-4_76.

[288] Anjali Wadhwa, Anuj Bhardwaj, and Vivek Singh Verma. "A review on brain tumor segmentation of MRI images". In: *Magnetic Resonance Imaging* 61 (Sept. 2019), pp. 247–259. ISSN: 0730-725X. DOI: 10.1016/J.MRI.2019.05.043.

[289] Stefan Bauer, Lutz-P. Nolte, and Mauricio Reyes. "Fully Automatic Segmentation of Brain Tumor Images Using Support Vector Machine Classification in Combination with Hierarchical Conditional Random Field Regularization". In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2011. Springer, Berlin, Heidelberg, 2011, pp. 354–361. DOI: 10.1007/978-3-642-23626-6_44.

[290] Darko Zikic et al. "Decision Forests for Tissue-Specific Segmentation of High-Grade Gliomas in Multi-channel MR". In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2012. Springer, Berlin, Heidelberg, 2012, pp. 369–376. DOI: 10.1007/978-3-642-33454-2_46.

[291] Bjoern H Menze et al. "Segmenting Glioma in Multi-Modal Images using a Generative-Discriminative Model for Brain Lesion Segmentation". In: Proceedings of MICCAI-BRATS. 2012, p. 8. URL: http://people.csail.mit.edu/menze/papers/menze_12_brats-discriminative.pdf.

[292] Nicholas J. Tustison et al. "Optimal Symmetric Multimodal Templates and Concatenated Random Forests for Supervised Brain Tumor Segmentation (Simplified) with ANTsR". In: Neuroinformatics 13 (2 Apr. 2015), pp. 209–225. ISSN: 1539-2791. DOI: 10.1007/s12021-014-9245-2.

[293] Spyridon Bakas et al. "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge". In: ArXiv preprint abs/1811.0 (2018). URL: http://arxiv.org/abs/1811.02629.

[294] Sérgio Pereira et al. "Deep Convolutional Neural Networks for the Segmentation of Gliomas in Multi-sequence MRI". In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - BrainLes 2015. Springer, Cham, 2016, pp. 131–143. DOI: 10.1007/978-3-319-30858-6_12.

[295] Konstantinos Kamnitsas et al. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation". In: Medical Image Analysis 36 (Feb. 2017), pp. 61–78. ISSN: 1361-8415. DOI: 10.1016/J.MEDIA.2016.10.004.

[296] Fabian Isensee et al. "Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge". In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - BrainLes 2017. Vol. 10670 LNCS. 2018, pp. 287–297. ISBN: 9783319752372. DOI: 10.1007/978-3-319-75238-9_25.

[297] K Kamnitsas et al. "Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation". In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - BrainLes 2017. Ed. by Reyes M. Crimi A. et al. Springer, Cham, 2018, pp. 450–462. DOI: 10.1007/978-3-319-75238-9_38.

[298] Andriy Myronenko. "3D MRI Brain Tumor Segmentation Using Autoencoder Regularization". In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - BrainLes 2018. Ed. by Alessandro Crimi et al. Springer, Cham, Sept. 2019, pp. 311–320. DOI: 10.1007/978-3-030-11726-9_28.

[299] Zeyu Jiang et al. "Two-Stage Cascaded U-Net: 1st Place Solution to BraTS Challenge 2019 Segmentation Task". In: Springer, Cham, Oct. 2020, pp. 231–241. DOI: 10.1007/978-3-030-46640-4_22.

[300] Fabian Isensee et al. "nnU-Net for Brain Tumor Segmentation". In: Springer, Cham, Oct. 2021, pp. 118–132. DOI: 10.1007/978-3-030-72087-2_11.

[301] Jeffrey D Rudie et al. "Emerging Applications of Artificial Intelligence in Neuro-Oncology." In: Radiology 290 (3 2019), pp. 607–618. ISSN: 1527-1315. DOI: 10.1148/radiol.2018181928.

[302] Philipp Lohmann et al. "Radiomics in neuro-oncology: Basics, workflow, and applications". In: Methods 188 (Apr. 2021), pp. 112–121. ISSN: 1046-2023. DOI: 10.1016/J.YMETH.2020.06.003.

[303] M. Monica Subashini et al. "A non-invasive methodology for the grade identification of astrocytoma using image processing and artificial intelligence techniques". In: Expert Systems with Applications 43 (2016), pp. 186–196. ISSN: 09574174. DOI: 10.1016/j.eswa.2015.08.036.

