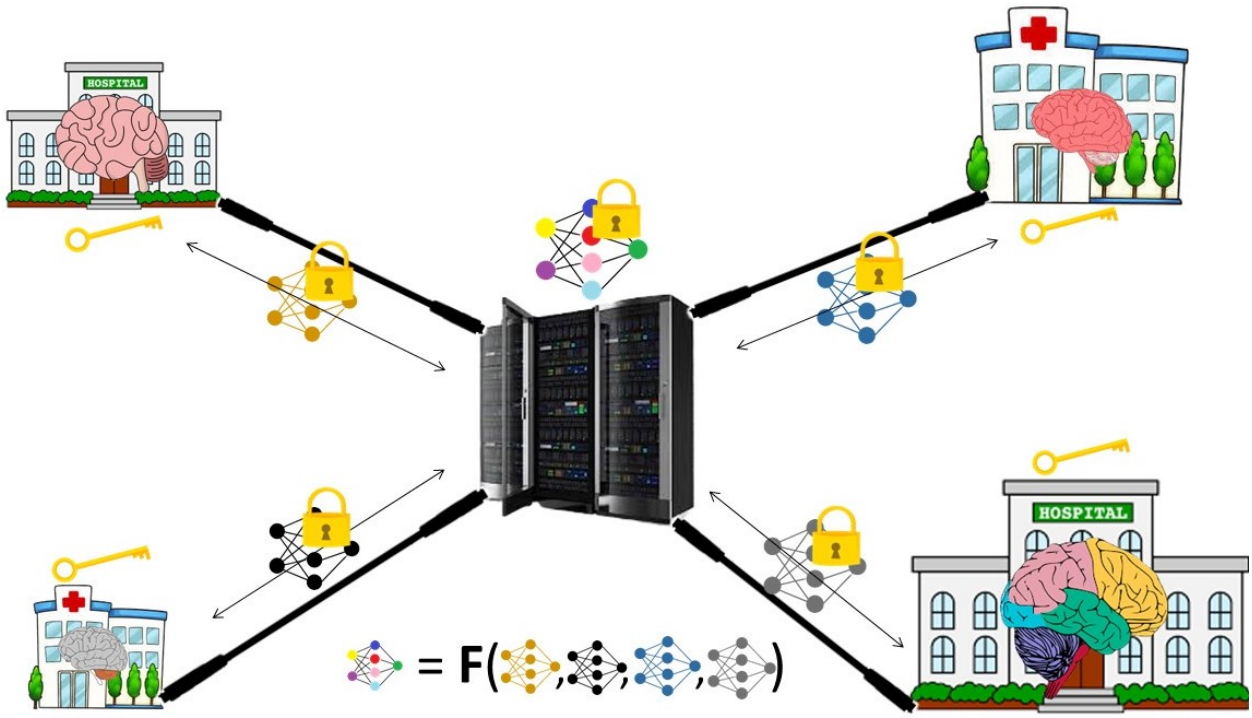
CHALMERS
UNIVERSITY OF TECHNOLOGY



Inter-hospital brain tumour diagnostics using Private Federated Learning

An empirical analysis of convergence in a heterogeneous, non-IID setting and a theoretical review of privacy mechanisms

Master's thesis in Complex Adaptive Systems, MPCAS

LUKAS NYSTRÖM

Department of Physics
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden, June 16, 2020

Master's thesis 2020

Inter-hospital brain tumour diagnostics using Private Federated Learning

An empirical analysis of convergence in a heterogeneous, non-IID setting and a theoretical review of privacy mechanisms

LUKAS NYSTRÖM

CHALMERS
UNIVERSITY OF TECHNOLOGY

Collaborators

Taipei Veterans General Hospital

National Chiao Tung University

National Yang Ming University

Department of Physics
Chalmers University of Technology
Gothenburg, Sweden, June 16, 2020

Inter-hospital brain tumour diagnostics using Private Federated Learning
An empirical analysis of convergence in a heterogeneous, non-IID setting and a theoretical review of privacy mechanisms
LUKAS NYSTRÖM

Supervisor: PhD Mats Granath, Department of Physics, Chalmers University of Technology
Supervisor: MD, PhD Wan-Yuo GUO, Department of Radiology, Taipei Veterans General Hospital
Supervisor: PhD Henry Horng-Shing LU, Institute of Statistics and Data Science and Engineering, National Chiao Tung University
Supervisor: PhD Yu-Te WU, Department of Biomedical Imaging and Radiological Sciences, National Yang Ming University

Examiner: PhD Mats Granath, Department of Physics, Chalmers University of Technology

Master's Thesis 2020
Department of Physics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Illustration of a four party Federated Learning collaboration. Each of the four hospitals has its own unique and heterogeneous data set. They all train individual models that are later sent to a central server for aggregation to create a joint, global model. This process is repeated until global convergence. The entire process conserves patient privacy thanks to the employed privacy mechanisms such as encryption and differential privacy.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden, June 16, 2020

Inter-hospital brain tumour diagnostics using Private Federated Learning
An empirical analysis of convergence in a heterogeneous, non-IID setting and a theoretical review of privacy mechanisms
LUKAS NYSTRÖM
Department of Physics
Chalmers University of Technology

Abstract

This study has investigated the possibility to achieve high performing brain tumour segmentation using Deep Learning, without breaching the strict privacy regulations such as GDPR that governs the use of medical data. This was achieved using a novel technique called Federated Learning (FL) in which models are shared across institutions rather than the sensitive raw data. The aim was to develop an autonomous AI system that can aid medical professionals in diagnosing patients more time and resource efficient. Reducing the cost of treatment is a crucial first step towards a more equal health care.

To achieve the objective extensive empirical experiments using the SOTA 3D ResNet U-NET models were carried out. The experiments were divided into three parts and used a data set comprising 3686 samples, making it almost ten times larger than the commonly used benchmark data set. First a model was developed to work in the common, centralised setting. It achieved human level performance and a median dice score of 0.87. It performs well for all analysed seven sub types of brain tumours as well as on data collected from several different sources. The latter is a key finding, since previous studies has struggled with this due to the large inter-institution heterogeneity in terms of data quality. The second experiment was to extend the centralised model to the federated setting. The data was distributed non-IID across five virtual hospitals. Each hospital first trained a local model on its own data, which lead to only 68.2% of the benchmark performance. Then the five sites trained a joint model using FL and the proposed novel technique adaptive momentum, which was shown to improve the current SOTA. This improved performance significantly, reaching as high as 88.6% of the conventional benchmark.

Finally, although FL does not share any raw data this study highlights several other privacy vulnerabilities as well as techniques for how to protect the system against them. It is shown that by using several layers of protection it is possible to provide complete privacy without any significant loss of performance. The layers include Differential Privacy, Homomorphic Encryption, Shamir's secret sharing, AES encryption and SHA-256 authentication. The study thus shows that it is indeed possible to get human level performance even in a federated, private scenario when the model is trained on non-IID and highly heterogeneous data.

Keywords: Federated Learning, Brain Tumour Segmentation, Computer Aided Diagnostics, Privacy Preserving Machine Learning, Differential Privacy, Deep Learning

Acknowledgements

This work would never have been possible without the support and encouragement from more individuals and institutions than could ever fit in this acknowledgement. To all my colleagues at VGHTPE and all others not mentioned - you know who you are: Thank you.

To the real hero of the story. I am at a lack of words for how grateful I am for what MD, PhD, Wan-Yuo GUO at VGHTPE has done for me. Thank you for believing in me enough to fly me half way across the world to pursue this project and for making my life in Taipei a walk in the park. Not many would have shown that kind of trust in someone after a single afternoon of talks. I will never forget it - Tack!

To MD Ying-Chou SUN, also VGHTPE, thank you for your support in setting up the computational TWCC environment and for aiding me in all hardware and software related matters. I have yet to see a computer science related issue that you do not seem to be able to solve. VGTHPE is lucky to have you on board.

To my supervisors, PhD Henry Horng-Shing LU, National Chiao Tung University, PhD. Yu-Te WU, National Yang Ming University and PhD. Mats Granath, Chalmers University of Technology. Thank you for your support throughout the project and for grilling me with you questions. Without your challenges this project would not have been at the level it is now.

To Ethan TU, PhD Tyng-Luh LIU and PhD Chia-Lin YANG and all the others at Taiwan AI Labs. Thank you for serving as a sounding board in the initial stages of this project and for giving me the idea to pursue Federated Learning.

Finally, to the National Centre for High-performance Computing in Taiwan. Thank you for providing me the facilities and computing power necessary to perform my research. Thanks to your intuitive user interface and flexible container environment I have been able to run tests and analysis in a matter of months that would otherwise take years - if at all - to complete.

The results published here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. A special thanks is also directed to the people behind the BlueLight Viewer: https://github.com/cylab-tw/bluelight/.

This work was supported by the Medical Scholarship Foundation in Memory of Professor Albert Ly-Young Shen, Taipei Veterans General Hospital (Grant Number: T1100200 and V109C-036), and Ministry of Science and Technology of Taiwan (Grant Number: MOST 108-3011-F-075-001).

Lukas Nyström, Taipei, June 16, 2020

Abbreviations

AES Advanced Encryption Standard
AI Artificial Intelligence
AL Active Learning
BRaTS Multi modal Brain Tumour Image Segmentation Benchmark
CAD Computer Aided Diagnostics
CAM Class Activation Map
CCPA California Consumer Privacy Act
CL Curriculum Learning
CNN Convolutional Neural Network
COPPA Children's Online Privacy Protection Rule
CRF Conditional Random Field
CT Computerised Tomography
DICOM Digital Imaging and Communications in Medicine
DL Deep Learning
DP Differential Privacy
DWI Diffusion-Weighted MRI Image
FedAvg Federated Averaging
FL Federated Learning
FLAIR Fluid-Attenuated Inversion Recovery MRI Image
GAN Generative Adversarial Network
GDPR General Data Protection Regulation
HD95 The 95th percentile of the Hausdorff distance
HE Homomorphic Encryption
HIPAA Health Insurance Portability and Accountability Act
IID Independent and Identically Distributed
MITM Man-In-The-Middle Attack
ML Machine Learning
MLP Multi-Level Perceptron
MRI Magnetic Resonance Imaging
PDPA Personal Data Protection Act
PISS Personal Information Security Specification
SGD Stochastic Gradient Descent
SHA Secure Hash Algorithm
SMC Secure Multiparty Computation
SOTA State-Of-The-Art
T1 T1 Weighted MRI Image
T1C T1 Weighted Post contrast MRI Image
T2 T2 Weighted MRI Image
TCGA The Cancer Genome Atlas
TL Transfer Learning
TWCC Taiwan Computing Cloud
VGHTPE Taipei Veterans General Hospital

Contents

List of Figures

List of Tables

1

Introduction

This chapter introduces the general background of the topic to be studied, namely brain tumour computer aided diagnostics. It further discusses why it is - quite literally - a matter of life and death and what the major difficulties are in real world applications. The problem discussion culminates in a declaration of the purpose of the paper and the devised research questions that it intends to answer. The chapter terminates in a discussion on the novelty and academic contribution that the paper provides as well as a brief discussion of what factors the study does not consider.

1.1 Problem background

1.1.1 Inefficient brain tumour diagnostics leads to inequalities and fails to save millions of lives

Each year 17 million people are diagnosed with cancer world wide [7] and it has been estimated that as many as 25% of all deaths in several countries are due to cancer [8]. Accounting for 30% of cases globally it is the leading cause of premature death [9] and it is estimated that the total cost incurred by the disease is several hundreds of billion dollars [7]. To make matters worse the cancer rates are expected to rise, due to an overall adoption of increasingly unhealthy lifestyles [7, 8, 9].

One of the most elusive types of cancer is brain tumours [9]. Although it is only the 17th most common sub type, with 300 000 cases per year, it is one of the world's deadliest diseases due to its much higher mortality rate [7, 9, 10, 11, 12, 13, 14, 15]. The 5 year survival rate in the US is just 33% [11, 15] and the situation is significantly worse in less developed countries that lack the same level of health care [9, 12]. Furthermore, several sub types of brain tumours have mortal outcome in less than a year on average [10, 14]. Besides its mortality, it is the only type of tumour to impact both the body and the mind meaning that even in the case of survival the chance of full recovery is low [9, 12]. Furthermore, the survival rate of brain tumours is strongly correlated to how early the disease was diagnosed, which motivates the need for efficient diagnostics tools [9, 12]. According to [16] the chance of recovery from certain sub types is as high as 90%, given that the tumour is diagnosed at an early stage.

After significant success in improving the survival rates for lung and breast cancer

patients by using MRI techniques for early stage diagnostics [7, 9], it has become the standard protocol for brain tumour diagnostics as well [9, 10, 14, 17]. The invention of different MRI sequences such as T1, T2, T1C, FLAIR, DWI etc. has provided the doctors with a wide array of tools to perform diagnostics since different tissue types, including tumours, show up in different contrast in each sequence. Furthermore, the MRI scan can be performed in axial, coronal or sagittal view meaning that tumours can be examined from several directions. Using these techniques it is possible to identify tumours as small as parts of a ml. However, this method produces a large amount of images for each patient which means that it is a time consuming and error prone work that requires experienced personnel in order to be successful [9, 10, 14, 17, 18].

A natural consequence of this is that the survival rate of brain tumours is strongly correlated to the availability of experience doctors [9, 12]. This leads to a mortal inequality between the rich and poor, as recovery is significantly more likely in high income countries [7, 9]. The survival rate in Denmark is close to three times as high as in Brazil, to illustrate the magnitude of the issue. Reports show that similar inequalities exist within nations as well, due to uneven resource allocation between rural and urban areas [9]. This is true for all types of cancer, but it is further amplified in the case of brain tumours since it is the most expensive type of cancer to treat [13]. The development of affordable, cost-efficient brain tumour diagnostics methods that can be used by less experienced doctors - or by experienced doctors remotely - would thus save millions of lives each year and help bridge the gap between the haves and the have nots [9].

1.1.2 Deep Learning as a tool for CAD is promising but it fails to perform due to lack of medical imaging

The severe inefficiency deficit of manual brain tumor diagnostics was realised already in the 1960's, when researchers turned to developing Computer Aided Diagnostics (CAD) tools as a complement to conventional manual methods [19]. However, the performance of CAD has yet to reach satisfactory levels. Motivated by the recent advances in Deep Learning (DL) [20, 21, 22, 23], a field which has proven very useful and even managed to surpass human level performance in a wide range of related tasks [24, 25, 26, 27, 28, 29], researchers have turned to investigate if it can be used to aid in this task as well [19, 30, 31, 32].

Since then DL has successfully been used for similar medical tasks [19, 33, 34], but it still fails to yield satisfactory results in the analysis of brain tumours in MRI scans despite recent advances in the young field [19, 30, 31]. This inadequacy is due to a number of challenges to be discussed in the sections to come. Nevertheless, the research field of DL for brain tumour diagnostics has grown exponentially over the last couple of years and it is expected to become an invaluable tool for medical professionals in the near future [19, 30, 31]. Several major advantages of DL over the manual diagnostics are quoted, such as that it allows for an objective assessment, increased process heterogeneity and efficiency and that it will be able to detect lesions that are smaller than perceivable by the human eye [19].

The first of several challenges with using DL is that it is a data driven approach [19, 35, 36, 20, 21, 37, 38, 33, 39, 34, 40], meaning that the model must be trained on large amounts of cases in order to properly identify abnormalities such as tumours. Several studies have shown the model performance increases significantly with the number of training samples per class in computer vision tasks [41, 31, 42, 43, 44, 35, 30]. Different authors provide different claims to what is a sufficient amount of samples per class, ranging from a few thousand [31] to orders of magnitude more [44, 43]. Domain specifically, a study on CT scans argue that a CNN model in the medical domain requires at least 5000 cases per class [31]. Some studies, such as [43], show that performance increases logarithmically with the data size. Others show similar monotonically improving results, such as [44] who show that the validation loss decreases as $\sim \frac{1}{\sqrt{\text{Data size}}}$. The general rule of thumb that most studies agree on is thus that more data is always better and that one should have at least on the order of 1000 samples per class to achieve relatively well. Further, the more complex the model is the more data is required. Conversely, if limited data is available one can not hope to fully leverage a sophisticated and complex model [41, 31, 42, 43, 44]. Furthermore, if the data is heterogeneous the model requires even larger training sets to converge [33].

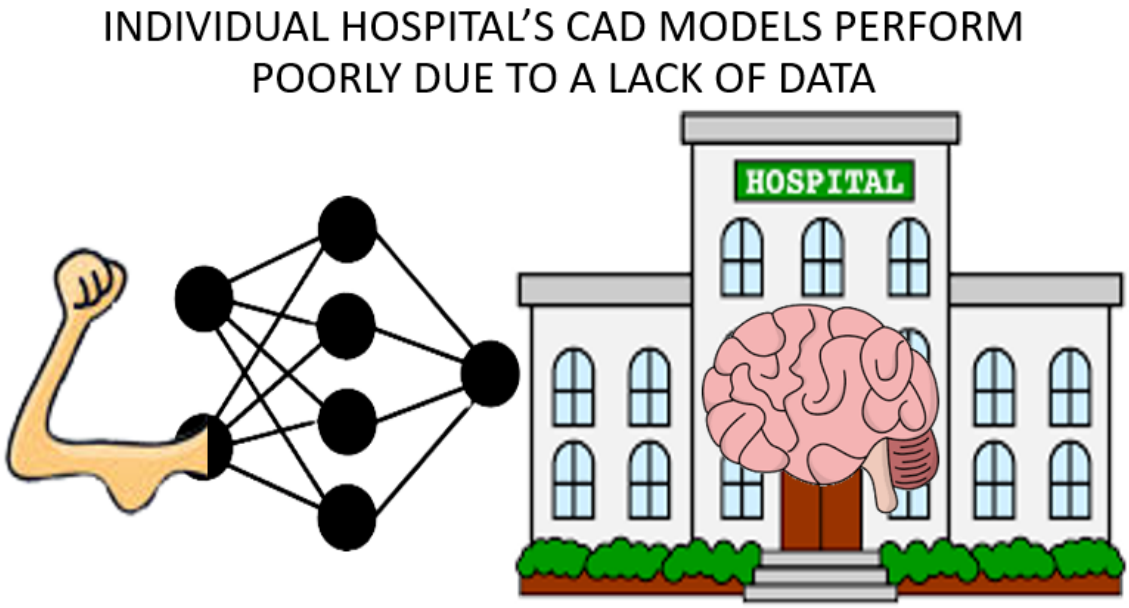


Figure 1.1: Computer Aided Diagnostics for brain tumour segmentation using Deep Learning has yet to yield satisfactory results due to that the local data sets at each individual hospital is too small for such a data driven approach.

The severe lack of training data in brain tumour imaging is emphasised by several sources [36, 21, 33, 45] and is referred to as one of the major reasons why DL has yet to be clinically satisfactory [35, 30, 9, 19]. There are several reasons behind this scarcity. Firstly, compared to natural images of objects that occur in the everyday life there is - thankfully - a relatively small number of brain tumour patients each year, which naturally limits the possible extent [9]. Secondly, it is significantly

more expensive and time consuming to gather annotated samples of brain tumours than of natural images since the former must be done manually by doctors [34, 30, 19, 35]. Furthermore, it is more challenging and subjective since determining tumour boundaries is much more difficult than determining if an image contains a cat or a dog [34]. This means that annotations must be considered as noisy [19, 35]. As a consequence of all these difficulties the leading data set for brain tumour segmentation, (BRaTS), only contains 461 samples [30]. This is well below the rule of thumb, which explains why the current state-of-the-art [6] performs much worse than models in related applications where data is abundant. It has been calculated that it is 30 times slower to collect a sample of an adequately annotated brain tumour patient than of a natural image, which is why BRaTS is 3000 times smaller than its equivalent in the latter domain (ImageNET [46]). Each image in (BRaTS) took 30 minutes to acquire [30].

1.1.3 Privacy concerns prevent attempts to crowd source more extensive data sets

Having established that a DL approach would likely perform well for brain tumor segmentation if sufficient amounts of data could be gathered, one might consider simply crowd sourcing the data from several hospitals to solve the problem. However, this has proven to be infeasible due to the rising public concern for privacy. Forbes performed a nationwide query which showed that 90% of the American population is concerned about their privacy online, which clearly shows the reluctance to share private information [47]. As the global population is becoming increasingly more concerned with its integrity, partially due to the rise of AI [48, 47], new legislative statutes have emerged that limit what the data scientist is allowed to do with the available data [19, 47, 48].

The most famous and restrictive such privacy act is the GDPR [49] which drastically tightens the regulation in the EU since 2018 [48, 19, 47, 36, 50, 45, 51, 52, 39]. However, it is far from the only new privacy act. China recently adopted the privacy act PISS [36], Singapore just passed PDPA [36] and as of this year California has the new CCPA [53] regulation in place [47, 52]. The latter was passed to further strengthen the two nation wide privacy acts HIPAA [19, 54, 51] and COPPA [52] that were already in place. Furthermore, in the wake of these regulations several other countries such as Japan, New Zealand, Russia etc. are now working on introducing new domestic privacy laws to provide its population with better protection as well [48].