[304]   Kevin Li-Chun Hsieh, Chung-Ming Lo, and Chih-Jou Hsiao.
         "Computer-aided grading of gliomas based on local and
         global MRI features". In: *Computer Methods and Programs in
         Biomedicine* 139 (Feb. 2017), pp. 31–38. ISSN: 0169-2607. DOI:
         10.1016/J.CMPB.2016.10.021.

[305]   Kevin Li-Chun Hsieh et al. "Effect of a computer-aided diagnosis
         system on radiologists' performance in grading gliomas with
         MRI." In: *PloS one* 12 (2 2017), e0171342. ISSN: 1932-6203. DOI:
         10.1371/journal.pone.0171342.

[306]   Ying Zhuge et al. "Automated glioma grading on conventional
         MRI images using deep convolutional neural networks". In:
         Medical Physics 47 (7 July 2020), pp. 3044–3053. ISSN: 0094-2405.
         DOI: 10.1002/mp.14168.

[307]   Ken Chang et al. "Residual Convolutional Neural Network for the
         Determination of IDH Status in Low-and High- Grade Gliomas
         from MR Imaging". In: *Clin Cancer Res* 24 (5 2018), pp. 1073–
         1081. DOI: 10.1158/1078-0432.CCR-17-2236.

[308]   Biqi Zhang et al. "Multimodal MRI features predict isocitrate
         dehydrogenase genotype in high-grade gliomas". In: *Neuro-
         Oncology* 19 (1 Jan. 2017), pp. 109–117. ISSN: 1522-8517. DOI:
         10.1093/neuonc/now121.

[309]   Hao Zhou et al. "Machine learning reveals multimodal MRI
         patterns predictive of isocitrate dehydrogenase and 1p/19q status
         in diffuse low- and high-grade gliomas." In: *Journal of neuro-
         oncology* 142 (2 Apr. 2019), pp. 299–307. ISSN: 1573-7373. DOI:
         10.1007/s11060-019-03096-0.

[310]   Kyu Sung Choi, Seung Hong Choi, and Bumseok Jeong. "Pre-
         diction of IDH genotype in gliomas with dynamic susceptibility
         contrast perfusion MR imaging using an explainable recurrent
         neural network". In: *Neuro-Oncology* (June 2019). ISSN: 1522-
         8517. DOI: 10.1093/neuonc/noz095.

[311]   Chandan Ganesh Bangalore Yogananda et al. "A Novel Fully
         Automated MRI-Based Deep Learning Method for Classification
         of IDH Mutation Status in Brain Gliomas". In: *Neuro-Oncology*
         22 (3 Oct. 2019), pp. 402–411. ISSN: 1522-8517. DOI: 10.1093/
         neuonc/noz199.

[312] Chandan Ganesh Bangalore Yogananda et al. "A novel fully automated MRI-based deep-learning method for classification of 1p/19q co-deletion status in brain gliomas". In: *Neuro-Oncology Advances* 2 (Supplement 4 Dec. 2020), pp. iv42–iv48. ISSN: 2632-2498. DOI: 10.1093/noajnl/vdaa066.

[313] Zeynettin Akkus et al. "Predicting Deletion of Chromosomal Arms 1p/19q in Low-Grade Gliomas from MR Images Using Machine Intelligence." In: *Journal of digital imaging* 30 (4 Aug. 2017), pp. 469–476. ISSN: 1618-727X. DOI: 10.1007/s10278-017-9984-3.

[314] Sebastian R. van der Voort et al. "Predicting the 1p/19q Codeletion Status of Presumed Low-Grade Glioma with an Externally Validated Machine Learning Algorithm". In: *Clinical Cancer Research* 25 (24 Dec. 2019), pp. 7455–7462. ISSN: 1078-0432. DOI: 10.1158/1078-0432.CCR-19-1127.

[315] Evangelia I Zacharaki, Vasileios G Kanas, and Christos Davatzikos. "Investigating machine learning techniques for MRI-based classification of brain neoplasms." In: *International journal of computer assisted radiology and surgery* 6 (6 Nov. 2011), pp. 821–828. DOI: 10.1007/s11548-011-0559-3.