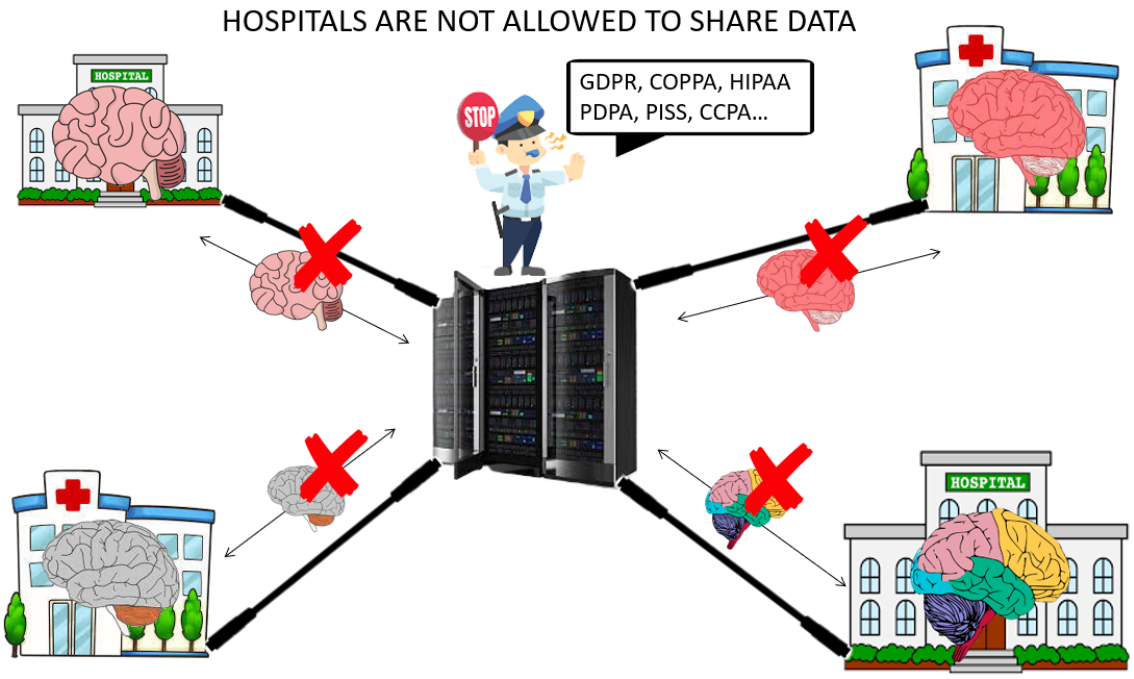


Figure 1.2: Hospitals are not allowed to combine their local data sets due to extensive privacy legislation. This means that it is not possible to overcome the issue of lack of data using this traditional collaboration idea, which implies that conventional Deep Learning is not the solution.

Due to the privacy sensitive nature of patient's medical journals [55, 56, 34, 36] it is thus not allowed to share data outside the hospital [57, 54, 37, 33, 32, 58, 34, 40, 35, 19, 31], see Figure 1.2 for an illustration. This makes collaborative data collection attempts impossible. The conventional Machine Learning approach - collecting massive data sets to one server and performing model training there before sending the final model back to the hospitals - is thus not a feasible solution [19, 59, 50, 60]. Consequently, in order to fully leverage Deep Learning techniques for brain tumor segmentation new approaches must be proposed.

1.1.4 Federated Learning is a promising attempt to solve this by using decentralised training

The current situation is that data stored at each hospital is much too small to be used individually to train the model [36, 21, 33, 45] and legislation prevents the data sets to be pooled together [57, 54, 37, 33, 32, 58, 34, 40, 35, 19, 31]. Similar scenarios in other applications has brought researchers to consider training their model in a decentralised fashion [61, 62, 63, 64, 65, 66, 67, 68, 69, 70], rather than in the conventional centralised setting. The general idea is to bring the model to the data instead of the other way around. This means that the sensitive data does not have to be shared across institutions, thus circumventing the privacy regulations, but the model is still exposed to all data sets thus effectively increasing the total training size [71, 39, 72, 73, 35, 19].

Different studies have proposed a number of different methods for such a decentralised DL framework for training on none-shareable data. [33] proposed one, [55, 36] used the Private Aggregation of Teacher Ensembles (PATE) design and [32] describes Split Learning, Large Batch synchronous SGD and NoPeekNN just to mention a few. Although they have all shown some success, none of them can be compared to the state-of-the-art (SOTA) solution [38, 37, 52, 40, 74, 50, 1, 60] referred to as *Federated Learning* (FL), initially proposed in 2016 by [71]. See Figure 1.3 for an introductory illustration of how the technique works.

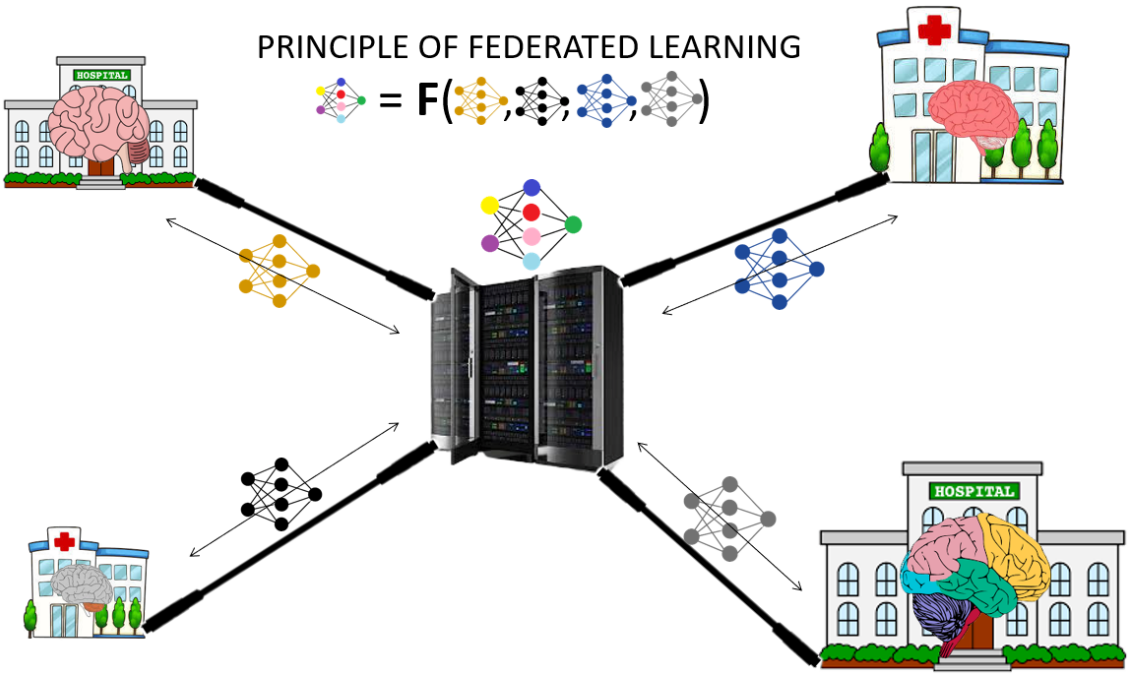


Figure 1.3: A holistic illustration of the principle behind Federated Learning. The idea is to share local models and aggregate them, instead of sharing the raw data since the latter is prohibited by privacy regulations.

Ever since, Federated Learning has been receiving a lot of attention. Tech giants such as Google, Facebook and NVIDIA [75, 76, 77] are already working on deploying FL in their consumer products. In the wake of the hype several start ups [78, 79, 80, 81] have been founded and a number of dedicated software frameworks [82, 76, 75, 77] are being released. More importantly, researchers have published in a number of promising publications on medical applications [63, 64, 65, 66, 67, 68, 69, 70, 61, 62]. Several sources predict that FL will become an invaluable CAD tool for health care in the near future, due to its ability to leverage large data sets without breaching privacy [83, 36, 54, 36, 33, 57, 32, 58, 23, 40, 84, 85].

With that said, the field of Federated Learning is still young and has several fallacies that must be studied more closely before its anticipated aptitude can become a reality in a clinical setting [35, 71, 30, 36]. Firstly there are a number of practical issues that must be solved on how to efficiently implement the framework in a real world scenario. These relate to the fact that geodistributed computation has

several drawbacks. To mention a few it incurs severe communication overhead, it has to be able to deal with synchronisation of autonomous clients as well as client heterogeneity. These and more are discussed at length in section 2.4.2. The second open question that must be solved is how to ensure that the model converges even when the data distributed across clients is non-IID and unbalanced, which has proven to be far from trivial. As described in section 2.5, this happens to be the case for brain tumor data due to demographic differences across hospitals. Studies have shown that the performance of FL can be degraded by as much as 70% if the data is heterogeneous enough. This is discussed in section 2.6. Lastly, it turns out that even though no data is shared across hospitals in FL there are still significant privacy threats that must be countered. Bare bone FL simply does not provide adequate protection of patients' integrity to be allowed in a clinical setting. The threats are presented in section 2.8 and proposed solutions are discussed in section 2.9.

1.2 Purpose and Research question

The study intends to implement a comprehensive framework for performing Privacy Preserving Federated Learning for computer aided diagnostics of brain tumours. The aim is to develop an autonomous AI system with human level performance that can aid medical professionals in identifying and segmenting tumours in MRI scans. It is envisioned that such a CAD model will be a step towards a cheaper, more accurate and more equal health care.

The study intends to investigate three separate aspects in depth. Firstly a comprehensive empirical analysis of different conventional, centralised Deep CNN models will be conducted to investigate which setup is ideal for the task at hand. The analysis will focus on how to overcome convergence issues caused by heterogeneous, unbalanced and low quality data across different participating hospitals. The model will be trained on 3D MR images of actual patients collected from multiple different sources in order to simulate the real world scenario of heterogeneous data across institutions. The expected outcome from these initial experiments is to find a model architecture that is high performing and robust which can serve as the backbone in the federated system. The first research question is hence:

How is model convergence in the ideal centralised setting impacted by the heterogeneous, unbalanced and low quality data that arise in real world collaborations across hospitals?

The second part of the study turns to analyse model performance in the federated setting. A detailed examination of different state-of-the-art techniques will be conducted in order to find a configuration that is adequately capable of handling all of the data related issues mentioned above in addition to the non-IID distribution across participating client. The focus will be on how to ensure global model convergence despite the statistical issues caused by the non-IID assumption. Since the focus of the study is on model convergence, it suffices without loss of generality to realise the framework in a virtualised container environment. This gives rise to the second research question:

Can Federated Learning yield adequate performance despite the non-IID distribution across participating hospitals, thus making it the best available option for real world applications?

The final pillar of the study concerns how to ensure that the strict privacy regulations are not breached by such a federated system. As will become evident this is mainly a computer science related matter that does not impact model performance at all, with the exception of Differential Privacy. This motivates a purely theoretical review of the subject. However, a minor empirical study on the impact of Differential Privacy on convergence will be conducted for completeness of argument. Besides this a comprehensive review of available and necessary privacy mechanisms will be conducted in order to evaluate which layers of protections must be put in place to ensure complete system integrity from malicious attackers. This includes amongst other things a discussion on different encryption and Secure Multiparty Computation techniques. The final research question then becomes:

Which privacy mechanisms are necessary in order to ensure complete protection of the federated system against malicious attackers?

1.3 Academic contributions and related studies

1.3.1 The study presents novel results that suggest that previous centralised studies are not realistic

The field of Deep Learning for brain tumour segmentation has been given significant attention over the last decade and several studies have been able to show that it is feasible to create models with human level performance. However, these have all been focused on the common benchmark data set BRaTS, with a few unsuccessful exceptions of minor studies on in-house data. This leads to the question of whether results on BRaTS are actually generalisable to real world conditions. The reason behind this relates to the matter of heterogeneous, low quality data discussed below.

The data set under consideration in this study has several characteristics which makes it significantly more applicable to real life scenarios than the BRaTS data. Firstly it is about ten times larger in size. Secondly it consists of cases collected from six different sources, including all BRaTS samples, and contains seven different tumour sub types all with their own unique characteristics. Importantly, it is the first study to consider Metastasises. This signifies that the model must be able to identify several lesions per patient as opposed to just one and that each lesion can be significantly smaller than before [9, 10, 11, 13, 14]. The BRaTS data only considers two very similar sub types and can be viewed as collected from a single source, due to the extensive attempts of homogenising. The study thus pioneers by considering a more extensive and inclusive data set.

Furthermore all cases in BRaTS have been quality controlled and cross checked by several independent sources, meaning that the annotation quality is ideal. This is not the case for the data set in this study where annotations have been done by

individuals without any synchronisation or oversight. Due to this, significant errors are present and a number of different tumour extent definitions, even for the same tumour sub type, have been used which leads to a non-negligible ambiguity. Furthermore the data set consist of both pre- and post surgery cases, which complicates matters even more. To conclude this implies that the data set in this study is more likely to represent real world conditions, as opposed to BRaTS which is an idealised scientific setting designed for model exploration.

The first research question then boils down to whether these data quality issues can be bridged using models developed for the BRaTS scenario. As it turns out, the results from this study leads to the conclusion that they actually can. The first academic contribution is thus that it is indeed feasible to build a centralised Deep CNN that can adequately handle real world data. However, the results are significantly lower than those on the BRaTS data. The second contribution is thus that the study provides evidence that suggests that the BRaTS data represents an idealised, non-realistic scenario.

1.3.2 The study provides the first proof of concept of Federated Learning in real world conditions

Turning to the second research question. As opposed to the centralised setting, the field of Federated Learning is young in general and requires extensive efforts before it proves as useful as conventional Deep Learning methods [71, 30]. Although several studies have initialised the research on how the technique might be employed in the medical domain [63, 64, 65, 66, 67, 68, 69, 70], there is still many obstacles to overcome before it can be clinically employed [30]. The exploration of its capabilities in the specific field of brain tumour segmentation is even further behind, with only two publications to date [61, 62]. This in itself motivates additional studies of the subject. Both the previous studies showed that non-private Federated Learning can perform on par with the centralised setting using the current SOTA model out of the box [61, 62]. However, as in the centralised case discussed above these experiments were only considering the BRaTS data. This study thus provides a novel contribution by being the world first at analysing how Federated Learning performs in real world conditions.

Both of the previous studies also studied the non-IID scenario, however they only considered a single tumour sub type. This implies that the data is only non-IID in terms of where the original source of samples were. However, as discussed above the inter-source difference in BRaTS is significantly less than in the data set considered in this scenario. Furthermore, this study distributes cases non-IID based on tumour sub types, motivated by the correlation of said sub types with demographic differences. This implies that the samples across federated clients is heterogeneous in terms of both tumour characteristics as well as data quality. The former relates to the fact that different institutions use different annotation protocols and MRI scanner, each with its own unique set of characteristics. A consequence of this is that the data distribution across clients in this study more closely emulates what a real world application would experience. The conclusion and contribution is yet

again that these more realistic settings lead to a significantly more difficult model learning scenario. It is shown that the federated model can still perform approximately at a human level, however the performance loss relative to the centralised setting is much greater than reported in the previous papers. A final note on the academic contribution is that the study proposes a novel federated aggregation technique termed adaptive momentum. According to the empirical results this method was superior to previously reported state-of-the-art methods.

1.3.3 The study contributes with a more comprehensive analysis of privacy threats and protection mechanisms

The final part under scrutiny is also the most previously analysed. The reason for this is natural, since the privacy threats and protection mechanisms are not as application dependant. This implies that ideas and insights from a wide array of fields can be used for this specific case. A consequence of this is that the theoretical review that this study lays out does not contain any novel elements per se, but should instead be viewed as an attempt to synthesise relevant prior findings in one place. This includes the unifying conclusion that based on the threats associated with a federated learning system for brain tumour diagnostics it is necessary to employ both Differential Privacy mechanisms, Homomorphic encryption and potentially Secure Secret Sharing techniques. By employing these complementary levels of protection the federated system is fully protected from malicious attackers.

Comparing these conclusion to the previous two related studies it is evident that they have not properly considered the full extent of privacy. The initial paper did not consider any additional privacy mechanisms at all [61]. The second paper is slightly better in the sense that they analyse how Differential Privacy and partial model sharing can prevent privacy leakage. Similar to this study they also analyse the trade off between performance and the level of Differential Privacy [62]. Their results are conclusive with this report in that excessive Differential Privacy leads to unacceptable model degradation. This results is what motivates the need for complementary privacy mechanisms, since DP on itself will never be enough if clinical level performance is required. Furthermore, none of the prior studies have considered the vulnerability to man-in-the-middle attacks [86] or the troubling assumption that a trusted aggregator exists. A novel contribution from this study is thus that these restrictive, unrealistic assumptions can be loosened.

1.4 Delimitations

The main goal of the study is to illustrate the feasibility of using Federated Learning in clinical brain tumour CAD system. The purpose is in no way to perform a comprehensive grid search over all possible parameters nor to examine a wide array of different model architectures. Instead the study relied on previous results and simply chose the current state-of-the-art model from the centralised setting. Since this was shown to perform relatively well, no further fine tuning was performed. However, since the federated aggregation is less explored, more extensive experiments and

tuning was done in this regard. To conclude the results should thus be viewed as an initial proof of concept that future work can rely on, rather than as the best practice.

All experiments were conducted in a virtualised environment in which practical issues such as synchronisation, stragglers, client autonomy etc. can be disregarded. As discussed this is due to that the study focused on convergence related matters, however it implies that additional studies on physical mock-ups must be performed in order go from theory to practice. The same goes for the presented privacy mechanisms. With the exception of Differential Privacy these have not been implemented and are instead left as theoretical constructs. However, these additional implementation steps are believed to be a relatively straight forward task since these are all well documented and standard protection techniques.

Another practical issue that the study does not consider concerns data acquisition. The used data was initially highly heterogeneous in terms of how it was stored, which formats were used and how annotations were provided. It required significant effort to convert everything to one homogeneous protocol that could be interpreted by the model. However, this process would have been much more problematic in the real world where the data scientist does not own the entire data set. Somehow all individual data owners would have to be in agreement and synchronised regarding how to physically store data. Thankfully this issue does not challenge any of the presented results, but it poses a significant practical obstacle before a true federated system can be rolled out.

A final and important notion concerns data quality. Since the data scientist in this study had complete ownership of the entire data set, it was relatively simple to remove the bad apples. However, in a real world scenario the hospitals are autonomous which means that no guarantees on their individual levels of data quality can be assumed. To move from a theoretical to a clinical federation this implies that additional mechanisms must be put in place to ensure that the local data qualities are adequate. Otherwise there is a risk that some participating hospitals have substandard samples with faulty annotation, which leads to a global model degradation. The theoretical review in this study presents some initial work on how to protect the federation from such inadequacies, but none of them are implemented. Besides, very limited work in general in this regards has been conducted by the academia. The last piece of the puzzle would thus be to study this in-depth and to evaluate such an implementation.

2

Theoretical Framework

As the name suggests this chapter present the theoretical framework required to answer the posed research question. The first two sections discuss and describe the state of the art model for conventional brain tumour segmentation in the centralised setting. They provide sufficient background to understand how Deep Learning can be applied as a CAD system. The five sections to follow turn to discuss the core of the study, namely Federated Learning (FL). The definition of FL is provided as well as a discussion on what practical issues must be solved in order to achieve good performance. Suggested solutions to these difficulties are presented, before the chapter turns to discuss the system design. The remaining four sections of the chapter are dedicated to ensure that privacy is preserved. The threat scenario is presented together with a number of proposed countermeasures. Primarily, it will be discussed how the use of Differential Privacy mechanisms and encryption can further strengthen the integrity of patients.

2.1 Conventional centralised Deep Learning for brain tumour segmentation

As discussed in the background, the goal of any CAD system is to autonomously identify all tumours in an MRI scan. In addition to identifying the lesion it is also important to outline the exact extent of the tumour, thus requiring the model to label each pixel as positive or negative. This pixel level detail is referred to as semantic segmentation. Translated to the domain of Deep Learning this means that the model objective is the following. Given a 2D or 3D image of a brain, output a 2D or 3D binary image that indicates if the pixel is abnormal or normal tissue. This study concerns seven different tumour sub types, but since the main objective is segmentation the model will only be tasked to make a binary prediction. As seen in Figure 3.6 the different tumours visually look relatively similar to a layman, which supports such an approach. The prediction for each pixel is a confidence score between 0 and 1 that represents the likelihood that the pixel is tumourous.