[316] Karoline Skogen et al. "Diagnostic performance of texture analysis on MRI in grading cerebral gliomas". In: *European Journal of Radiology* 85 (4 Apr. 2016), pp. 824–829. ISSN: 0720-048X. DOI: 10.1016/J.EJRAD.2016.01.013.

[317] Qiang Tian et al. "Radiomics strategy for glioma grading using texture features from multiparametric MRI". In: *Journal of Magnetic Resonance Imaging* 48 (6 Dec. 2018), pp. 1518–1528. ISSN: 10531807. DOI: 10.1002/jmri.26010.

[318] Yang Yang et al. "Glioma Grading on Conventional MR Images: A Deep Learning Study With Transfer Learning." In: *Frontiers in neuroscience* 12 (2018), p. 804. ISSN: 1662-4548. DOI: 10.3389/fnins.2018.00804.

[319] Jinhua Yu et al. "Noninvasive IDH1 mutation estimation based on a quantitative radiomics approach for grade II glioma". In: *European Radiology* 27 (8 Aug. 2017), pp. 3509–3522. ISSN: 0938-7994. DOI: 10.1007/s00330-016-4653-3.

[320] Hideyuki Arita et al. "Lesion location implemented magnetic resonance imaging radiomics for predicting IDH and TERT promoter mutations in grade II/III gliomas". In: *Scientific Reports* 8 (1 Dec. 2018), p. 11773. ISSN: 2045-2322. DOI: 10.1038/s41598-018-30273-4.

[321] P Chang et al. "Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas". In: *AJNR. American journal of neuroradiology* (May 2018). ISSN: 1936-959X. DOI: 10.3174/ajnr.A5667.

[322] Saima Rathore et al. "Multi-institutional noninvasive in vivo characterization of IDH, 1p/19q, and EGFRvIII in glioma using neuro-Cancer Imaging Phenomics Toolkit (neuro-CaPTk)". In: *Neuro-Oncology Advances* 2 (Supplement 4 Dec. 2020), pp. iv22–iv34. DOI: 10.1093/noajnl/vdaa128.

[323] Yuqi Han et al. "Non-invasive genotype prediction of chromosome 1p/19q co-deletion by development and validation of an MRI-based radiomics signature in lower-grade gliomas". In: *Journal of Neuro-Oncology* 140 (2 Nov. 2018), pp. 297–306. ISSN: 0167-594X. DOI: 10.1007/s11060-018-2953-y.

[324] Donnie Kim et al. "Prediction of 1p/19q Codeletion in Diffuse Glioma Patients Using Pre-operative Multiparametric Magnetic Resonance Imaging". In: *Frontiers in Computational Neuroscience* 13 (July 2019), p. 52. ISSN: 1662-5188. DOI: 10.3389/fncom.2019.00052.

[325] Stijn Bonte. "Artificial intelligence in medical imaging for the diagnosis of primary brain tumours". PhD thesis. Ghent University, 2018, pp. XVIII, 222. ISBN: 9789463551687.

[326] Milan Decuyper, Stijn Bonte, and Roel Van Holen. "Binary Glioma grading : radiomics versus pre-trained CNN features". In: Medical Imaging Summer School 2018 : Medical Imaging meets Deep Learning. Favignana, Sicily, Italy, 2018, pp. 13–13. URL: http://iplab.dmi.unict.it/miss/posters.htm.

[327] Milan Decuyper, Stijn Bonte, and Roel Van Holen. "Binary Glioma Grading: Radiomics versus Pre-trained CNN Features". In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2018. Ed. by Alejandro F. Frangi et al. Springer

International Publishig, 2018. DOI: 10.1007/978-3-030-00931-1_57.

[328] Geethu Mohan and M. Monica Subashini. "MRI based medical image analysis: Survey on brain tumor grade classification". In: *Biomedical Signal Processing and Control* 39 (Jan. 2018), pp. 139–161. DOI: 10.1016/J.BSPC.2017.07.007.

[329] Spyridon Bakas et al. "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific Data* 4 (Sept. 2017), p. 170117. ISSN: 2052-4463. DOI: 10.1038/sdata.2017.117.

[330] William D. Penny et al. *Statistical Parametric Mapping: the Analysis of Functional Brain Images*. 1st edition. Academic Press [Imprint], 2006. ISBN: 9781493300952.

[331] Russell T. Shinohara et al. "Statistical normalization techniques for magnetic resonance imaging". In: *NeuroImage: Clinical* 6 (2014), pp. 9–19. ISSN: 22131582. DOI: 10.1016/j.nicl.2014.08.008.

[332] Hugo J W L Aerts et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." In: *Nature communications* 5 (June 2014), p. 4006. ISSN: 2041-1723. DOI: 10.1038/ncomms5006.

[333] J. M.Y. Willaime et al. "Quantification of intra-tumour cell proliferation heterogeneity using imaging descriptors of 18F fluorothymidine-positron emission tomography". In: *Physics in Medicine and Biology* 58 (2 2013), pp. 187–203. ISSN: 00319155. DOI: 10.1088/0031-9155/58/2/187.

[334] Milan Decuyper et al. eng. In: *Medical Imaging with Deep Learning: MIDL 2020 - Short Paper Track*. Ed. by Christopher Pal et al. Montréal, Canada, p. 5.

[335] Milan Decuyper et al. "Automated MRI based pipeline for segmentation and prediction of grade, IDH mutation and 1p19q co-deletion in glioma". In: *Computerized Medical Imaging and Graphics* 88 (Mar. 2021). ISSN: 08956111. DOI: 10.1016/j.compmedimag.2020.101831.

[336] Fabian Isensee et al. "No New-Net". In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - BrainLes 2019. Ed. by Alessandro Crimi et al. Springer, Cham, Sept. 2019, pp. 234–244. DOI: 10.1007/978-3-030-11726-9_21.

[337] Spyridon Bakas and Chiharu Sako. *MICCAI-BraTS 2019 Leaderboard.* 2019. URL: https://www.cbica.upenn.edu/BraTS19/lboardValidation.html.

[338] Yan Shen and Mingchen Gao. "Brain Tumor Segmentation on MRI with Missing Modalities". In: Information Processing in Medical Imaging - IPMI 2019. Springer, Cham, June 2019, pp. 417–428. DOI: 10.1007/978-3-030-20351-1_32.

[339] Tongxue Zhou et al. "Brain Tumor Segmentation with Missing Modalities via Latent Multi-source Correlation Representation". In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020. Springer, Cham, Oct. 2020, pp. 533–541. DOI: 10.1007/978-3-030-59719-1_52.

[340] Michael L. Cheng et al. "Clinical tumour sequencing for precision oncology: time for a universal strategy". In: *Nature Reviews Cancer* 18 (9 Sept. 2018), pp. 527–528. ISSN: 1474-175X. DOI: 10.1038/s41568-018-0043-2.

[341] Lisa Scarpace et al. "Radiology Data from The Cancer Genome Atlas Glioblastoma Multiforme [TCGA-GBM] collection". In: (Jan. 2016). DOI: 10.7937/K9/TCIA.2016.RNYFUYE9.

[342] Nancy Pedano et al. "Radiology Data from The Cancer Genome Atlas Low Grade Glioma [TCGA-LGG] collection". In: *The Cancer Imaging Archive* (Jan. 2016). DOI: 10.7937/K9/TCIA.2016.L4LTD3TK.

[343] Bradley Erickson et al. "Data From LGG-1p19qDeletion". In: *The Cancer Imaging Archive* (2017). DOI: 10.7937/K9/TCIA.2017.dwehtz9v.

[344] Michele Ceccarelli et al. "Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma". In: *Cell* 164 (3 2016), p. 550. DOI: 10.1016/J.CELL.2015.12.028.

[345] Martinus P G Broen et al. "The T2-FLAIR mismatch sign as an imaging marker for non-enhancing IDH-mutant, 1p/19q-intact lower-grade glioma: a validation study." In: *Neuro-oncology* 20 (10 2018), pp. 1393–1399. ISSN: 1523-5866. DOI: 10.1093/neuonc/noy048.

[346] Sohil H Patel et al. "Cancer Therapy: Clinical T2-FLAIR Mismatch, an Imaging Biomarker for IDH and 1p/19q Status in Lower-grade Gliomas: A TCGA/TCIA Project". In: *Clin Cancer Res* 23 (20 2017). DOI: 10.1158/1078-0432.CCR-17-0560.