A large number of models and methods have been proposed to solve this objective. This section is dedicated to describe and compare these different approaches. The aim is to survey the results from the research community in order to find the current state-of-the-art model that will serve as the backbone of the Federated Learning

Framework. As described briefly in the background FL extends the conventional, centralised model training by training on multiple local clients in parallel. However, the local training at each FL client is still identical to to the conventional setting. Without loss of generality - assuming all other things equal - one can thus argue that whatever model performs well in the conventional setting should have a good chance at performing well in the federated setting.

2.1.1 U-Net models are the state of the art

As discussed, virtually all research uses the BRaTS data set [87]. From the extensive survey in this study only three papers were found that used other, external data sets [88, 89, 90]. Another survey that investigates the progress and history of the field during the last five years they also conclude that BRaTS is indeed the most commonly used [16]. As a result, the appropriate benchmark when comparing models is their performance on said data set. The current standing is presented in [6] which reveals that the state of the art model has a dice score of 0.913. Furthermore, a total of 27 papers present performance above a score of 0.9. This shows that several good models are available for the task at hand. However, as discussed briefly in section 2.2, the results on BRaTS might not be adequately representative for real world conditions. Unfortunately several of the result in the leader board are presented with pseudonyms, which means that their papers are not available. Nonetheless, several important papers have been found. The winner and the fifth place of the challenge in 2018 is [4] and [91] respectively. The second and fifth place for 2019 is [92] and [5]. These four articles will be referred to as the leading models throughout this section and as will be shown they all have significant similarities.

From the five year survey it is clear that virtually all kinds of models, both machine learning and handcrafted, have been proposed to solve the objective [16]. However, during the last years most of the emphasise has been placed on Deep CNN models [16], which have won the BRaTS challenge all of the last five years [93]. The most commonly used model is the U-NET proposed in 2015 by [94] specifically to deal with the intricate difficulties in medical imaging. It is a multi scale CNN based model that builds on the realisation that segmentation requires the model to extract both large scale (global) and small scale (local) information about the input. A visual illustration of the model is presented in Figure 3.18. The general idea is to sequentially down sample the input and process it in the lower dimensional space, since this allows for more computationally exhaustive techniques. However, the model also retains all higher dimensional representations as well and finally concatenate them. Thanks to this multi-scale processing the model manages to identify a wide range of relevant features [95, 90, 94, 96].

This approach proved to outperform all other proposed models, especially when limited data is available, and its extensions have been the state of the art ever since. Due to its proven aptitude it is used by [97, 98, 92, 99, 100, 96, 101, 91, 61, 93, 102, 94, 95, 89, 103, 90, 104, 105, 91, 4, 106, 62, 107, 108, 109]. According to [100] practically all research teams use the U-NET in some manner, a statement which is supported by [91] who claim that it is the 'classical approach'. Further motivation is

given by [4, 90] who claim that even a generic, baseline U-NET provides competitive results for a wide array of segmentation tasks. Its aptitude is further supported by the fact that all of the leading models use it as their backbone [92, 4, 5, 91].

Despite U-NET's majority share, it is not the only proposed method currently in use. Another common model is called DeepMedic [96, 101], which is a multi-path CNN where each path is designed to extract information on different scales. Such a multi-path CNN approach is besides U-NET the most common method, used and praised by [110, 111, 112, 96, 101, 104]. In general it is clear that the success to high performing segmentation is to properly extract both local and global information. Other less performing methods that are found in the literature are Random Forest based models [88, 113], Fully Connected CNN [113], Growing CNN [88] and sliding window based ideas [110]. Besides Machine Learning approaches a variety of hand-crafted, statistical methods have also been proposed [110, 114, 90, 88, 113, 16], such as the Superpixel Segmentation Algorithm [90] and NS-EMFSE [114] or to extract features with Stationary Wavelet transform [88]. However, these have been proven to perform worse than their machine learning counterparts [90]. Furthermore, they are slow at inference time [90] and might take up to 100 minutes per image [110].

2.1.2 Leveraging ResNet, ensembles and 3D inputs further improve model performance

Having established that U-NET is the state of the art, researchers started to extend and improve the model further. One of the first extensions was to modify it to take 3D inputs, since MRI scans are mostly three dimensional [99, 95, 93]. This is further motivated by [95, 99, 93] who all claim that neighbouring 2D slices contain redundant information which makes the model less efficient. Additionally, as discussed by [93] the 3D model can leverage richer spatial context. Although isolated studies claim that 2D models are better [95], the general consensus is that 3D models are indeed superior [99, 93, 98, 104], which is supported by the fact that all of the leading methods use 3D inputs [92, 4, 5, 91]. The downside of 3D models is that they are computationally expensive [93, 98] and require a lot of RAM [5]. The latter implies that the batch and resolution has to be reduced. Due to these drawbacks the research field is divided, with approximately half working with 2D U-NETs [95, 99, 93, 61, 94, 89, 103, 90, 104, 105] and half on 3D-U-NET [95, 96, 97, 91, 4, 106, 92, 98, 62, 107, 108, 109]. However, if computational power is not a bottle neck the 3D version is superior.

A second improvement to U-NET is motivated by the recent advances of the famous ResNet model. Several authors [102, 62, 4, 92, 97, 108, 107, 99, 104, 105] have proposed to replace the convolutions in U-NET with residual blocks as a way to handle the vanishing gradient problem [105, 104, 99, 97] and provide better scale invariance [105]. This has indeed proven to be a successful approach, which is supported by the fact that a majority of the leading methods use residual blocks [92, 4, 5, 91]. A final improvement, again motivated by its success in other related tasks, is to use ensembling to capture the individual traits of several models [91]. This is used successfully by several papers [92, 4, 5, 109, 91, 101] once again including

all of the leading methods.

Besides using 3D inputs, ResNet and ensembling a plethora of other extensions have been proposed. Once such technique is to introduce multiple loss functions to explicitly encourage specific model behaviours. A successful approach used by the winner 2018 [4] is to append a variational auto-encoder to the U-NET, thus forcing the model to recreate the input. The intuition behind this is that it serves as regularisation, since it encourages the encoder part of U-NET to compress more detailed information about the input. Similar ideas are proposed by [96, 98] who both append various auxiliary loss functions at the lowest resolution in the model. Again, the underlying idea is to force the encoder part to create more useful encodings. Another proposal by [5] has another take. They instead append a 3D extension of supervised active contour loss to amplify model regularisation. However, the efficiency of all these methods has yet to be proven. So far the performance boosts are only minor, if at all.

A more general notion is that deeper networks have been shown to work better [102], similar to what one expects from other related tasks. An interesting study by [99] further attempts to increase the U-Net's ability to identify small lesions by introducing the Attention mechanism. The method has so far only been able to present minor improvements, but considering its efficiency in other fields it might be worth looking into in future work. Other proposals are to modify the U-NET to have an asymmetrically large encoder part [4], to include both down sampled versions of feature maps and the input at each stage [92] and to create a unique model for each tumour sub type and to first classify the input and then feed it to the respective segmentation model [107]. However, they barely impact performance. A final approach that is worth mentioning was recently introduced by [109]. They argue that tumours have some common traits, which means that it should be beneficial to introduce a tumour prior. The intuition behind their proposed algorithm is that besides feeding the input image to the model, they also feed a 3D heat map that represents the likelihood of finding a specific sub type at each location. This heat map is an aggregate of all tumour masks in the training set. Although this idea makes intuitive sense, it has yet to show anything but minor improvements.

2.1.3 Image preprocessing is vital whereas postprocessing might provide small improvements

As discussed in the background chapter the data is often collected from different sources with different protocols, resolutions and quality. In an attempt to make the data set more homogeneous, which in turns simplifies model fitting and improves performance, researchers commonly use some form of preprocessing of images before feeding them to the model [112, 111, 4, 98, 97, 113, 96, 107, 103, 91, 90, 110, 109, 88, 114, 99].

The vast majority of studies use simple intensity normalisation by subtracting the mean and dividing by the standard deviation, thus at least forcing all inputs to have the same approximate intensity magnitude [99, 112, 111, 97, 113, 96, 107, 103, 90,

110, 109]. Some authors expand on this by only calculating the mean and standard deviation based on the pixels encapsulated in the actual brain, thus disregarding the background [4, 98, 91]. Yet another study took an even more sophisticated approach by using linear neighbour interpolation based normalisation (SEB) [88]. Furthermore, another paper instead simplified the normalisation by just dividing by the median of the pixels encapsulated in the brain [93]. A final notion is that some authors remove pixel outliers by clipping before doing normalisation [99, 109]. Nevertheless, to conclude some form of intensity normalisation is almost always performed.

Beyond intensity normalisation some researchers propose the use of N4-ITK bias field correction [112, 97, 107, 90, 110] or less commonly Wiener filtering [114, 88] to remove noise and produce smooth edges [88]. Others argue that skull-stripping, by the use of handcrafted methods, should be performed [88] as has been done for the BRaTS data. A final and more extreme proposed preprocessing method is to threshold the grayscale input into a binary image [88], however this is the only paper that references this method.

Whereas preprocessing is adopted to make the input data more homogeneous, post-processing instead works on the output in an attempt to further boost performance. It is a less common technique than the more or less mandatory preprocessing step, but it is still used in some form by several authors [112, 100, 96, 101, 111, 16, 90, 110, 105]. The idea is to process the output segmentation mask to remove false positives [110] and to smoothen predicted tumour regions [112, 100] using prior knowledge of how tumours look and where they occur. The efficiency of post processing is ambivalent, with some authors reporting minor improvements [96] whereas others claim that it does not impact performance at all [4].

The most common form of post processing is to use a handcrafted method called Conditional Random Field (CRF) [96, 101, 90] or its extension fully connected CRF [111, 16]. Since is is handcrafted method it is not ideal, and its efficiency is debated [96, 4]. Another less common technique is to use trainable MLP or Random Forest based methods to overcome this [96]. Yet another paper proposes to do postprocessing by simple thresholding, in the sense that tumours that are larger or smaller than a set proportion of the image are discarded [105]. Lastly, [110] propose a method that they refer to as "erosion and dilation". To conclude, however, the common point for all these postprocessing methods is that their efficiency is debatable. Whether they work or not seem to be application and model dependent. However, especially CRF based methods might be worth considering if it is important to squeeze out the last percentages of a model's performance.

2.1.4 Augmentation makes the model more geometrically and intensity invariant

As discussed by [112, 94] a CAD model must be geometrically and intensity invariant in order to generalise. This is especially true for medical imaging due to the deformative nature of abnormal tissue [94]. A common method to mitigate this issue is to use

image augmentation [89, 108, 95, 97, 4, 90, 91, 107, 92, 112, 94, 98, 106]. This means that an image is transformed by different techniques in order to artificially amplify the variability of the training data. Most studies show that this indeed improves performance [95, 90, 112, 89], however excessive augmentation degrades it because it leads to overfitting [90]. Furthermore, not all types of transformations are useful, as discussed by [5] who argue that it suffices to use naive and basic transforms since their extensive study of more sophisticated techniques did not show any additional improvements. Lastly, although the vast majority of studies only do augmentation during training it is also possible to do it during inference [4]. The rational is that by averaging the result from several transformed versions of the unseen image, the chance increases that some of them are closer to any of the training samples that the model was fitted to thus improving the prediction. In their study [4] showed that this indeed slightly improved test results.

As mentioned augmentation can either be geometric or intensity based, with the former being more common. Geometric versions were reported to be used by [89, 108, 95, 97, 4, 90, 91, 107, 92, 112, 94, 106] compared to that the latter was used by [106, 89, 4, 98, 108, 91, 95, 112]. The geometric methods intend to provide invariance against rotation, translation, shear, scale, skew etc. Intensity methods on the other hand manage noise, obfuscation, enhancement, sharpness etc.

The most commonly used geometric augmentation method is mirroring [89, 108, 95, 97, 4, 90, 91, 107, 92, 112], in which the image is flipped across any of the three axis. The second most common method which is to do arbitrary rotations [108, 95, 97, 92, 90, 107, 112]. To manage different scales it has further been proposed to do random cropping [89, 4, 95, 108], up sampling [108] or down sampling [107, 108]. Less common varieties is lastly to use different kinds of non-linear transformations or elastic distortions [106, 90, 94, 112].

When it comes to intensity based augmentation the most common method is to add Gaussian noise to regularise the model [91, 95, 112]. Other popular methods are to do intensity shifting [89, 4, 98] or to scale pixel values by some factor either uniformly across the entire volume [89, 4, 98] or non-uniformly for each slice [108]. Less commonly reported, researchers also use sharpening and emboss [112], edge enhancement [89, 112] or pixel shuffling and obfuscation [106].

2.1.5 Optimisation is most efficient if Adam, soft dice loss and regularisation is used

For segmentation tasks there are primarily two choices of loss functions. One option is to use the classical cross entropy loss [99, 109, 93, 96, 111, 104, 94, 97], but more recent papers have proposed to replace it with the soft dice loss [91, 90, 103, 89, 95, 107, 92, 4, 5, 62]. Some authors argue that the dice loss is better at optimising an unbalanced problem [103], whereas others on the contrary are concerned with that it is less robust [97]. However, the fact that all of the leading methods use the soft dice loss serve as evidence that it is indeed superior [91, 92, 4, 5]. The reason for this is likely the fact that one of the major difficulties with brain tumour segmentation is

that the class imbalance is very severe. This is illustrated by that 98.46% of pixels in the BRaTS data set are of non-tumorous matter [99].

The binary soft dice loss is defined in equation 2.1 below [91, 90, 103, 89, 95, 107, 92, 4, 5, 62] where $p \in [0,1]^{\text{dim}}$ is the flattened vector of predictions and $y \in \{0,1\}^{\text{dim}}$ is the ground truth. ϵ is a small constant to avoid division by zero.

$$\text{Soft Dice Loss(p, y)} = \frac{-2\sum_i p_i y_i}{\sum_i p_i + \sum_i y_i + \epsilon} \tag{2.1}$$

Several papers propose to use weighted loss functions [99, 94, 99, 104, 93, 97, 109] to impose some prior on the result. The work by [109] attempts to mitigate the severe class imbalance by weighting the loss such that the non-tumorous regions can never have more than 3 times the impact of tumorous regions on the back propagation. Another proposal is to weight the pixels in the region close to tumour boundaries higher as a way to force the model to properly identify the tumour edges [94]. However intuitive these methods might seem, their efficiency has yet to be proven since none of the leading models use any weighting. When it comes to the choice of optimiser the vast majority propose to use Adam [108, 98, 89, 4, 61, 62, 90, 91, 92, 109], although some settle for momentum SGD [99, 93, 94] or even simple SGD [110, 96]. However, all leading models use Adam [4, 92, 91, 5].

Besides that the loss function has to be adequate it is important to consider model regularisation. The most common, and successful, approach is to use drop out [109, 4, 92, 61, 110, 111]. Furthermore different kinds of normalisation is often used. The conventional Batch Normalisation is used by [110, 5], however due to memory constraints that limit the batch size it is not very efficient [4]. Rather, the research community suggests to use Instance Normalisation [91, 92, 5] or Group Normalisation [4, 5, 109]. Lastly, some of the leading papers suggests to use L2 norm regularisation [4, 109].

2.1.6 Model evaluation uses the Sørensen–Dice coefficient

The model performance can be evaluated using several different metrics. Which metric is preferable depends on the application, the class imbalance, the task etc. Although it might sound like a trivial question this is relevant to consider in order to properly quantify how well a model is suited to a certain task, in order to avoid cognitive bias or over/under appreciation of the results. When it comes to semantic segmentation the most commonly used metric is the Sørensen–Dice coefficient [107, 91, 90, 110, 61, 89, 96, 16, 105, 100, 108, 111, 104, 93, 113, 109, 97, 98, 92, 5, 99, 95], commonly referred to as just the dice score or the F-measure [108]. Another common metric is to use the 95th percentile of the Hausdorff distance (HD95) [91, 113, 109, 92, 5, 99, 95]. Both the dice score and HD95 are metric that describe how similar two sets are, which makes them useful for comparing a predicted tumour mask to the ground truth. However, since they are complementary this study will settle with only using the more occurring dice score for simplicity. The dice score is defined in equation 2.2 [107, 91, 90, 110, 61, 108, 111, 104, 109, 97, 92, 98, 99, 95, 93, 89, 96, 16],

where X and Y are the predicted and ground truth binary masks respectively and $|X|$ is the cardinality of set X.

$$\text{Dice score} = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2.2}$$

Another set of metrics that are commonly used are sensitivity (also referred to as recall) [90, 110, 16, 96, 105, 108, 111, 104, 113, 95, 97, 92, 5], specificity [110, 96, 108, 111, 113, 95, 97, 92, 5, 104, 16] and precision [90, 16]. These are metrics to describe performance on instance level, such as to present how many of the tumours were found. They are less suited to report how well the segmentation is but it is a good complementary tool to the dice score/HD95 if the intention is to describe how well the model would work in a clinical setting. However, for this study these metric will not be included for simplicity as the main focus is on achieving good segmentation.

2.2 The main obstacle for conventional methods is the lack of high quality data

The research community is largely in agreement on why Deep Learning has yet to perform satisfactory in the field of brain tumour segmentation. As discussed in the background the models, no matter how sophisticated, are severely degraded if the data availability is low. Unfortunately this is the case for the task at hand [36, 21, 33, 45, 35, 30, 9, 19]. The reason behind the limited data sets is that data annotation is time consuming, expensive, difficult and error prone even for experienced doctors [93, 108, 96, 107, 90]. Different sources provide slightly different estimates to quantify this. According to [30] a sample takes on average 30 minutes to label and according to [93] it can take up to 45 minutes for even an experienced oncologists. More over, the problem is amplified by the fact the different hospitals use different protocols for their MRI scans [93]. As discussed in the background the scan can be taken in several views and with different modalities, all of whom have unique characteristics. Since the model must be fed with images from the same modality and view, the collection is restricted to only hospitals that follow the same protocol which according to [93] poses a large practical obstacle.

Besides the fact that data sets are small, it turns out that the labelling is highly subjective and depends on the doctor's personal experiences and bias which means that the existence of a gold standard is not even always guaranteed [96, 61, 107, 93, 90]. The variability of tumour appearances, its diffuse edges and the difficulty of simultaneously analysing several different MRI modalities leads to large inter-rater disagreement [107, 110]. Studies have shown that the labelling of two experience doctors only overlap with a dice score of 0.74-0.85 [93, 61], which should be compared to that the state-of-the-art model achieves a dice score of 0.92 on the same data set [92]. The data quality issue can also be seen by comparing results on the leading benchmark data set BRaTS to that on other, private sets. Models on the former

tend to achieve around 0.9 dice score, whereas models on private - less processed and cleaned - data only performs around 0.6-0.7 [112, 88, 89, 90]. This could of course depend on a number of reasons, but since the same models are used it still suggests that BRaTS might not be representative for the global data quality. Rather, it seems to be too clean compared to practically available data sets.

The problem at hand is thus trying to train a DL model on limited and noisy labelled data. Although the research community has developed a plethora of models, the field is reaching a point where only minor improvements are made - as can bee see by the fact that most models described above actually perform very similar despite that their level of complexity can range by many orders of magnitude. This has lead researchers to focus on managing the data deficiency, rather than simply tweaking model architectures. As discussed, augmentation has been shown to help some [95, 90, 112, 89] but it is limited. This section is dedicated to describing more sophisticated, and drastic, proposed solutions to how small data training can be achieved.

2.2.1 Curriculum & Active Learning decrease the need for data by using it more efficiently

Curriculum Learning (CL) and Active Learning (AL) are two relatively new areas of research that both try to make the training process more efficient. The two fields attempt this in slightly different, but related manners. The former does what the name implies, namely attempt to be pedagogical during the training process [95, 108, 110]. The idea is to start training on the simple training samples and then gradually increase the level of complexity, similar to how humans learn new topics. The intuition is that this should facilitate convergence and it is indeed shown to improve performance [95, 108, 110]. The latter works in a fundamentally different way. Whereas Curriculum Learning is primarily concerned with ordering the existing training set in a clever way, Active Learning is a technique to efficiently increase the total data set [103]. The two methods are thus complementary.

The question remaining for Curriculum Learning is how do we rank samples by order of increasing complexity? Different authors propose different methods. An exploratory study by [95] suggest to start the training on low resolution images and then gradually increase the level of details. The study by [110] focus on trying to mitigate class imbalance issues by starting training with an artificially sampled equal distribution and then incrementally move towards the skewed unbalanced truth. Another proposal by [108] is to divide training in three stages. In the first, only the original data samples are used. In the middle stage they introduce extensive augmentation and in the final stage they incorporate a Focal Loss, to encourage the model to focus on the most difficult cases. Focal Loss is defined in equation 2.3 [107, 108].

$$\text{Focal Loss}(p, y, \gamma) = -y(1-p)^{\gamma} log(p) - (1-y)p^{\gamma} log(1-p) \tag{2.3}$$

where $p \in [0, 1]$ is the prediction and $y = 0, 1$ is the ground truth. The Focal Loss is an extension of the standard cross entropy loss in which an additional parameter γ is introduced. When $\gamma = 0$, the Focal Loss is equal to the cross entropy but when $\gamma > 0$ the loss provides more weights to bad predictions thus mathematically forcing the model to focus more on the hard samples. By tuning the γ parameter one can trade off how much focus should be integrated. All three proposed Curriculum Learning methods present promising results.