[347] N. Shah et al. "Data from Ivy GAP [Data set]". In: *The Cancer Imaging Archive* (2016). DOI: https://doi.org/10.7937/K9/TCIA.2016.XLWAN6NL.

[348] Ralph B. Puchalski et al. "An anatomic transcriptional atlas of human glioblastoma". In: *Science* 360 (6389 May 2018), pp. 660–663. ISSN: 0036-8075. DOI: 10.1126/SCIENCE.AAF2666.

[349] Sharan Narang et al. "Mixed Precision Training". In: International Conference on Learning Representations - ICLR 2018. 2018. URL: https://arxiv.org/pdf/1710.03740.pdf.

[350] Laurens Van Der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605. URL: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[351] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (85 2011), pp. 2825–2830. URL: https://www.jmlr.org/papers/v12/pedregosa11a.html.

[352] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Oct. 2017, pp. 618–626. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.74.

[353] Matthew D. Zeiler and Rob Fergus. "Visualizing and Understanding Convolutional Networks". In: European Conference on Computer Vision - ECCV 2014. Springer, Cham, 2014, pp. 818–833. DOI: 10.1007/978-3-319-10590-1_53.

[354] Julius Adebayo et al. "Sanity Checks for Saliency Maps". In: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). 2018. URL: https://arxiv.org/abs/1810.03292.

[355] John E. McManigle, Raquel R. Bartz, and Lawrence Carin. "Y-Net for Chest X-Ray Preprocessing: Simultaneous Classification of Geometry and Segmentation of Annotations". In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, July 2020, pp. 1266–1269. ISBN: 978-1-7281-1990-8. DOI: 10.1109/EMBC44109.2020.9176334.

[356] Ahmed Harouni et al. "Universal multi-modal deep network for classification and segmentation of medical images". In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, Apr. 2018, pp. 872–876. ISBN: 978-1-5386-3636-7. DOI: 10.1109/ISBI.2018.8363710.

[357] Philipp Lohmann et al. "Feature-based PET/MRI radiomics in patients with brain tumors". In: Neuro-Oncology Advances 2 (Supplement 4 Dec. 2020), pp. iv15–iv21. ISSN: 2632-2498. DOI: 10.1093/noajnl/vdaa118.

[358] Qianye Yang et al. "MRI Cross-Modality Image-to-Image Translation". In: Scientific Reports 10 (1 Dec. 2020), p. 3753. ISSN: 2045-2322. DOI: 10.1038/s41598-020-60520-6.

[359] Alex Kendall, Yarin Gal, and Roberto Cipolla. "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics". In: 2018, pp. 7482–7491. URL: https://openaccess.thecvf.com/content_cvpr_2018/papers/Kendall_Multi-Task_Learning_Using_CVPR_2018_paper.pdf.

[360] Ting Gong et al. "A Comparison of Loss Weighting Strategies for Multi task Learning in Deep Neural Networks". In: IEEE Access 7 (2019), pp. 141627–141632. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2943604.

[361] Erik R. Ranschaert et al. "Advantages, Challenges, and Risks of Artificial Intelligence for Radiologists". In: Artificial Intelligence in Medical Imaging. Springer International Publishing, 2019, pp. 329–346. DOI: 10.1007/978-3-319-94878-2_20.

[362] Bibb Allen, Robert Gish, and Keith Dreyer. "The Role of an Artificial Intelligence Ecosystem in Radiology". In: *Artificial Intelligence in Medical Imaging*. Springer International Publishing, 2019, pp. 291–327. DOI: 10.1007/978-3-319-94878-2_19.

[363] J. Raymond Geis et al. "Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement". In: *Radiology* 293 (2 Nov. 2019), pp. 436–440. ISSN: 0033-8419. DOI: 10.1148/radiol.2019191586.

[364] Sara Gerke, Timo Minssen, and Glenn Cohen. "Ethical and legal challenges of artificial intelligence-driven healthcare". In: *Artificial Intelligence in Healthcare*. Academic Press, Jan. 2020, pp. 295–336. ISBN: 9780128184387. DOI: 10.1016/B978-0-12-818438-7.00012-5.

[365] Vishal Patel. "A framework for secure and decentralized sharing of medical imaging data via blockchain consensus." In: *Health informatics journal* 25 (4 Dec. 2019), pp. 1398–1411. ISSN: 1741-2811. DOI: 10.1177/1460458218769699.