The motivation behind Active Learning is that collecting data is expensive, so we should only every collect samples that are maximally efficient in training. By being clever in how to collect data we can thus limit the amount of redundant samples, which would decrease the total annotation cost. Again, the question is how would one do this in practice? The authors of [103] propose the following workflow for this. First train a seed model on a limited initial training set. Next, use this model to predict on a set of N unlabelled samples. For each such sample, calculate an uncertainty score and a similarity score. The former measures how confident the model is in its prediction on this unseen sample. The latter is a measure of how similar the unseen sample is to any of the already labelled samples - this can be quantified efficiently by comparing the encoded, down sampled versions of samples. The fourth step is to rank all the unlabelled samples based on their two scores. Next, pick only the top L ranked samples in the sense that they have high uncertainty and low similarity scores, meaning that the sample is drawn from a very different distribution than all training samples. Now label only these L highly informative samples and retrain the model on the initial training data joint with these new samples. This process is then repeated until satisfaction or until no additional informative unlabelled samples exist. The second step, in which N unlabelled samples is selected, can be made more efficiently if the entire unlabelled set is clustered first meaning that only the most distant N cases are selected.

In the initial study by [103] they ran 10 such iterations and managed to get on par performance by only using 13% of the entire data set, which proves that there is indeed significant redundancy in training data. Besides being data efficient they also claim that the Active Learning framework mitigate class imbalance issues, since only the most under-represented samples are used for training. However, a downside with Active Learning is that the workflow is relatively slow. Each iteration - excluding the training time - took 45 minutes. This is of course significantly faster than annotating close to 8 times more data, but it magnifies the fact that it is still an immature research field that requires more scrutiny. Nevertheless, the initial results are very promising which motivates it to be left for future studies.

2.2.2 Weak or Mixed Supervision are promising ways to decrease the dependency on expensive annotations

As already discussed the main bottle neck to performance is the fact that 3D pixel wise annotation is excessively expensive. One approach to circumvent this is the described Active Learning, which attempts to limit the amount of samples that need to be annotated. Another take on the issue is to keep the number of annotations large,

but instead attempt to decrease the time spent on annotating each sample. This is the topic of Weakly Supervised models. The field started to gain popularity in 2015 when [115] introduced the concept of Class Activation Map (CAM) based models. The underlying idea is that when a model is trained on one task, it accidentally learns relevant features for related tasks as well. More concretely they managed to show that by training a model to classify images, the model learns to identify regions in the input that are relevant for the classification. In short, they extracted heat maps (CAMs) on the input that described which pixels correlated must strongly with the predicted class. These heat spots were then cleverly extracted as segmentation masks for said class. This idea was later improved by [116] in 2019 when they introduced the generalised GradCAM model. Both papers showed that by training on only samples that were labelled with a class, they managed to get impressive segmentation performance - although not quite as good as for fully supervised training. The term weakly supervised arise because only partial annotation is required, which saves a lot of time and resources.

In the context of brain tumour segmentation the concept of weak supervision means that it would suffice to only label a patient as healthy, or with tumours - a much simpler and less subjective task. The work by [93] is the first and only to investigate this to date. To further leverage the fact that we do indeed have some fully annotated samples already, they propose the term mixed supervision. Their model is trained jointly on both fully and weakly annotated samples, however with a much larger proportion of weak samples. They managed to show that by training on only 5 fully annotated and 223 weakly annotated samples they got 89% of the performance when the model was trained on all 228 fully annotated samples. This represents a 98% reduction in annotation, at just 11% performance loss. However promising, the work has yet to be completely explored. The study suggests that the method has diminishing returns as the proportion of weak samples increases, but they argue that in order to draw any conclusions one would have to investigate the case when weak samples are in the millions, not hundreds as in their initial study. Furthermore, they have only explored 2D models so this would be another natural extension.

2.2.3 Traditional Transfer Learning is not currently feasible, but novel methods are promising

Several major fundamental differences exist between Deep Learning papers concerning medical imaging and other applications on what is commonly referred to as normal images. The latter signifies common, everyday objects such as for example animals and landscapes. The perhaps biggest such difference is that models in the normal imaging domain almost always use Transfer Learning (TL) techniques, whereas in the case of brain tumour segmentation virtually no authors mention this at all [101]. Out of all the surveyed papers in this study only two models employ Transfer Learning [117, 110], but they are constrained to 2D models. This is surprising since Transfer Learning has been proven time and time again to significantly boost performance, especially in applications where domain specific imaging is limited [106] - such as for brain tumours. However, there is a natural explanation for

this.

Firstly, there is no large publicly available data set for general medical imaging that could be leveraged for pre-training, compared to the natural imaging domain that has both ImageNet [46] and COCO [118]. Stanford announced in 2017 that they are indeed planning to release what they refer to as Medical ImageNet [119] to solve this, but it has yet to be made publicly available. Furthermore, as discussed in a previous section, most state of the art models for brain tumour segmentation use 3D inputs which makes it difficult to pre-train on 2D data [106]. It would this be necessary to find a massive 3D public data set. Two possibilities are Sports1M [120] and Kinetics-700 [121], but they are far from ideal since they consist of video data. Another possible option could be the Medical Segmentation Decathlon data set [122] but with its 2633 3D samples, spread across ten body parts, it is many orders of magnitude too small. Since no surveyed paper mentions either of these two, or any other proxy, the conclusion stands that there exists no applicable proxy data that can be used for Transfer Learning. Hopefully Medical ImageNet will solve this in the future.

The conventional transfer learning methods are thus not feasible at this time. However, researchers have proposed novel approaches to circumvent the data fallacy issue. One paper attempts to artificially increase the data set by using GAN's, with promising results [105]. Perhaps more interestingly, a paper by [106] propose another method called Model Genesis that attempts to do transfer learning without access to labelled data. The underlying idea is that there exists a massive amount of 3D medical imaging available, but without annotations. They use a 3D U-NET model and show by pre-training the model as an auto-encoder, i.e fully self-supervised, they manage to boost the segmentation performance slightly. The workflow is to take an unlabelled 3D image, augment it and then feed it through the U-NET but rather than trying to decode the tumour region they make the model recreate the original, non-augmented image. After running this self-supervised pre-training over unlabelled data they simply extract the U-NET and fine-tune it for the segmentation task. More importantly, they show that this pre-training can even be done on data from another domain. This research is still immature and need more scrutiny, but it presents very intriguing possibilities since it would decrease the need for annotated samples significantly.

2.3 Definition of Federated Learning

In the preceding sections all discussions have been on the conventional centralised setting. The remainder of the chapter will now turn to the core of the study, namely the decentralised federated setting. The term Federated Learning (FL) was first introduced in 2016 by [71] as a tool to perform Machine Learning (ML) in a decentralised, privacy preserving manner [71, 35, 36]. As briefly discussed in the introductory chapter, the general idea is to share model weights rather than raw data across parties in an attempts to protect privacy [71, 39, 72, 73, 35, 19], see Figure 1.3 for an illustration. Locally at each hospital the DL model is still trained

in the conventional sense, meaning that all knowledge and experience from this domain can be transferred to the FL setting as well. The revolutionary difference that FL provides is that all local models are then pooled and aggregated, thus creating one joint model that has been exposed to a dramatically enlarged data set [71, 39, 72, 73, 35, 19] to capture the large scale performance boosts that DL has been proven to have [41, 31, 42, 43, 44, 35, 30]. This joint model is then redistributed to all parties to perform what is referred to as another *round of training* [123, 84] and the process repeats until convergence [71, 39, 72, 73, 35, 19]. For clarity, the following nomenclature will be used throughout the study: *server* referrers to node that does not perform any training but is responsible for coordination and model aggregation. The computational nodes are interchangeably referred to as *clients*, *devices* or occasionally *hospitals*. All nodes are jointly referred to as *parties*.

A rigorous mathematical definition of Federated Learning (FL) can be provided by [35, 73]. They states that it is a ML framework in which N clients denoted by $F_1, ...F_N$ collaborate to train a final, joint model $M_{central}$. Concretely this means that the data set D is distributed over all clients, where each client holds a subset denoted by D_n such that $D = \bigcup_{n=1}^{N} D_n$ [35, 73]. A joint initial model is distributed to each client who then trains the model on its local data set. The result of each local training, often in the form of a gradient update or a locally improved model, is then pooled and aggregated at a central server to form the next generation model. Said model is then distributed back to the clients and another identical training round is repeated. This process is repeated until the global model converge to form $M_{federated}$. Convergence is determined by evaluating the model on each client's local validation set and aggregating the validation results [71, 35, 73]. The model is said to converge when the global performance $V_{federated}$ seizes to improve. Moreover, one can evaluate the performance loss of the FL approach compared to the traditional centralised methodology by comparing the performance difference of the model trained in both settings. The FL algorithm is said to have a δ-accuracy loss if the difference in performance between the central and federated model satisfies equation 2.4 [35]. Note that this is a purely academic measure, since most real world applications does not allow for a centralised model to be constructed [19, 59, 50, 60].

$$| V_{central} - V_{federated}| < \delta \tag{2.4}$$

One can further define sub types of FL by considering what kind of data is available and how it is distributed [35, 36]. They define three sub types: horizontal, vertical and transfer learning. The different sub types will be explained in terms of the application of this study for clarity. This means that each clients is a hospital, a sample is a patient and the feature space is the data stored about each patient. The horizontal setting is the perhaps most intuitive and refers to when all hospitals store the same data but on different patients [35, 36]. The vertical setting means that hospitals have the same patient, but they have stored different data on said patient [35, 36, 39]. The transfer learning sub type is the most extreme case in which the union of both patients and features is limited across different hospitals [35, 36]. The nature of the application at hand in this paper - brain tumour segmentation

in MRI scans - naturally falls into the horizontal FL framework [61, 62] which is why the focus henceforth will be on this setting. This relies on the assumption that all hospitals store brain MRI scans in the same way and that they have unique patients. This is the same assumption previously used in the related studies [61, 62], but nevertheless its validity and the potential problems with these assumptions is discussed in section 3.2.

It should be noted that Federated Learning is by definition a collaborative mechanism, which means that all parties can be better off by collaborating and taking part in the joint training [35, 71]. An interesting aspect is that parties that contribute more data will benefit more, since the model has been exposed to more samples originating from said hospital's unique distribution and might thus be biased to perform better in this specific domain. Further, the model aggregation is usually implemented as a weighted average, discussed in section 2.6, which means that local updates from large contributors influence the model more [71]. Consequently, there is an incentive for parties to increase there contribution share to capture this bias. This arms race results in a positive spiral that is mutually beneficial for all parties involved because of the economy of scale associated with data size in DL [35].

Lastly, it should be mentioned that there are several similarities between FL and another research area called Distributed Machine Learning. However, that field is mostly directed towards efficient parallelism of computation and does not consider privacy preserving mechanisms nor does it handle the heterogeneous division of data across clients. Contrary to the FL domain the server has complete control of all clients and data. These differences and their implications are discussed in section 2.4, but the severity of them arguably provide enough reason to endorse the evolution of this new field [35, 71].

2.4 The Federated Learning framework has several fallacies that has to be solved

2.4.1 Federated Learning fails to converge if the data is non-IID and unbalanced

The undisputed largest obstacle to overcome before FL can become truly invaluable in clinical settings is that it has historically been shown to be non-robust to data heterogeneity [74, 34, 124, 37, 54, 20, 125, 45, 73, 59, 23, 50, 60, 52, 33, 123, 36, 72, 71]. More precisely it has been shown repeatedly that class imbalance and non-IID distribution across clients severely impede model performance [124, 37, 20, 45, 50, 123, 36, 72, 71]. This means that if there are statistical differences between the hospitals in regards to what types of tumours occur in patients, and to what extent, then the Federated algorithm will suffer. As will be discussed in section 2.5 this happens to be the case due to biological traits of brain tumours [9, 10, 12, 13, 14, 11], demographic differences [9, 10, 13, 14, 15, 11, 7, 8] and the fact that certain hospitals care for specific sub-populations [36, 33, 45, 8, 9, 10, 11, 15].

It is a known fact from conventional ML that data sets with severe class imbalance impede model performance [45, 52]. This property naturally transfers to the Federated Learning setting as well, since local models will fail to properly converge. However, this issue is further enhanced in FL because in this setting the model does not only have to be robust to the imbalance of one data set, but across several with potentially very different characteristics [124, 37, 20, 45, 50, 123, 36, 72, 71]. Additionally to class distribution differences across clients the decentralised model has to be robust to the fact that local data sets can be order of magnitude different in size [45], due to that hospitals in urban areas care for a lot more patients than those located in rural districts. This implies that the federated model might be overly biased to the larger hospitals, thus counteracting the aim of producing a more socioeconomically equal healthcare. This issue is amplified by the fact that the feature dimension might be higher than the number of samples at the smaller hospitals, which makes local model training difficult due to overfitting [45, 72]. An implication of this is that the local model contributions from small hospitals might be less accurate in a generalisable sense. Lastly, as will be discussed in section 2.5, data quality likely varies significantly across institutions due to the different level of experience of the local doctors [34]. A FL model should thus be robust to the potential presence of low quality data at some sites.

It can be shown that if the data distributed across clients in the FL training is IID, then the model is guaranteed to converge in the same way as it would if the data had been collected at one central server [20, 71]. This is not surprising, since if the data is IID then there is no difference at all, in a statistical learning sense, between the centralised and decentralised training. However, such a theoretical guarantee does not exist for FL if the IID assumption fails [20, 71]. An intuitive understanding of this proof can be seen from analysing the most commonly used model aggregation method in FL: Federated Averaging (FedAvg) [51, 126, 83, 127, 124, 37, 54, 123, 36, 73, 125, 85, 84, 23, 52]. Although a more detailed discussion of FedAvg and its extensions is deferred to section 2.6, it will be defined in this section to clarify the convergence analysis.

The objective of FedAvg is to find the joint model weights w that minimise the global loss function. This loss is a weighted average, by sample size, of the local losses incurred at each client as defined in equation 2.5. Subscript k refers to a client, n_k and n respectively refer to the number of samples at client k and in total [71, 125, 60, 123].

$$\text{Global Loss}(w) = \sum_k \frac{n_k}{n} \text{Local Loss}_k(w) \tag{2.5}$$

However, the model updates using gradient descent techniques - such as Adam, SGD, AdaGrad etc. - are performed on each local loss, since data can not be shared. This means that FedAvg can only be evaluated on the global loss, but the actual convergence has to rely on local data [71, 125, 60, 123]. This is where the non-IID assumption starts to cause difficulties. If the data is IID across the clients this implies that $E[\text{global\_loss}(w)] = E[\text{local\_loss}_k(w)]$, meaning that the

local gradient descent steps are unbiased. This implies that the local updates are guaranteed (in expectation) to jointly converge to the global optima. However, if the data is non-IID this equality no longer holds and the resulting global model might diverge even if local models converge to local optima.

Although the first paper on FL by [71] did not report any such significant issues, it has since then been empirically shown by several studies that the non-IID setting indeed is a severe inhibitor [124, 37, 20, 45, 50, 123, 36, 72, 71, 35, 128]. A recent study show that a model trained on non-IID data has 51% worse performance than when it was trained on IID data [50, 124]. Another study even reports a 70% performance drop [37]. The same study claim that if more than 40% of the aggregate data is biased then the FL model becomes completely unusable. They further emphasise that limited work has been done to describe this deficiency and it requires more research before the issue can be properly understood in a quantitative meaning [37]. Besides model depreciation it has been shown that non-IID data causes the model to converge slower and that it requires more data to do so in an adequate manner [123]. As already discussed, the data availability in the medical domain is limited which means that additional effort must be applied to counteract this. Additionally, as discussed by [41], this lack of data further amplifies the convergence difficulty due to the class imbalance.

Although this section might suggest that FL is dead in the water for the application at hand, it turns out that one does not to be that pessimistic. The non-IID difficulty is an inherent difficulty in any decentralised setting, but thanks to recent advances in FL it has emerged as the the state-of-the-art in this regard [38, 37, 52, 40, 74, 50, 1, 60]. More sophisticated algorithms than FedAvg have recently been proposed and shown to be more robust to non-IID data, to be discussed in section 2.6, so there is still hope. Furthermore, as discussed by [36], the presence of highly non-IID data might not be all bad. Yes, it makes convergence harder, but at the same time it allows the model to be exposed to a more complex learning scenario which forces it to generalise [36]. In essence, if the difficulty can be overcome than the end result will be a less biased model that is better off for everyone.

2.4.2 Practical difficulties with Federated Learning systems

Besides the problematic convergence in Federate Learning, there are a number of practical issues that must be solved in order to make implementation feasible. These are discussed in this section for completeness. It will be revealed that this concern is less severe in the hospital collaboration case compared to the more commonly studied mobile phone setting. Nevertheless, it is relevant to provide a brief discussion in order to understand what elements have been the driving factors in the attempts to improve FL historically. Further, it provides motivation to why certain common FL features can be overlooked in this application.

As mentioned, the historical motivation behind proposing FL was for it to be used for mobile phones [71]. This meant that the number of clients was expected to be in the millions, compared to the much smaller setting of hospitals [36]. A commonly quoted

issue with decentralisation in general and FL in particular is that the managing server does not have complete control, due to that clients are autonomous [20, 36, 36]. This means that the server can not assume that all clients are always available for the next round of training, nor that they would be able to complete the round that they started. This means that significant effort has been spent historically in the FL community to cope with stragglers and devices that drop out [20, 84, 85, 123, 36]. Commonly this is solved by simply hard coding that if the client takes too long to respond, it is simply dropped from the round [123]. Naturally, this synchronisation problem is proportional to the number of clients. Luckily, the hospital application has a relatively small amount, meaning that this factor is less important.

Another factor that impacts the costs associated with synchronisation is the extent to which the clients are reliable. In the case of hospitals this is hardly an issue, since they have a stable and fixed power supply and reliable network connection. For mobile phones, on the other hand, the story is very different [84, 85, 51]. Firstly, mobile phones are highly heterogeneous with regards to computation power, meaning that some versions might take longer to train on - or might even not able to at all [54, 50, 36, 21, 20, 50, 60, 36]. Furthermore, they can only be allowed to be used in a training round if they have sufficient battery levels or are plugged in, have a reliable wi-fi connection, have the newest software update etc. [84, 85, 51]. Additionally, the common FL setting requires that the system is robust to unreliable and heterogeneous infrastructure [84, 85, 51]. Although arguably none of these factors are significant threats to the hospital case, they must still be considered in a final version.

The last traditional concern is regarding the high communication overhead and latency associated with performing Federated training [129, 124, 37, 54, 20, 125, 45, 84, 85, 59, 50, 123, 36]. This has been seen as one of the largest deficits and a large effort to mitigate it has been seen from the academia. Again, the motivation is largely that the mobile phone application assumes mostly wireless and limited network connection, in which it is inadequate to send the large model updates over and over [125, 123, 71]. The common solution has been to either propose different model compression techniques or to limit the number of training rounds by making each iteration more efficient [125, 20]. This network bottleneck is less severe if the system contains less than 100 clients [84], which is reasonable to assume in this case. Furthermore hospitals can be assumed to have a stable and fixed network connections. To conclude, this practical issue is not significant either. Nevertheless, there are arguments to why communication should be minimised anyway. Firstly, it is of course wasteful to not do so if it can be done without other major deficits - after all the hospital's bandwidth is still finite. More importantly, increased networking increases the risks associated with Man-In-The-Middle attacks [86] to be discussed in section 2.8.

2.5 MRI scans across hospitals are heterogeneous

As already hinted in section 2.4.1 the data considered for a decentralised learning setting on brain tumour patients exhibits strong non-IID and class imbalance, which has been shown to severely degrade FL performance. In this section evidence to support the former statement is provided. To fully understand why this is one must first explain the biological traits of the disease.

Brain tumours comes in many different sub types, each with its unique characteristics and survival rates that differ by orders of magnitude [9, 10, 12, 13, 14]. The fact that the patient's chance of survival is tightly bound to what sub type is present means that the CAD model must be adequately performant at not only identifying lesions, but also correctly classify them in order to allow proper treatment protocols to be initiated. Importantly, this means that multi-class training must be performed. On a brighter note, thanks to advances in MRI scanning techniques it is possible to visually distinguish the different sub types by comparing contrast and patterns in different views and sequences [9, 10, 14, 17, 18], as discussed in section 1.1.1. Although this is a none-trivial task, it at least provides a theoretical chance that a computer vision model can do the same.

These sub types can be categorised as either primary - the most common types being Meningiomas, Pituitary Adenomas, Neuromas and Gliomas - or secondary tumours. The latter are tumours that have spread from other parts of the body and they are referred to as Metastasises. These different sub types have significantly different occurrence rates in general. For example, Metastasises occur fours times more often than all primary tumours combined. Furthermore, the relative occurrence rate of different primary tumours vary by as much as a order of magnitude [9, 10, 11, 13, 14]. This illuminates the magnitude of class imbalance that the CAD model must be able to overcome. To make matters even more complex, Metastases patients tend to have more than one lesion - it is not uncommon with a handful - meaning that the accumulated number of this sub type is further enhanced. Lastly, they are much smaller than primary tumours which signifies that the model must be able to detect lesions as small as parts of a ml [9, 10, 14, 17, 18].

So far the discussion has only shown that an intrinsic class imbalance exist, which is just as problematic in conventional ML as it is in FL. However, by also considering demographic factors it will be made clear that the distribution also differs across hospitals, which is where the real FL problems arise. The reason why the IID assumption required for conventional distributed training is invalidated in this domain is because the risk of getting a certain type of brain tumour depends on age [7, 8, 9, 10, 15], race [8, 9, 10, 14, 15], sex [7, 8, 9, 10, 11, 14, 15], life style [7, 8, 9, 10], genetics [8, 9, 10], standard of living [7, 9] and the geographic location [8, 9, 10, 11, 15] of the patient. For example, Meningiomas are 50% more common in females than in males [10] and brain tumours in general are 4 times more common in high HDI areas than in poorer regions [7]. As discussed by [45] it is even possible that some hospitals do not have a single case of certain sub types. All this means that the distribution of tumour types differs significantly across different hospitals

due to that they attend to different sub-populations, in different geographical areas with its unique demographic footprint and socioeconomic standard [36, 33, 45]. Besides this class imbalance there is a significant difference in the total number of samples at each hospital, due to their difference is size, which means that the FL model must cope with a severely unbalanced data set as well [36].

The final point to discuss, brought forth by [34], concerns that the data quality differs between hospitals. They also considers the matter of collaborative training amongst hospitals for tumour prediction and they argue that "there may exist non-negligible gaps in the quality of data among different hospitals since a rich-experienced chief physician with advanced medical devices in a high-rate hospital will be more likely to produce accurate data than an junior physician with low-end of devices in an ordinary hospital" [34]. This view is supported by [19, 35] who claim that data annotation in medical applications is non-trivial and partially subjective. Another study by [33] agree and claim that patient journals are heterogeneous from different doctors. Furthermore, to make matters even worse, there is a plethora of different MRI scanners all of whom give rise to images with different characteristics [130]. All of this means that not only does data quality differ across hospitals, but it will likely also do so within the same hospital due to that different doctors or brands of MRI scanners are involved. Naturally, and as supported by [36], training a model with noisy, or low quality data is a major inhibitor which further complicates convergence.

It should be noted that this quality heterogeneity is despite extensive global efforts to standardise the medical imaging domain. For example, most hospitals store data according to the DICOM standard and regulation enforces that the images are of high quality [31]. To put it plainly: the situation could have been much worse. On another note completely, related to the practical implementation of FL, this standardisation actually simplifies a real world FL collaboration since data is physically stored in the same way. Although a number of reasons cause the model fitting itself to be complicated, it suggests that it is at least practically feasible to implement such a collaboration between institutions.

2.6 FL can be improved by considering more sophisticated model aggregation algorithms

In general the matter of convergence is a complex one that depends on a number of parameters, especially in an unbalanced and non-IID setting. For examples it has been shown that some DL models are inherently better suited for non-IID data than others [37] and that even if adequate models and algorithms are used, the success is still sensitive to proper hyper parameter tuning [83]. Furthermore, since DL is a data driven approach [19, 35, 36, 20, 21, 37, 38, 33, 39, 34, 40] the data scientist might even find itself with the perfect model, perfect algorithm and perfect hyper parameters but still end up with useless performance due to the simple fact that data is too scarce. In that case the common approach is to pre-train on proxy data, which has been proven time and time again to improve the situation [84, 59, 21]. However, if no appropriate proxy data is available one must rely on other techniques.

Furthermore, as discussed in section 2.4.1, the matter of convergence is even more complicated in FL than in conventional DL. This is especially true for applications such as the one under consideration in this study, since brain tumor data is non-IID and unbalanced - see section 2.5 for a detailed discussion on this - a scenario which has been show to severely degrade FL performance. As described, the cause for this deficit is mainly inadequate model aggregation techniques that fail to converge if the data is non-IID. However, as discussed FL is a very promising technique well equipped to deal with several of the privacy constraints that have impeded DL historically. Combining these statements it is thus not surprising that significant research has been conducted to improve on the FL model aggregation in order to mitigate the convergence fallacy and to unlock the invaluable potential that FL possess in sensitive applications. This section is dedicated to presenting these extensions and improvements.

2.6.1 Federated Averaging is the common benchmark but several options have been proposed

Ever since Federated Learning was proposed in 2016 [71] the research community has come up with a plethora of algorithms to improve its performance. The initial version, termed FedSGD [71, 52, 129, 123, 73], has since then been largely replaced by more sophisticated aggregation techniques. Today the most commonly quoted algorithm is FedAvg [125, 85, 84, 23, 52, 51, 126, 83, 127, 124, 37, 54, 123, 36, 73], but others such as FedProx [127, 57, 60, 73, 72], DGC [129, 37, 58, 36], FEDPER [50], LoAdaBoost FedAvg [50], CMFL [50], eSGD [50], SVRG [73], FAug [124], FedAsynch [51], AFL [74], SecProbe [34], a one-shot model [59] and a momentum based FedAvg [73, 72] are all proposed by various authors. Consequently there is no shortage of algorithms to choose from in an attempt to tackle the characteristics of any application at hand.

Interestingly enough though, despite the discussed convergence fallacies most applications are still only using the bare bone Federated Averaging algorithm introduced in section 2.4.1, equation 2.5 [125, 85, 84, 23, 52, 51, 126, 83, 127, 124, 37, 54, 123, 36, 73]. Although it is arguably the most naive algorithm - just a simple weighted average - and that several studies have shown that it is non-robust to non-IID data [127, 57, 60, 73, 72], it is still widely used and preached. However, rather than using this fact to justify a claim of FedAvg's superiority over the other propositions, this is a sign that illustrates just how young the field of FL is. The studies that are using FedAvg are mostly interested in examining other unexplored aspects of FL than its convergence or absolute performance, such as minimising communication overhead [129, 37, 36, 124] or allowing for stricter privacy constraints [51, 59]. FedAvg is thus being used a convenient benchmark [127, 57, 60, 73, 72, 129, 124, 37], but as this section will soon show several of the other propositions are better suited to handle a variety of tasks.

One can largely categories the FedAvg extensions as either attempts to address the network bottleneck, system design improvements or managing unbalanced, non-IID convergence. The reason why these factors must be analysed was discussed in section

2.4.2 and section 2.4.1.

2.6.2 Asynchronous or bandwidth efficient algorithms solve several problems but at the cost of accuracy

As described in section 2.4.2 a large historical concern has been that FL incurs too much communication overhead. This has lead researchers to develop increasingly lighter communication algorithms, starting with that FedSGD was replaced by FedAvg for this very reason [71, 52]. Since then CMFL [50], eSGD [50] and FAug [124] have all been proven to improve this further. However, they all do so at the expense of model performance [124, 50]. The general idea is to use different compression techniques in order to decrease the size of model updates before sending them [20, 50]. This has proven very successful due to that model updates are very sparse - up to 99.9% redundant [50] - thus making them highly compressible without trading off too much accuracy [50, 129]. Consequently, the current state-of-the-art algorithm for network constrained applications is Deep Gradient Compression (DGC) [129, 36, 37, 58].

Although DGC successfully decreases communication significantly, is too does so at a performance cost relative to FedAvg [129, 37] in both IID and non-IID settings [37]. This inadequacy is increased with model complexity, thus making DGC a bad choice for complex learning scenarios - unless the latency and network constraints are the limiting factor [37]. Lastly, some studies claim that DGC is useful as a privacy protection mechanism because the compression works as an obfuscator [58], so it might be a tool worth considering in sensitive applications for this reason. However, as discussed in section 2.9, there are other privacy mechanisms that does not trade off performance. This suggests that unless network is a limiting factor FedAvg is still a better choice than DGC.

Another more experimental algorithm is the one-shot model proposed by [59]. It takes network constraints to the extreme by only allowing a single round of training, arguing that one can consider each local model as part of an ensemble. Although the performance naturally decreases, it is still only 10% less than the centralised baseline model. This quite extraordinary feat is accomplished by carefully analysing each local model and weeding out those that have poor validation performance or inadequate training data, thus creating an ensemble of the best local models [59]. However, this technique requires more research to verify its usefulness before it can be considered in this study. Nonetheless, it is an interesting and instructive take on the problem. Especially, the idea that inadequate models should be excluded is a good take away that might be useful for improving other algorithms as well.

Another scenario that has received attention by several researchers [131, 20, 36, 38] is the possibility of fully decentralised Federate Learning, in the sense that no server has to be trusted to manage the system. This is discussed more in section 2.7. One of many problems with implementing this scenario is that the FedAvg algorithm assumes a synchronous, server controlled system [51]. To this end the first asynchronous model aggregation algorithm was proposed by [51]: FedAsync. They

show that it is indeed possible to perform FL in a decentralised fashion, which has the upside that it allows for more flexibility and increased scalability [51]. Although FedAsynch is shown to be robust to hyperparameter decisions, 10 times slimmer in network communication and to converge faster than FedAvg it does so at the expense of performance. However, limited analysis of the algorithm has been performed and it has yet to be used for a deep network on complex data. To date, it has only been used with a shallow CNN on CIFAR-10 [51]. Consequently, FedAsynch is not deemed appropriate for this study. However, due to the many upsides of decentralisation this proof of concept is intriguing for future extensions.

2.6.3 Convergence in non-IID settings can be improved if momentum and regularisation is employed

Besides optimising the FL framework to better handle latency and system architecture, several studies have considered the even more pressing matter of convergence. Although FedAvg is more robust to non-IID and unbalanced data than its predecessor FedSGD [73], it still has severe deficits as discussed in section 2.4.1. There are authors that still claim that FedAvg is indeed robust to this [125], but the vast majority of research suggests otherwise. In order to capture the full capability of FL researchers have thus proposed extensions to FedAvg that aims to solve this.

One suggestion on how to mitigate the non-IID convergence issue is to limit the extent to which one attempts to generalise across local distributions. This is the approach used in the FEDPER algorithm described by [50]. The idea is that by only sharing the base layers of the model during training, it is possible to learn low level features with the help of all local distributions but to personalise the high level information extraction on a local level. The base layers are trained jointly, but the high level layers are only trained locally thus allowing for each local model to specialise to its unique distribution. Note that this is similar to the argument behind conventional pre-training, where the base layers are trained on proxy data. The intuition is that low level features are less non-IID, which suggests that FEDPER should suffer less in this regard. Indeed, the study showed an improved robustness and increase performance relative to FedAvg [50].

Another insight is that the convergence problem in the non-IID setting is largely caused by that the local models diverge relative to each other due to that they are exposed to widely different distributions, which causes the aggregate model to misbehave [60, 37]. One might hope to mitigate this by limiting the impact of each local update in each round, thus forcing them to stay closer to each other [60, 37]. This is the idea behind the second improvement to FedAvg, proposed by [60]: FedProx. It introduces a proximal term in the local loss function, a trick that is shown to improve convergence both theoretically as well as in practice. The study shows a 22% performance boost relative to FedAvg [50] - making it the state-of-the-art solution to date for FL on non-IID data [127, 57, 60, 73, 72, 50]. More precisely, FedProx adds a L2 regularisation term to the old local loss function. The norm that is punished is the delta relative to the previous global model, which encourages small local deviances to be sent back for aggregation. The proximal loss function is

defined in equation 2.6.

$$\text{Local Proximal Loss}(w) = \text{Local Loss}(w) + \frac{\mu}{2}||w - w\_global||^2 \tag{2.6}$$

From equation 2.6 it is plain to see that FedProx is a strict extension to FedAvg, where the latter corresponds to FedProx with $\mu = 0$. Everything else regarding the two algorithms is the same. By properly tuning the μ-parameter it is possible to balance convergence speed and robustness to non-IID divergence. Another novelty introduced by the FedProx inventors [60] is that they allowed for heterogeneous hyper parameters across clients. Previous studies had used the same number of training epochs at each client during a round, but by loosening this constraint they allowed for better utilisation of client heterogeneity. For example, some clients might have a more powerful GPU which means that they can afford to train more epochs in the same time span without lagging the joint training. Furthermore, this personalisation meant that clients with less training could train for fewer epochs to avoid overfitting. By the same argument they allowed the proximal parameter μ to be client dependent, since some local distributions will be more divergent from the average than others which calls for a higher L2 punishment [60].

The next improvement is brought forth by [73]. They took inspiration from the fact that conventional DL experienced improved performance at the introduction of momentum based optimisers, such as Adam, AdaGrad and RMSProp [132]. They recognised that the global model update, resulting from the aggregation of the local updates, is the analogue to the gradient in conventional SGD. Since adding momentum improved SGD, it is reasonable to assume that similar techniques might work for FL as well. Their contribution was thus to update the global model using a moving average of previous global models, as described in equation 2.7. Subscript k refers to a client, w(t) is the model weights at time t and η is the momentum parameter.

$$w(t+1) = w(t) - \eta \sum_k \frac{n_k}{n}(w(t) - w_k(t)) \tag{2.7}$$

Again, this is a strict extension of FedAvg which corresponds to $\eta = 1$. By allowing tuning of the momentum parameter η, they managed to get significantly improved performance relative to bare bone FedAvg [73, 72]. The intuition behind the improvement is that the global model delta at each step is the result of local training on a random subset of clients, similar to the random samples that constitute a batch in conventional DL, and the momentum term thus limits the impact such a random subset can have on the final model. In the non-IID setting this is crucial, since random selection of clients might produce a round where the local data sets are not representative for the global average and thus it would be fatal not to limit this update's impact [73].

It should be noted that this momentum approach does not assume the use of any specific local objective. Importantly, this means that the FedProx algorithm can

be used in conjuncture with this method to produce an even more sophisticated algorithm. To date no study has used these two methods together, but since they are both strict extensions of FedAvg this approach must be at least as good. However, due to the increased flexibility by tuning μ and η it is reasonable to assume that it will actually improve the convergence performance.

Lastly, a natural extension of FedAvg/FedProx with momentum would be to introduce the FedAvg/FedProx Adam equivalent. Similarly to how Adam was shown to outperform momentum in the conventional, centralised learning scenario [132] it is reasonable to assume that it would be beneficial in the federated setting as well. This relies on the aforementioned interpretation that each local model update in FL is the analogue of a batch gradient update in the conventional setting. However, such an aggregation technique has not been cited in literature to date and is left for future work in this study as well.

2.6.4 Global model performance can be increased by selectively only including well behaving local updates

One of the arguments in favour of momentum is that it limits global degradation caused by clients with low quality data. This aspect of data heterogeneity is also the topic under scrutiny that lead to the proposal of SecProbe by [34] and LoAdaBoost FedAvg by [50]. As discussed in section 2.5 different hospitals will likely have data sets of greatly varied utility and quality, which implies that the local model updates from these locations will likely be noisier and less performant. The general idea that the authors behind SecProbe and LoAdaBoost FedAvg discuss is that since this data heterogeneity is known to exists, it is naive to treat all local model updates in the same way during global aggregation as is the case for all previously discussed algorithms. Instead they propose a schema where the global model discriminates against poorly performing local models in order to avoid it being tainted by their lower quality. Both studies show that this indeed improves performance relative to the baseline FedAvg [34, 50].

The difference between the two models is how they quantitatively determine which local clients are under performing. LoAdaBoost FedAvg does this by examining the loss progress. Practically, they compare the local validation loss with the median validation loss during the last round and choose to discard all models that exceed this level. This intuitively means that the global model only incorporates local contributions that are better than the average from last round, which suggests that the next generation should improve [50].

SecProbe is implemented in a slightly different way and relies on a *utility score* for the local model. The exact definition of how this utility score is calculated depends on the task at hand, but the general idea is that the local model update is validated at the server on a blind validation set and better performance equals a higher utility. Finally, they choose K of the M local updates in each round based on their utility and aggregate these to form the next generation global model. However rather than choosing the K best performing they randomly draw K clients without replacement,

each with a probability proportional to its utility. Similar to how Differential Privacy mechanisms works, as discussed in section 2.10, this stochastic element ensures that the privacy of the client is preserved since it is not possible to conclusively determine which model updates where used. Importantly this means that it is not possible to determine which of the clients has lower quality data, which might be a privacy sensitive feature [34].

Although both of these methods are still young, largely untested and requires more thorough research before they can become standard practice, the authors present impressive preliminary performance boosts. For example, it is shown that even in a setting where 70% of clients hold only 40% high quality data and 60% random garbage, the final model performance is barely degraded compared to the ideal baseline case if SecProbe is used [34]. This suggests that the SecProbe algorithm is robust to unreliable participants. This result is promising two-fold. Firstly, it means that the FL ecosystem can be made robust to lower quality participants. Secondly, it also means that the system is protected against malicious participants that wish to actively degrade the global model by sending faulty updates, since these will be discarded based on their low utility score [34]. This type of threat, called model or data poisoning is discussed in section 2.8.3.

An end note on the problematic existence of noisy or low quality data on some clients is that this issue is mitigated if the number of clients grow, since that limits the extent to which one individual client can impact global performance [85, 126]. This is of course only true if the newcomers have high quality data. A corollary of this is that the FL convergence becomes more robust if a larger proportion of clients is selected at each round. This is the analogue of why conventional SGD appreciates with a growing batch size, since the updates become less noisy [85, 73, 83]. A rule of thumb is provided by [73] who claim that one should never use less than 10% of the clients at each round. Besides improving performance this also speeds up convergence [73].

2.6.5 The non-IID issue might be an artefact of naive assumptions - Agnostic FL attempts to solve this

A final note on convergence in a non-IID setting concerns a radically different idea brought forth by [74] and [45]. They argue that the only valid approach to FL is actually to consider everything concerning distributions a black box on which nothing can be assumed. They argue that since in reality we can not know anything for sure about the distributions, all and any assumptions are false. They thus urge researchers to view the learning scenario as completely unknown. Although [45] only states this without providing any solutions, the paper by [74] actually studies this abstract idea in full and proposes a practical implementation to make the vision a reality. This section is dedicated to describing said work.

The general idea behind the paper by [74] is that not only can we not assume that data is IID across clients, but we can not even assume that the joint aggregated distribution is representative for the real world distribution. This implies that the

global objective function obtained by the simple weighted average of local objectives in itself is biased. In such a scenario, test performance on external data will be lacking even if the FL algorithm manages to converge. To put it plainly, they claim that even if all clients pooled their data to one server and trained in the conventional centralised manner the resulting model would be biased and under performing on third party data - especially if the aggregate data size if small [74]. This is equivalent to stating that a subset of hospitals can not present an unbiased representation of all global patients unless the collaboration constitute a very large proportion of all institutions - which is a reasonable statement. Furthermore, since the data across clients is non-IID there is no guarantee that the weighted average distribution is representative for the individual client's data. This means that the jointly trained model is optimised for an abstract global distribution for which it is not intended to do inference, which might cause it to under perform in production at the actual clients [74].

To summarise, the article by [74] address the fact that previous work has been too naive in its assumptions on what distributions are real, meaning that they have all optimised for a faulty loss function. In order to mitigate this flaw, the authors propose a model called *agnostic FL* (AFL) [74]. The idea is that rather than training a model to optimise on the naive simple weighted distribution one should create the model that is ideal for *any* mixture of the local distributions. They define this mathematically using a mixture vector $\lambda \in \mathbb{R}^N$, where N is the number of clients in each training round. Further, they impose that $\sum_k \lambda_k = 1$ and $\forall k \lambda_k >= 0$. They thus propose that the model should be optimised for the distribution D in equation 2.8, where subscript k refers to a client [74].

$$D = \sum_k \lambda_k D_k \tag{2.8}$$

Note that this is a strict extension of FedAvg which naively assumes that $\lambda_k = \frac{n_k}{n}$. The authors of AFL argue that there should exist at least one such λ which produces a global distribution D that constitutes a learning scenario which more closely resembles real world conditions. The objective of Agnostic FL is now to find the best model regardless of what λ this is, meaning that the training is completely agnostic of what the true distribution really is - hence the name. Mathematically this objective function is finally defined in equation 2.9 [74].

$$\min_w \max_\lambda \sum_k \lambda_k \text{Local Loss}_k(w) \tag{2.9}$$

Agnostic FL thus tries to find the best model for the worst possible scenario, which [74] argue is the most reasonable setting since nothing can be assumed about the true distribution. This argument is supported by the fact that they empirically find that AFL outperforms FedAvg. Additionally they show that besides that AFL increases performance it is also more fair in the sense that it tends to be less discriminating to individual clients. The latter is a consequence of that AFL does not assume

anything on any distribution, thus preventing the final model to be biased towards any particular individuals [74].

The Agnostic framework is arguably a more idealistic setting that more closely mirrors real world conditions. It thus seems reasonable that it might be better than FedAvg also on brain tumor CAD. However, it is a very young field that still poses many practical obstacles that have yet to be researched. Firstly, the minimax optimisation problem is significantly more complex than the conventional minimisation objective [74]. For this reason [74] report that the algorithm is still very slow, especially if the number of clients is large because that causes the λ space to blow up. One way to mitigate this might be to incorporate some prior on the data, thus limiting the search space. However, this has never been done and it is not clear what that prior would look like, nor is it established how this would impact performance [74]. Implementing a large scale AFL model is thus still infeasible in terms of computation, at least until more efficient algorithms are proposed.

Secondly, the paper by [74] only theoretically prove that convergence is guaranteed to improve relative to FedAvg in the convex case. Yes, they show an empirical improvement in a non-convex case as well, but it is still limited since they only considered a shallow model on Fashion MNIST. To date there is still no theoretical or empirical proof that AFL works on more complex, multi-modal data sets that require more sophisticated models. No work has been done on DL in AFL for example, which leads to the conclusion that this technique is still to undeveloped to be considered for this brain tumour study at this time. However, the paper by [74] has provided ample arguments to support that it is indeed a groundbreaking, promising methodology which is why it would be interesting to study further at a later date. It is thus left as future work.

2.7 System design and privacy assumptions

After a comprehensive analysis and discussion of convergence in Federated Learning it is time to consider more practical matters concerning implementation. It has to be decided how hardware should be configured, how communication should be controlled and what assumptions can be made concerning participants in the collaboration. This is the topic under scrutiny in this section.

2.7.1 The system can be either fully decentralised or managed by a trusted server

The first design choice concern how the computation should be performed and coordinated. By design Federated Learning is a decentralised approach, but one still has to decide how this should be managed. The system can either be completely decentralised, in the sense that no single party coordinates training meaning that all participants are equal, or partially centralised [51, 38, 58, 20, 36]. The vast majority of existing implementations follow the latter by allowing a central server to coordinate all operations [50, 36, 37, 54, 84, 85, 123, 51]. The peripheral clients in this

design perform all training, but only the server perform model aggregation. This design imposes the restriction that clients can only communicate directly with the server and not at all with each other. The network topology is a star. Plainly this scenario means that the server works as a master node, responsible for initiating all training rounds, and clients are simple worker nodes responsible for training on their local data to provide model updates to the server. All data is located at the clients, with the exception that some proxy validation data might exist at the server [50, 36, 37, 54, 84, 85, 123, 51]. Although this is an intuitive and simple solution - which explain why it the most widely used - it also requires some problematic assumptions.

The star topology makes the server a critical and powerful player in the system, which makes it a much greater threat if it is compromised and starts acting maliciously. Essentially, by concentrating all this control to one node the system becomes more vulnerable to more powerful attacks [131, 38]. This and other threat scenarios are discussed at length in section 2.8. The server is such a scenario is referred to as a *trusted aggregator*, because all partied must trust that it does not misbehave [131, 38, 85]. For this reason some researchers have started to question if it is really possible to assume that such a trusted, powerful but yet secure party can be found in many applications [131, 38]. In the example of brain tumour diagnostics this is equivalent to asking the question whether all hospitals can agree on who should be allowed to be much more powerful and virtually obtain perfect control of the system. Not only must this chosen hospital be trusted by all others to not break this confidence, but everyone must also agree that it is properly protected against third party attackers. Some authors argue that hospitals do not have perfect trust, meaning that this is impossible in practice [38]. One solution might be to allow a third party, such as government, to have this role instead but that might not be feasible either. Consequently, in a scenario where collaborators lack the blind trust necessary to unanimously assign one party as the *trusted aggregator* the only other solution is to look towards a fully decentralised solution instead [131, 38, 20, 36, 50, 45].

Although the vast majority of studies have assumed the existence of a *trusted aggregator* [131, 50, 36, 37, 54, 84, 85, 123, 51], some work has been extended to consider the more extreme scenario where no trust at all can be assumed between parties [131, 38, 45]. As discussed, this requires a completely decentralised design in which all parties are equally powerful and all perform both training and aggregation [20, 36]. Although this would provide significant privacy upgrades [131, 38], it is still an unsolved and very complex problem [36]. However, some proposals exists.

The paper by [131] attempt to solve this by using Blockchain and cryptography technologies. They propose a full mesh topology where all clients communicate with each other, but thanks to encryption and Blockchain they do so completely invisibly thus preserving perfect privacy. This makes for something of a utopia for FL but unfortunately the implementation is still inadequate. The authors have only considered a simple learning scenario because due to the severe communication overhead the technique is still infeasibly slow. This rules out the possibility of using it for complex deep models on multi modal data at this time. However, they show that

on their simple scenario they manage to get performance that is on par with server controlled FL [131]. This makes it a very interesting topic for future work. Another similar proposition is the one by [38]. They suggest a model called *TorMentor* which works by using the Tor and Onion protocol. Similar to Blockchain, these are also techniques for secret, invisible IP communication. This ensures that the sender's identify is unknown by the receiver. This is also very premature work that has yet to be shown to work for more complex scenarios, but the authors explicitly state that they believe that *TorMentor* is ideal for FL in the medical domain due to its sensitive nature [38]. This is thus also left as interesting future research.

So far most arguments have been in favour of fully decentralised systems, due to their superiority in terms of privacy. However, there are reasons for choosing a server central design as well, beyond the fact that it to date is the only feasible option. As discussed in section 2.4.2 a practical issue in FL is that clients are autonomous and unreliable. A server controlled design has at least one reliable point - the server - whereas a fully decentralised system is completely left in the dark [20]. This lack of control is further amplified if the decentralisation schema involves obfuscating all participant's ID, as in the proposals by [38] and [131]. As a result, it is significantly easier to design fault-tolerate algorithms in the server case, since it can be given the authority to define all operations [20]. To date it is not clear how a server-less design would even handle heterogeneous clients, drop-out, stragglers and a number of other control mechanisms that are required in a real world scenario where perfect availability can not be assumed [20]. Additionally, a decentralised design further incurs significant communication overhead which makes it infeasible for network limited applications [20]. To conclude this section it is thus reasonable to claim that the only practically viable option to date is to use a server controlled design. However exciting the idea of Tor or Blockchain may be, it is simply not practical yet.

2.7.2 Training is commonly performed synchronously, but some works propose an asynchronous solution

The second design decision when building a FL framework is whether the model aggregation should be synchronous or asynchronous. Concretely this means whether model aggregation must wait until all participants have responded, or if it can be done on the fly as updates come in continuously [20]. The traditionally most employed choice is the former [50, 20, 37, 54, 126, 84, 85, 123, 51] although a few studies have considered the latter as well [51, 38, 131]. There are a number of reasons behind why synchronous is so vastly more common. The first two are purely a statistical consequence. Firstly it is because the server controlled FL setting naturally employs this schema, which as discussed above is the most common setting. Secondly, the by far most common FL algorithm FedAvg is synchronous.

However, there are intrinsic arguments for a synchronous design as well. Firstly, it is technically less demanding since it is the arguably most intuitive way to perform aggregation. This increased implementation complexity has made asynchronous solutions virtually non-existent in cases with less than 50 nodes, sine the gain of

asynchronous solutions increase with the number of participants [20, 51]. Additional arguments against asynchronous solutions is that it is not clear how they would handle stragglers or drop-outs, whereas a synchronous design can define that all clients that have yet to responds after a set amount of time is simple skipped [20, 71, 36, 51]. To conclude, the most suitable solution seems to be to employ a synchronous design. The only reason to attempt to solve all the additional difficulties related to asynchronous designs is to capture its superiority if a large number of clients is used. According to [51] the synchronous solutions seize to be efficient if the number of participants exceed a few hundred. Thus, for the brain tumour and hospital case no compelling arguments are found to support anything else than the synchronous decision.

2.7.3 The perceived threat level determines the setting

So far the discussions have mostly been on how to practically achieve good performance in Federated Learning, but another just as important aspect is doing so without exposing patient's privacy. The succeeding sections are dedicated to discus how these privacy concerns can be managed. However, before moving to this discussion it is important to establish common nomenclature and to explain how previous work has considered the matter. This is presented in this section.

Protecting privacy means that sensitive information must not be leaked to unauthorised or untrusted parties. However, it is not always trivial to determine what parties can be assumed to be trustworthy. As discussed in section 2.7.1 some studies have assumed that the server can be trusted whereas other have not. It is a subjective, application and incentive dependent question that requires careful consideration since if the wrong assumption is made that might leave the system vulnerable to attacks.

Besides that one has to establish if a *trusted aggregator* exists, there are a number of additional questions to be answered. Firstly, since the FL training involves IP networks one must establish if the communication channel can be assumed to be secured or not [40]. If the application is such that the latter is true then one has to employ protection against this, such as encryption or secure hardware solutions. See section 2.8 and 2.9 for a detailed discussion on this. However, if the application happens to be in a scenario where the channels are perfectly safe the this extra step can be disregarded.

The second decision that must be made is how much the different parties can be trusted. In the most naive setting where everyone is assumed to be perfectly trustworthy, then no privacy concern would exist except from third parties. This assumption thus only requires the designer to make sure that the system is secure against third party breaches. However, in most scenarios such a perfect trust is unlikely to be reality [131, 38]. Instead, the most common assumption is the honest-but-curious setting [39, 36, 50, 83, 1, 34]. In this scenario it is assumed that all collaborating parties will obey to the common rules that are laid out, thus not actively cheating. However, at the same time they are curious in the sense that they will try to maximise the amount of information they can extract from every situation [36].

Importantly, this implies that if the protocol accidentally leaks any unintended information then the parties will not hesitate to use this even if it happens to be sensitive and private data. Under this assumption the system must not only be protected from third parties, but it must also make sure that the common protocol is bullet proof from a privacy preservation standpoint.

However, there are those that claim that even this loosened honest-but-curious assumption on trust is too naive [50, 45]. These authors instead claim that the only valid assumption is to assume that no trust at all exist between any parties. In this extreme case, the risk of actively malicious clients is considered to exist [36]. This forces the system to have an additional level of security that ensure that no sensitive information what so ever leaves the client, neither via the common communication protocol nor arbitrarily aggressive attacks. This extremely restrictive assumption is rare in practice, but some researchers still use it. For example, a study by [45] assumes that the data is so sensitive that it is not allowed to either share it, nor train a joint model on it. In such a scenario FL is not possible, instead they were limited to only perform hyper parameter sharing across clients using a model called Restrictive Federated Model Selection (RFMS).

The final decision to make is concerning if the collaborating parties can be assumed to be non-colluding or not, with the former being the most common [34]. If they can, the the protection mechanisms can be slightly relaxed. However, if the threat of colluding parties is considerable then one has to make sure that privacy can not be breached even if an arbitrary number of clients share all their knowledge about the party that they wish to attack. This of course significantly more difficult.

From this discussion it is clear to see that the number of precautions that have to be designed grows exponentially as the assumptions become less and less naive. Conversely this means that if the designer is too restrictive in these assumptions - considering threats that are not plausible - this causes significant unnecessary overhead in the development. This motivates why it is so important to start the design by carefully considering what threat scenarios are relevant to the current application.

2.8 Federated Learning has to be protected from a number of privacy threats

Whenever dealing with medical data, which is highly privacy sensitive [21, 54, 57, 33, 56, 32, 58, 34, 55], it is crucial to consider what potential threats exist that might compromise patient's integrity. This section is devoted to providing a detailed description of what these risks are. It turns out that the bare bone FL design is susceptible to a number of fallacies with regards to privacy [45]. The next section, 2.9, naturally follows by discussing what techniques can be employed to protect the FL system from these threats.

The first issue concerns that multiple studies have shown that traditional anonymisa-

tion of data is not sufficient, as attackers are still able to extract private information. This is discussed further in section 2.8.1. The second threat debunks one of the common arguments in favour of Federated Learning, namely that it is safe because it only involves sharing model updates which are ephemeral [85, 71, 133]. Although FL is more secure than conventional ML because data is not shared [85, 71, 133], it turns out that a plethora of studies have shown that it is indeed possible to extract significant information on the underlying data by cleverly analysing the model updates. This is discussed further in section 2.8.2. Lastly, it is described that the FL system is severely vulnerable if any party - internal or external - decides to break the honest-but-curios assumption. This is discussed in section 2.8.3.

Another threat is related to the fact that FL by design is decentralised. This means that the system is vulnerable to adversarial third party agents that intercept the communication between clients and servers [52, 34, 38, 40] in so called man-in-the-middle (MITM) attacks [86], see Figure 2.1. As discussed in section 2.9 this suggests the use of different encryption techniques, see Figure 2.4. A final and related privacy concern is relevant in the most extreme sensitive cases where the identity of clients and servers can not even be known to each other during communication [38]. This scenario is beyond the scope of this study but suggested solutions include using Blockchain techniques [36, 131] or the Tor and Onion protocol [38].

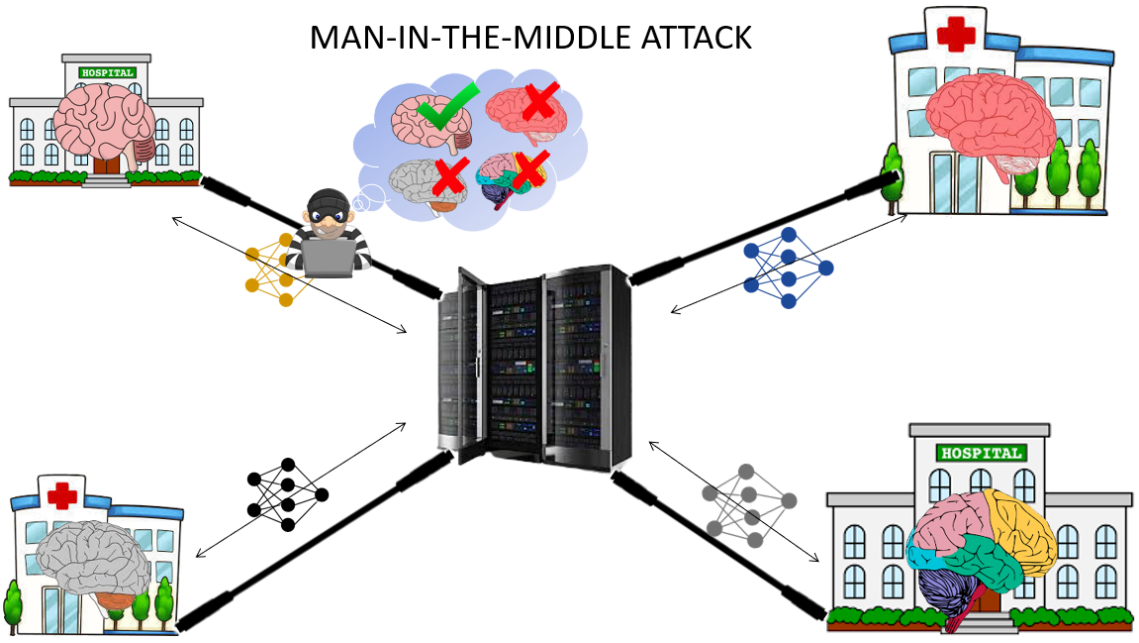


Figure 2.1: Federated Learning and all other decentralised systems is naturally vulnerable to Man-In-The-Middle attacks in which an attacker intercepts data between participants.

2.8.1 The death of anonymisation

The naive and traditionally most employed method to ensure the privacy of individuals is to manually anonymise all data before providing it to the data scientist [134].

However, there are significant drawbacks to such an approach [71, 47]. Firstly, it is time consuming and prone to human error. Secondly, it is difficult to know before hand what features must be anonymised. The latter might seem trivial, but as proven by [135, 136, 137, 23, 55, 85] it is far from easy due to the large amount of secondary public data that is available. In [135] they prove that it is possible to identify individuals in the famous Netflix data set, even though it is anonymised, with the help of publicly available data that is correlated to the Netflix data set. Simply put, research has found that conventional anonymisation is insufficient in today's big data world due to the presence of correlated external data sets [135, 136, 137, 23, 55, 85], see Figure 2.2 for an illustration. It is thus necessary to rely on other techniques. This is discussed at length in section 2.9.

Figure 2.2: Several studies have shown that it is possible to infer private information from anonymised data by analysing its correlation with publicly available secondary data. Due to the abundance of big data in the modern society anonymisation in itself thus fails to provide sufficient security guarantees.

2.8.2 Federated Learning does not share raw data but information is still leaked in the model updates

A commonly used argument for Federated Learning is that it is more secure because only model updates are shared [54, 57, 33, 58, 32, 52, 40, 85]. However, several studies have shown that these model updates actually also contain significant amounts of compromising information on individuals on the training data [131, 40, 52, 50, 54, 38, 34, 1, 58, 56, 23, 83, 138, 35]. These studies prove that the previous belief that privacy is protected if the raw data is not shared, is simple not true [58]. The initial study that revealed this fallacy showed that it is possible to reconstruct a recognisable face from the input data by just having access to the

final model in a facial recognition system [56]. Their attack worked by rendering the image that maximised the a priori probability of the model output. Furthermore, they showed that the technique was also able to reconstruct the original faces from blurry, or only partial, images [56]. Since the initial proof by [56], more sophisticated model attacks have been developed that pose even stronger threats by revealing more detailed information on individuals [1]. An instructive illustration of just how good the reconstructed images can be is shown in Figure 2.3. These particular results are from the study by [1] and they very clearly show how much information is actually leaked from the model.

Figure 2.3: Several studies have successfully been able to show that a malicious adversary can reconstruct the original images in a data set by simply analysing the model, thus proving that models themselves leak sensitive information. This implies that Federated Learning requires additional levels of protection to be fully secure. The illustrated results are from [1]. A passive attack means that the adversary is playing by the rules, in the sense that it does not meddle with the training process at all. The adversary only uses information that the others are willingly giving up by sharing their model updates. The reconstructed images in this scenario represents the intrinsic information leakage in bare bone Federated Learning. An active attack on the other hand is even more severe, but this requires that the adversary is actively trying to break the system. This scenario discussed in section 2.8.3.

Another study showed that it is possible to completely recover, pixel by pixel, the original image from observing the shared gradient even in the case of very complex training images [58]. Other studies have shown similar results using GAN's [34, 2] or model inversion attacks [38]. Even more disturbingly, another study showed that it is possible to infer information from the original images that goes beyond the learning objective of the model. They exemplified it by showing that with a model originally trained to identify the gender and ethnicity, based on a face, their attack could accurately reveal if the person was wearing glasses or not - even after that

several models had been aggregated [83].

The described model attacks simulate the threat posed if either the server [1] or any of the clients [58] decide to pursue a malicious goal. Further, several attacks work without the adversary having any prior information at all on the data, model or application [58, 138] meaning that the system is susceptible to third party attacks as well. This makes it important to make a distinction between what is referred to as a *black box* or a *white box* attack. The former means that the adversary only has access to the output of the model whereas the latter is more severe because it means that the attackers has full access to everything about the model - architecture, parameters, intermediate outputs etc. [138, 56]. It has been shown that several of the attacks work in both scenarios [138, 56], but that the risk is amplified in the *white box* scenario which is problematic since the conventional FL setting provides *white box* access to even third party interceptions [138].

A comprehensive study by [138] examined the accuracy of their *membership attack* - meaning that they wanted to infer if an image was or was not part of the training data - in both settings. Their analysis used CIFAR-100 and the current state-of-the-art model, DenseNet, in an FL setting using FedAvg. They showed that a passively malicious server could perform the attack with 79% accuracy, whereas a client did it with 72% accuracy. As discussed more in depth in section 2.8.3, the threat is amplified if the adversaries are actively malignant. In that scenario the accuracies increases to 87% and 77% respectively [138]. Lastly, they showed that the attack even worked if the model was pre-trained on another data set and only fine tuned on the target data [138]. This attack quantitatively describes how naive the traditional view is on that sharing model updates is safe. What is even worse is that several of the attacks were shown to work completely invisible to the other parties, meaning that it is impossible to realise when the threat is a reality [1, 138].

A detailed discussion about how to protect against these kinds of attacks is provided in section 2.9, however it is relevant to briefly discussed it here as well. Firstly, model leakage is naturally proportional to the number of parameters in the model meaning that complex models are more susceptible [138]. Further, an overfitted model leaks more information, meaning that regularisation and running fewer epochs provide more protection [138]. Further, leakage is naturally more severe the more homogeneous the data set is since the reversion is statistically less difficult to perform [83]. The results by [138] also illuminates the risk associated with centralising too much power to the server, which is why some studies have started to look into Blockchain solutions [36, 38, 131]. Another common technique that has been proven to mitigate the leakage is Differential Privacy [23, 34], discussed in section 2.10.

2.8.3 Actively malicious agents enhance the threat level

In the previous discussion it was described that the model updates themselves contain compromising information that any party can leverage, without breaching the honest-but-curious scenario. This setting, referred as a passive, means that parties follow all protocols during training, but they do not hesitate to do any model leak-

age attacks on the final model to extract more information [36]. However, if this honesty assumption fails to hold several studies have shown that even more severe privacy deficits are exposed [34, 131, 72, 38, 84, 126, 138, 83, 1, 50, 36]. This fact is illustrated by Figure 2.3 which shows that an active adversary can reconstruct images that are even closer to the original, unseen data. This setting, referred to as no-trust, is especially dangerous if the server is compromised because it has access to the most amount of information [36]. Similarly, the threat posed by clients that possess a larger proportion of the training data is also enhanced [138].

Studies have suggested two types of active so called *backdoor attacks* that both reveal substantial private information on individuals: *data poisoning attacks* [83, 84] and *model poisoning attacks* [34, 131, 72, 38, 126]. In the former the attacker purposely feed the model with deceiving data during training, thus biasing it towards some malicious goal to extract information [83, 84]. The latter is more severe and means that the attacker purposely sends deceiving model updates, thus allowing for more tailor-made outcomes. One type of *model poisoning attack* was proposed by [1] to simulate a scenario where the server is compromised. They showed that by purposely isolating a client, in the sense that it and it alone was part of the training rounds, they could create a highly biased model that leaked even more information on the target client than discussed in section 2.8.2.

Two other studies by [72, 126] provide even more powerful *model poisoning attacks*. They show that it is possible to virtually replace the entire global model with their own model by analysing the aggregation scheme and reversing it. This statement is worth pondering for a second, since it means that even though potentially millions of clients are part of the FL training the entire outcome can be decided by a single actor [83] [72, 126] further show that this allows for an attacker to replace the originally intended model with one that is designed to purposely misclassify selected images. To make matters worse this attack was 100% successful even though only one client was adversarial and it was completely invisible to all other parties in the study by [126]. In the brain tumour CAD application this has devastating consequences, since an attacker for example can make the model purposely classify all tumour patients as healthy.

These attacks make it clear to see why the no-trust setting presents a major threat. The fact that several of them [138, 126, 1] are invisible to the other parties and that model aggregation provides some anonymity to the adversarial client [72] makes matter even worse. Even though the attacks are discussed to be more difficult on highly non-IID data [126], there are still significant arguments to add layers of protection against these active adversaries. One such proposition is using forms of encryption [1]. This and more tools are discussed in the next section.

2.9 Ensuring privacy

2.9.1 How much privacy is enough privacy?

Federated Learning provides much more privacy than conventional Machine Learning due to that the raw data is not shared [54, 57, 33, 58, 32, 52, 40, 85] and that privacy risks are ephemeral [85, 71, 133], making it the State-of-the-Art privacy mechanism [131]. However, as seen in the previous section there are still significant threats that must be considered before FL can be fully used in a clinical setting. With that said, the extent to which different sources are concerned about these additional security threats differs significantly. It ranges from those that does not consider any extra privacy mechanisms at all [60, 85, 71, 61] to those that claim that FL must be considered a system that by design has to be protected against arbitrarily malicious behaviour [72]. The need for privacy is of course application dependant, but it is clear that the young research field requires more work to find common ground. The application considered in this study however is one of the most privacy sensitive [21, 54, 57, 33, 56, 32, 58, 34, 55], which motivates a discussion on how to prevent patient integrity breaches.

The intrinsic contradiction between privacy and ML is that models are ultimately designed to extract information about the data [55, 83]. This means that if privacy restrictions are looser, meaning that the model has more information to leverage, performance tend to increased [38, 127, 72, 57, 54, 40]. Hospitals must thus carefully consider the inherent trade-off of performance and risk. This makes for a difficult task because there is currently no theoretical, quantitative description of this trade off [54]. The pragmatic approach then becomes to carefully make sure that no excess privacy restrictions are imposed [36]. Unfortunately, it has proven complicated to determine what is *enough* protection. This can be exemplified by that historically it was thought that data anonymisation is sufficient, which several studies has debunked [135, 136, 137, 135, 23, 55, 85]. Next, professionals thought that sharing models or gradients was secure but that has since then, as previously discussed, also been overturned [131, 40, 52, 54, 50, 38, 34, 1, 58, 56, 83, 23, 138]. This raises the concern that today's privacy mechanisms might be disproven in the future.

An additional abstraction of the concept of privacy relates to what assumptions on the system can be made, based on the threat scenario. As discussed in a previous section different authors assume either no trust at all between parties [45, 38, 131] or that all parties are honest-but-curious [39, 36, 50, 83, 1, 34, 34, 85]. The former setting requires what is defined as *local privacy*, in the sense that no sensitive information at all can leave the site [36]. The latter, more common setting, is defined as *global privacy* and only requires that data is protected against third party adversaries [36]. Furthermore, one could imagine a scenario where privacy concern is heterogeneous across parties, such as if they are geographically located in regions with different privacy legislation [36]. This scenario has not received a lot of attention yet and has proven difficult to implement [36]. Which setting is the most realistic for the hospital case is debated and boils down to if different institutions trust each other [32]. The majority of studies have considered that global privacy

is sufficient, but there are those that claim that it is unrealistic that hospitals could share that kind of trust [38]. Again, it becomes an application dependant decision, partially decided by current relevant legislation.

2.9.2 Common techniques to protect integrity in FL

Even if the level of required privacy is still under debate, the fact that some additional layer of protection to the baseline FL framework is required is commonly agreed upon [127, 57, 54, 39, 40]. Several different approaches to achieve this has been proposed. The most commonly used method is a technique called Differential Privacy (DP), advocated by [45, 84, 34, 52, 138, 127, 60, 1, 57, 54, 50, 83, 40, 131, 36]. This technique is used to make individual patients invisible in the sense that the model does not reveal if the patient was part of training or not, a concept termed *membership inference* [83]. DP is discussed at length in section 2.10. However, it does not provide any protection against malicious parties, which brings us to the the second most common technique: Secure Multiparty Computation (SMC), advocated by [84, 52, 34, 38, 72, 58, 60, 57, 50]. This is a tool that is designed to deal with a scenario where parties do not blindly trust each other by obfuscating their contributions to the global model. SMC will be discussed in section 2.11. In order to protect the system against both threats, most studies argue that both techniques must be employed to provide complementary protection [127, 57, 54, 39, 40]. This constitutes the current state-of-the-art method [40].

There are multiple methods to perform SMC. The two most commonly occurring techniques use either Homomorphic Encryption (HE), advocated by [40, 34, 58, 57, 54, 50, 36, 39], or Shamir's secret sharing, advocated by [131, 36, 54, 1, 57]. These are discussed in section 2.11.1 and 2.11.2 respectively. They have their own pros and cons and different authors claim that one is better than the other, but the jury is still out. An argument against HE brought up by the advocates of Shamir's secret sharing [57, 1] is that it does not protect the system against malicious clients, since they must all share the same encryption key. They thus argue that it only protects against a malicious server. On the other hand HE is promoted by [58, 40, 39] who claim that it is more secure since the encryption protects against third party adversaries. This makes the combination of DP and HE the state-of-the-art, according to [40], since the threat posed by third party attackers is more severe than from other clients. In the case of the hospital case, this seems to be a reasonable assumption. Lastly, in an application with very high privacy concerns it is possible to leverage all three methods to get even stronger protection [36].

Although DP and SMC collectively ensures relatively good protection against malicious behaviour from other parties [36], there are still downsides and fallacies. A few critics of DP claim that it can only provide sufficient protection if it trades off significant performance [72], to be discussed in section 2.10. A less disputed claim is that SMC provides protection against data related attacks, but not model poisoning [72]. Further, fact is that both HE and Shamir's secret sharing incur large communication overhead [34, 57, 40, 84], making them infeasible in several settings where network is a bottleneck [38]. Another study claims that it is infeasible if the

number of clients exceed 100 [84]. These arguments have lead to the proposal of a plethora of other privacy preserving techniques that are either complementary or substitutions to DP and SMC.

The most common complementary technique, and perhaps most obvious, is to encrypt all communication between parties and to enforce an authentication protocol, see Figure 2.4. Commonly [34, 38] AES [139] is used for channel encryption and SHA-256 [140] for authentication. This protects the FL system from MITM, model- and data poisoning attacks [84, 34, 38]. However it does not protect against malicious servers, see Figure 2.5, which is why it should be complemented with other techniques. Another solution is to used specialised secure hardware, but this is too expensive and impractical in most applications [38]. Recent advances are experimenting with training and inferencing on completely encrypted models [34, 39, 30], such as DeepSecure or Microsoft's CryptoNets [39, 141]. However, this incurs significant computational overhead [34, 39] and it is in a early development stage and has yet to show good performance [39]. Another intriguing method is the use of Blockchain technologies proposed by [36, 38, 131], with the upside that parties are completely invisible to each other. However, this is also a premature technology that requires more research to be feasible.

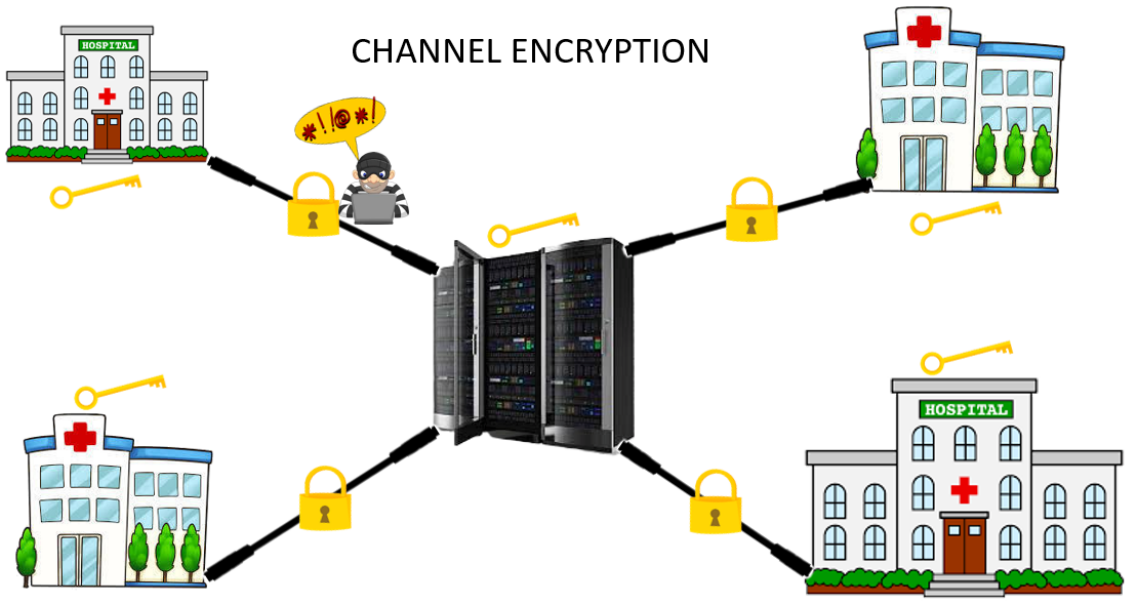


Figure 2.4: Channel encryption protects a system from third parties, such as the Man-In-The-Middle attack illustrated in Figure 2.1.

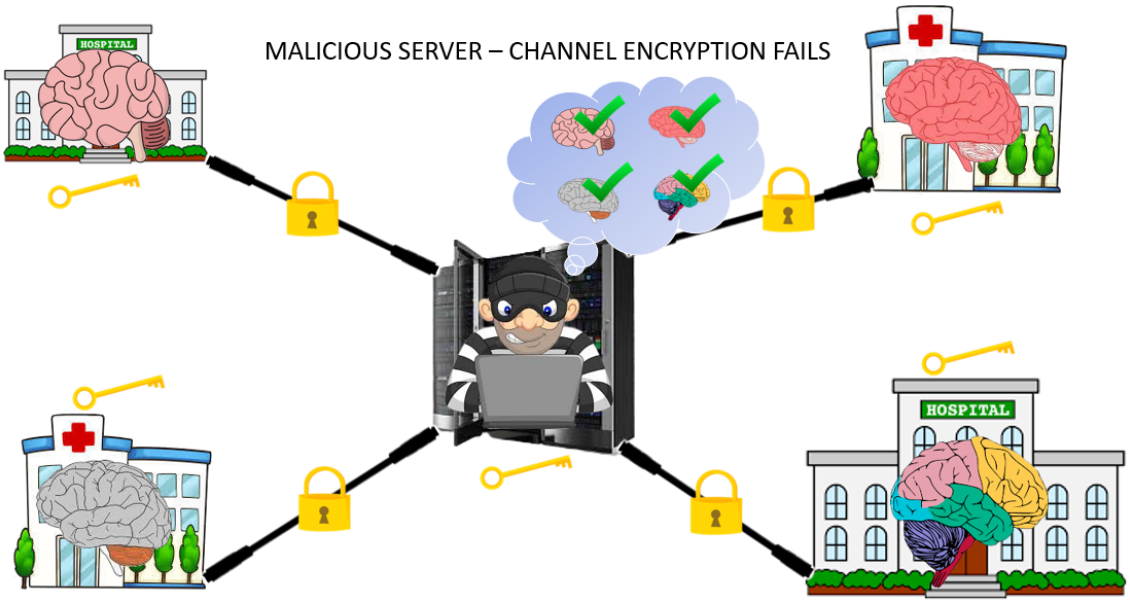


Figure 2.5: If the server is compromised, or actively acts maliciously, the threat level increases significantly since it has access to information from all clients. In such a scenario simple channel encryption, illustrated in Figure 2.4, does not suffice since the attacker has access to the encryption key. This implies that it is necessary to employ more sophisticated privacy mechanisms such as SMC or DP, see section 2.9 for an in-depth discussion.

Three final notes of caution on model aggregation. Firstly, although aggregation is necessary to protect the local data by making the individual party more anonymous [138, 72], it presents a contradictory threat as well. The fact that the client becomes more anonymous actually makes it harder to identify where malicious activity originates from [72]. The same argument highlights a deficit in a Blockchain system. Secondly, the only thing being shared in FL is the model updates [85, 71, 133]. This means that the privacy risk is directly proportional to the number of parameters in the model, meaning that a more complex model poses a greater risk [138]. This is further amplified by the fact that an overfitted model leaks more information than a generalised [138]. This motivates the use of the least complex model necessary. Consequently, in order to mitigate overfitting, the leakage can be decreased by increasing the batch size [58], input size [58] and decreasing the number of training epochs [138]. Surprisingly, dropout has not been shown to influence the leakage [83]. Furthermore, the privacy protection is naturally increased if the number of clients increase [131, 83, 138]. Lastly, another means of decreasing the effective model dimension that is sometimes used is model pruning [58] or model clipping [138] but these are less common and will not be considered further in this study.

2.10 Differential Privacy

Differential Privacy is as discussed the most common technique to provide privacy protection in Federated Learning. Intuitively, differential privacy can be though of

as the technique of adding noise as a way to obfuscate details. Figure 2.6 provides a visual intuition of the concept. In this section it is mathematically defined and it is described how it can be used in practice.

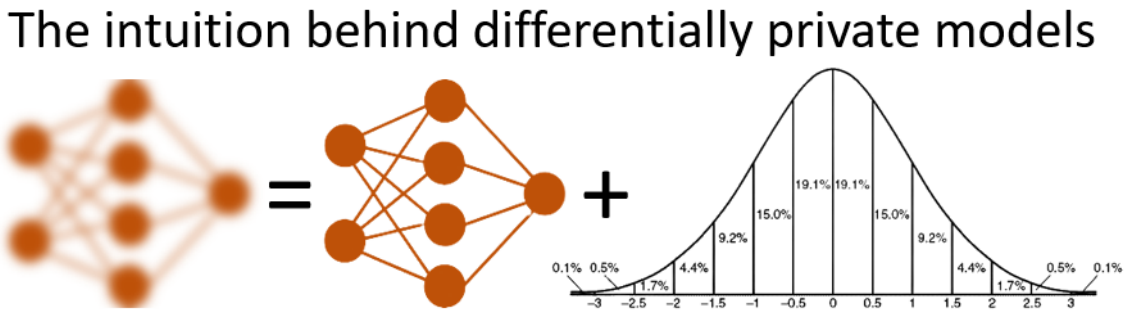


Figure 2.6: Differentially private models attempt to overcome the issue illustrated in Figure 2.3, namely that plain text models leak sensitive information. The idea is to protect sensitive details by obfuscating it in properly tuned noise.

2.10.1 Definition of Differential Privacy

There are two commonly used definitions of Differential Privacy that are used in the academia [23, 127]. The initial definition, referred to as ϵ-DP, was first proposed by [134] and has since then been quoted by several studies [23, 127, 34]. The second definition, referred to as ϵ, δ-DP, is a strict extension to the initial version and was first proposed by [142] and has since then been quoted by [40, 23, 127, 143, 144, 55]. The latter is the "the gold standard of privacy" [55] and will thus be the focus of this study.

Differential Privacy provides privacy guarantees on a stochastic algorithm A. It is related to the concept of neighbouring data sets D and D' which is defined as any two sets that differ in at most one element. The algorithm A is said to be ϵ, δ-DP if it satisfies equation 2.10 for all D and D' [142, 40, 23, 127, 143, 144, 55].

$$Pr[A(D') = S] \leq e^{\epsilon} Pr[A(D) = S] + \delta \tag{2.10}$$

In the setting of FL the algorithm A refers to the model, trained on the two neighbouring data sets D and D'. The definition thus states that the probability distribution describing the possible converged states of the model does not differ significantly. The deviation is quantified by ϵ. Plainly this means that if a model is ϵ, δ-DP then an attacker who got its hands on the final model can not determine if it was trained on D or D'. More precisely, the attacker can not guess it without having at least a $\frac{100}{1+\epsilon}$ % chance of being wrong [127]. Thus if a model has $\epsilon = 0$ then it is completely impossible for an attacker to determine which data set was used for training. Importantly this means that the individual who was not part of one of the neighbouring data sets is completely invisible to the attack, thus having perfect privacy protection [134]. Thus, by designing the FL model to have a low ϵ-DP patient integrity can be secured.

The description so far has not considered the second parameter, δ. This is the original ϵ-DP definition provided by [134]. The extension ϵ, δ-DP used in this study simply means that the probability distributions in 2.10 are allowed to deviate more than prescribed by the ϵ-parameter with a small probability δ [142]. This extension was introduced to provide additional flexibility and performance gains, at a small δ cost. Common practice is to allow $\delta < \frac{1}{N}$, where N is the number of clients used for training [23, 55]. Similar rules of thumb are provided for appropriate levels of ϵ, but this is more disputed because it is hard to intuitively gauge what certain ϵ, δ levels actually mean in terms of privacy in practice [38]. One author argue that ϵ should range between 0.1 and 3, depending on the sensitive nature of the data [145]. Medical data would thus fall in the lower end of this spectrum. Another claim that $\epsilon \sim 1$ correlates to "very strong privacy" [55]. A less conservative third study allows ϵ to be as high as 8 [143]. Even if the academia is in a slight disagreement, the order of magnitude is unison and can serve as a rule of thumb. Further, the socially acceptable level depends on the perceived risk of an attack occurring. In a small scale decentralised system where there is more control and all clients share the same objective, meaning that there is limited incentive for any insider party to attack the system, this risk is lower meaning that a higher ϵ can be allowed [145]. It seems reasonable to assume that the considered hospital setting fall into this category, assuming that the model is protected from third party attacks using encryption. In general it is valid to allow higher ϵ levels if multiple layers of protection are used such as SMC or channel encryption [127].

It should be noted that Differential Privacy does not equal absolute privacy. It has been shown that if the data set is sufficiently rich, these DP techniques are not sufficient to provide full integrity protection. However, the current general consensus is that the social gain from using the data out weights the limited, but existing, privacy risk [145]. This statement is further supported by the wide use of DP in academic studies, already discussed. Furthermore, this reason is why most authors propose that several layers of protection are used, such as SMC.

2.10.1.1 Definition of L1 Sensitivity

A closely related concept in the Differential Privacy domain is the L1 Sensitivity. This quantity will be used in the next section to quantitatively determine how much privacy an algorithm provides [40, 23, 34], which motivates the need to properly define it. Given an algorithm A the L1 Sensitivity, ΔA, is defined in equation 2.11 [134, 145, 142, 143, 34].

$$\Delta A = \max_{D,D'} |A(D) - A(D')| \tag{2.11}$$

where D and D' are all neighbouring data sets. The definition states that the sensitivity of an algorithm is the largest difference in output that can be achieved by removing or adding a single entry from a data set, thus quantifying the maximum impact an individual can have on the result.

2.10.2 Implementing a Differentially Private model

Having established that FL models should also consider Differential Privacy this raises the question of how that is analytically determined and practically implemented. The goal is thus to modify the model *M(D)* to be ϵ, δ-DP. The common approach is to simply add zero centred random noise, thus creating *M'(D) = M(D) + noise* [35, 143, 144]. The intuition behind said method is that the underlying data set *D* is masked by noise, thus making it difficult to reveal any individual data points. Empirical proof of this statement is illustrated in Figure 2.7 which shows that a DP-model leaks less information than its non-DP counterparty. The question that remains is what kind of noise, and how much, must be added to achieve the sought after ϵ-DP. Further it must be determined if the noise should be added at the client or at the server. Once those questions are answered it is possible to create a differentially private federated learning system, as illustrated in Figure 2.8.

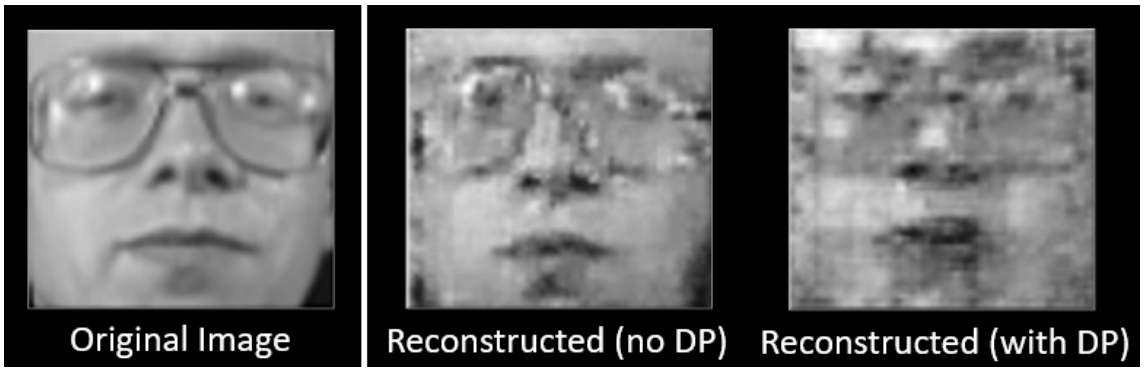


Figure 2.7: The illustrated results are from a study by [2] and they show that differentially private models leak less information. Notably in the illustrated example is that adding DP makes the attacker unaware of whether the subject is wearing glasses or not, which might be considered sensitive information. However, differential privacy does not equal absolute privacy as seen from the de facto reconstructed image even with DP. This implies that additional layers of complementary privacy mechanisms are necessary to achieve full protection.

Firstly, concerning the type of noise. The most commonly used method is to add Gaussian noise [127, 55, 40, 57, 54, 72, 50, 58], with Laplacian noise coming in on second place [40, 34, 57, 58]. One author further proposes the use of Binomial noise [57]. One argument to why Gaussian noise is preferred is that it is less fat-tailed than the Laplacian distribution, meaning that it perturbs the original model less causing it to converge faster [55]. This precedence was shown to be amplified if the number of classes was increased [55]. Another study claims that the type of noise is irrelevant and that the amount of noise is the determining factor [58]. Again, no clear consensus can be reached but this study will follow the majority and use Gaussian noise. It should be mentioned that more experimental studies are trying more advanced methods [127], but they have yet to be properly proven to be superior and are thus not considered in this study.

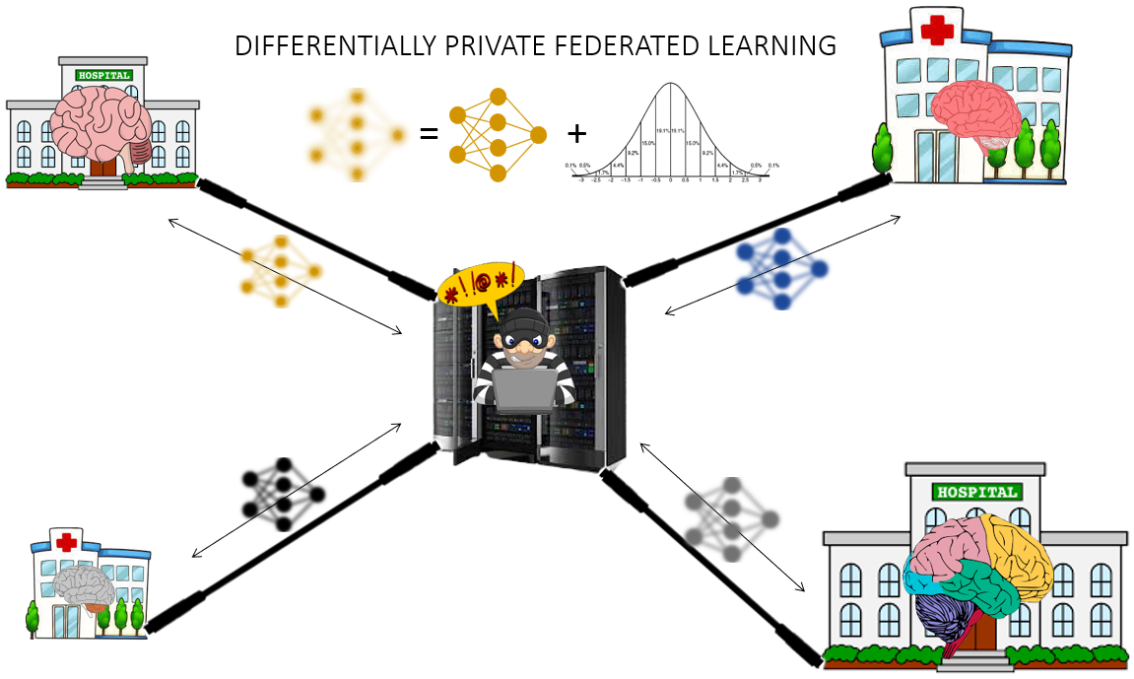


Figure 2.8: By adding carefully calibrated amounts of noise to the local model updates before sending them to the server for aggregation it is possible to create a differentially private federated learning scenario. Importantly this means that the system is protected, in a differential sense, from even malicious behaviour by the server. However, as illustrated in Figure 2.7, DP in itself does not equate full protection and should be complemented by other privacy techniques.

The method of adding Gaussian noise in order to achieve DP is commonly referred to as the Gaussian mechanism [143, 142]. Pioneering work by [134] started to quantify how much noise must be added using the model's L1 Sensitivity. It has been proven that by adding Gaussian noise with σ the model M achieves (ϵ, δ)-DP if equation 2.12 hold [143, 40, 23]. Although it is non-trivial to appreciate the practical impact of the mechanisms, some unifying underlying intuition for the Gaussian mechanism can be derived from the inequalities. Firstly, a model with higher sensitivity requires more noise in order to be DP. This is in line with the intuitive statement that if an individual has a bigger impact on the final model, then more noise must be added to obfuscate the individuals contribution. A second, simpler conclusion is that adding noise with greater variance provides more protection of integrity.

$$\frac{\sigma}{\Delta M} = \frac{2\sqrt{\log 1.25/\delta}}{\epsilon} \tag{2.12}$$

A practical concern is that the assurances provided by equation 2.12 requires that the L1 sensitivity is bounded. In the FL setting this translates to that the model updates sent from the client to the server must be bounded. The common approach to achieve this is to perform model clipping, meaning that the L2 norm is bounded by some constant C [127, 38, 40, 57, 54, 143]. It should be noted that choosing C too

low degrades performance [82], whereas a C that is too high fails to provide sufficient privacy protection. A proposed choice of C is to use the median norm of all model updates [23]. The full common algorithm for the Gaussian mechanism is thus finally to bound the model update and then add Gaussian noise [143, 127, 38, 40, 57, 54].

Introducing the Gaussian or Laplacian mechanism is commonly done locally at each client before being sent to the server, thus providing client-level protection [127, 38, 40, 57, 54]. Another, less common and more relaxed approach is to aggregate all local models at the server before using the Gaussian mechanism [23]. This second approach assumes either a trusted server, or that other layers of protection such as SMC is used. The gain however, is that less noise must be added to provide the same level of DP [23]. A final notion concerns what is referred to as the privacy amplification theorem which states that the DP level is improved if at each FL training round only a subset of clients are randomly sampled with a probability q. It can be proven that this diminishes the privacy leakage by a factor of q [144, 55, 23]. The intuition is that the random sampling of the input adds a layer of difficulty of inferring what input data was used.

However, the discussion so far and the assurances provided by equation 2.12 only relate to a single pass of the data set [144, 142]. In the iterative Deep Learning training this notion thus has to be extended to work for several epochs. A naive corollary to the privacy guarantee provide for a single pass is that the leakage grows linearly with each iteration [34, 144]. However, this fails to account for that each iteration has diminishing leakage because the data has already been seen once before. This insight has lead to that several tighter bounds have since then been proven. To date the best bound is obtained by a method referred to as the *Moments Accountant*, with which it can be shown that the total leakage after T epochs, using random client sampling with probability q, satisfies $O(\epsilon q\sqrt{T}, \delta)$-DP [55, 23, 143]

Having established how the Gaussian mechanism, it is important to consider the trade offs associated with its usage. There is a fundamental trade-off between DP and model performance because the increased noise makes convergence difficult [38, 127, 72, 57, 54, 40, 36, 39, 58]. An application specific decision must thus be made on what level of privacy is required and how much accuracy loss is acceptable [35, 143, 82]. In some studies it has been shown that in order to achieve acceptable levels of privacy, excessive noise must be added that severely degrades performance [72, 58]. On a more positive note, several studies have shown that it is indeed possible to achieve good results in both regards [23, 38, 143, 62]. Again, no consensus can be reached but the wide usage of DP suggests that the more positive view is more reasonable. With that said, it is important to keep in mind that no excess noise beyond the required should be added.

The purpose of this study is not primarily to provide a detailed, quantitative understanding of the privacy leakage in the FL training. The exact derived bounds in this section are thus not the main contribution. Rather, they will be used to provide an intuitive rule of thumb on how different design choices impact the DP performance of the model. For completeness, these will now be summarised. Firstly, privacy

decays with increased model L1 bound and decreases as the root of the number of training iterations. It is improved by adding more variance and increases linearly with random sampling of clients.

2.11 Secure model aggregation using Secure Multiparty Computation techniques

As discussed in section 2.9, researchers have proposed a number of solutions to ensure that the local model updates can be aggregated in a secure way, without needing to rely on a trusted server or third party aggregator. The motivating idea is that since the model updates themselves leak private information from the client's data, as discussed in section 2.8, it is preferable if the aggregation can be performed in a way that does not reveal any of the individual model updates. This gives rise to the extensive field of Secure Multiparty Computation (SMC), dedicated to solving aggregation tasks in a private manner. The task is to aggregate data from several sources, without any party learning anything more than the final aggregation [84, 52, 34, 38, 72, 58, 60, 57, 50]. Importantly, this means that there is no need for a *trusted aggregator* [141], see section 2.7.3.

In the field of Federated Learning there are mostly two different SMC techniques that are used for this end: Homomorphic Encryption [40, 34, 58, 57, 54, 50, 36, 39] and Shamir's Secret Sharing [131, 36, 54, 1, 57]. This section is dedicated to explaining these techniques. From the literature review is is clear that Homomorphic encryption is a slightly more common technique than Shamir's Secret Sharing, which is why it will be the main focus. The secret sharing technique is only briefly touched upon, for completeness of the discussion.

2.11.1 Homomorphic Encryption

An important distinction in the field of encryption is to differentiate between symmetric and asymmetric methods. In section 2.9 the encryption method termed AES was mentioned as a tool for channel encryption [139]. This is a symmetric method, meaning that both sides of the communication can encrypt and decrypt the message. The reason behind this is because the encryption and decryption key are the same and shared by both sides of the channel [139]. This is ideal for protection against third party attackers, since they do not have the key. However, it is insufficient from a secure aggregation perspective, since the goal is that only the client should be aware of what the model update is - not the server. This means that the server can not be given the decryption key. However, the server must still be able to perform the aggregation operation on the encrypted model update. This idea of performing operations on encrypted data is solved by the *homomorphic property* [146, 141, 39, 40].

The *homomorphic property* relates to a specific operation, for example addition or multiplication. Given an operation $\otimes$ an encryption algorithm is said to be $\otimes$-homomorphic if the following holds. Given any two original messages α and β and

their encrypted analogues α^* and β^*, we have that $\alpha \otimes \beta = \text{Decrypt}(\alpha^* \otimes \beta^*)$. In short it means that the operation can be performed either on the encrypted or the decrypted messages, the end results after final decryption is always the same [146, 141, 39, 40]. This property means that the client can encrypt the message and send it securely to the server. The server can then take all these encrypted models, aggregate them while still encrypted, and send back the final model to each client. The client can now decrypt the new generation model, knowing that the entire process was secured both from snooping third parties and the server. The entire process is depicted in Figure 2.9. Especially note that this means that AES channel encryption is no longer needed.

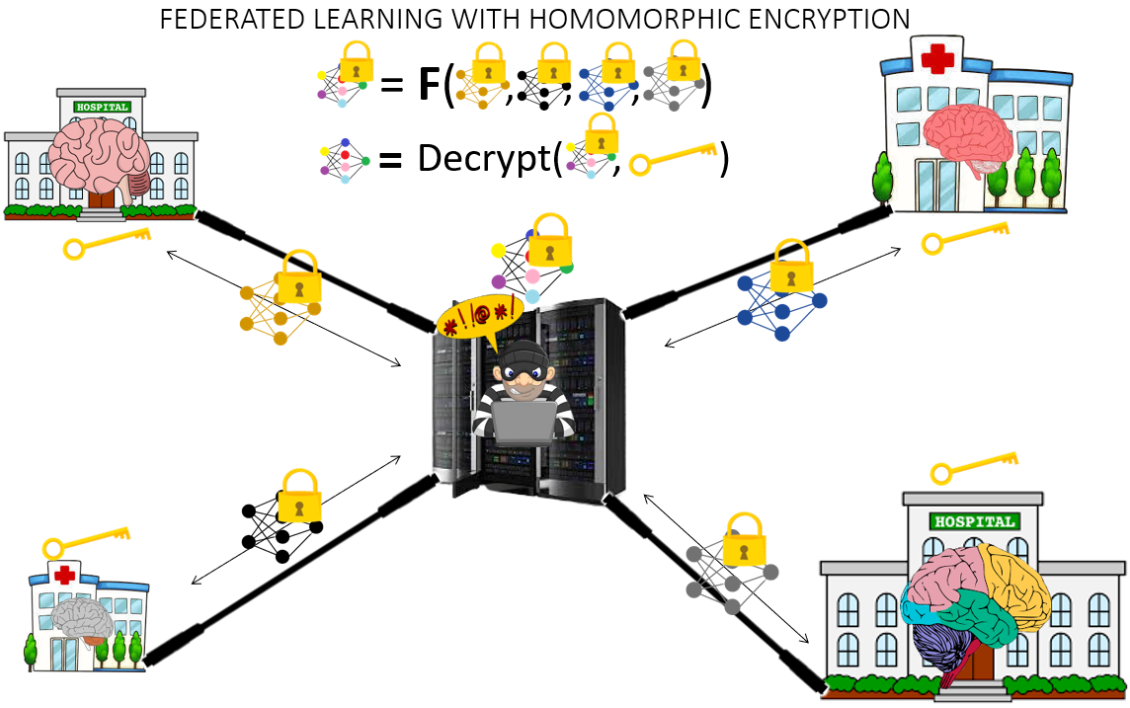


Figure 2.9: It is possible to create a federated learning system that is fully protected against malicious servers by using homomorphic encryption. This is an asymmetric encryption technique in which only the clients have access to the key. Practically it works as follows. Clients encrypt the model updates before sending them to the server. Thanks to the homomorphic property the server can still perform mathematical operations such as aggregation on the encrypted data. The still encrypted new global model is then sent back to each client who decrypts it and continues the process as usual. An important upside of HE compared to DP is that the former does not in any way depreciate model performance, whereas excessive DP can cause the model to diverge.

There are primarily two different homomorphic encryption algorithms that are used in the literature. The most commonly quoted is Pallier encryption [141, 39, 40]. Another technique is LWE encryption [141]. Others use a threshold variant of Pallier encryption which requires that t out of n clients must agree at each decryption stage, which means that the protocol also protects against collusion between a subset of

clients thus providing even greater protection [40]. However, at the end of the day all of the HE protocols achieve the sought after protection, but with its own pros and cons [141].

On an ending note, a general downside with using homomorphic encryption is that it incurs computational overhead since it has to encrypt and decrypt all data, which depending on the encryption protocol might be substantial. Perhaps more crucial is that it increases the communication cost by a factor of 2-3, depending on the protocol, thus making it impractical if network constraints are present [141].

2.11.2 Shamir's Secret Sharing

Another way to solve the privacy issue with model aggregation is provided by [133]. Instead of encrypting the data sent from clients to the server, they rely on a technique called secure sharing. The proposed solutions is designed to be robust to the inevitable case that one of the clients drops out during aggregation [133]. They use Shamir's t-out-of-n Secret Sharing, which means that each client splits its data into n pieces. The data can only be constructed if a party has at least t of these pieces. The algorithm works as follows, explained using an example where the data is a single scalar for simplicity. Each client randomly defines a polynomial f of degree t-1, such that f(0) is equal to the data. The client than send the tuple (i,f(i)) to client i, for all of the n clients. This means that only if at least t of these shared points (i, f(i)) are known can we recreate the initial data since a polynomial of degree t-1 is uniquely determined by t points. We now note that the aggregate result that we are interested in is the sum of each clients polynomial intersection with the y-axis. If each of the clients now add all of the pieces that it got from the other clients, thus effectively shifting the polynomial, and send the sum to the server then the server can use these sums to recreate the aggregate polynomial and thus find the true aggregate of the initial data points that we wanted to share. The algorithm thus allows the server to calculate the aggregate, without any party knowing any of the other parties' local model [133].

If we compare the secret sharing technique to the homomorphic encryption technique we realise that both of them leave the final outcome model unaffected. The latter, however, requires all clients to share the same encryption key thus increasing the attack surface. The secret sharing technique further requires that at least t clients are compromised in order for privacy to leak. Further, the paper by [133] show that they can implement the method in such a way that network cost is increased with less than a factor of two, thus making it more communication efficient than the homomorphic encryption schema. The conclusion is thus that both HE and secret sharing protect the FL system against malicious servers, thus serving the sought after purpose. Which one is better is application dependent and a matter of taste.

3

Methods

This chapter presents the experimental setup used to answer the posed research questions. Firstly, a detailed description and discussion of the used data set is presented, as well as the image processing techniques that have been used. The chapter then moves on to present the local model before turning to discuss the federated setting. After this the section naturally moves on to discuss the implemented privacy mechanisms. Each of the local, federated and privacy sections of this chapter include a detailed breakdown of the experimental approach that was used to analyse how different parameters impact performance. Finally, the chapter is concluded with a description of the hardware and software configuration.

3.1 Data management and image processing

3.2 Description of the used data set

The data used in this study was collected from six independent sources and contained seven different tumours sub types, including all the most commonly occurring. These were Meningioma, Neuroma, Pituitary Adenoma, Low and High Grade Glioma, Metastases and a final seventh class referred to as 'Other' that contained a mix of all less commonly occurring tumours. The total extent of the data set was 3146 unique tumour patients, each represented by a 3D MRI volume and corresponding ground truth segmentation mask. As discussed in the theoretical review the aim of the study was to achieve good segmentation, not classification, which meant that the objective was framed as a binary problem. The ground truth is thus a binary mask where a 1 represent a tumourous pixel. Besides these positive cases another 540 cases of healthy patients were collected. These were not used during training, instead they were later used as a way to evaluate the false positive rate of the model. All in all the total extent of the data set was thus 3686 samples.

The majority of cases are from Veterans General Hospital in Taipei (VGHTPE). Henceforth these cases are referred to as In-House. Roughly half of these cases are annotated with premium quality since the annotation masks were intended to guide the Gamma Knife treatment, which requires a mm precision. The other half originate from a pathological database. The annotation quality for these cases are adequate in general, but considerably less detailed than the Gamma Knife cases for

natural reasons. The In-House data included cases of all seven different tumour sub types. Since the goal of the study was to create a data set that emulates real world conditions to the greatest extent possible, these In-House cases were complemented with external, publicly available data sets. These were all downloaded and used with permission and in accordance with the licensing agreements. All In-House cases were collected with the patient's permission.

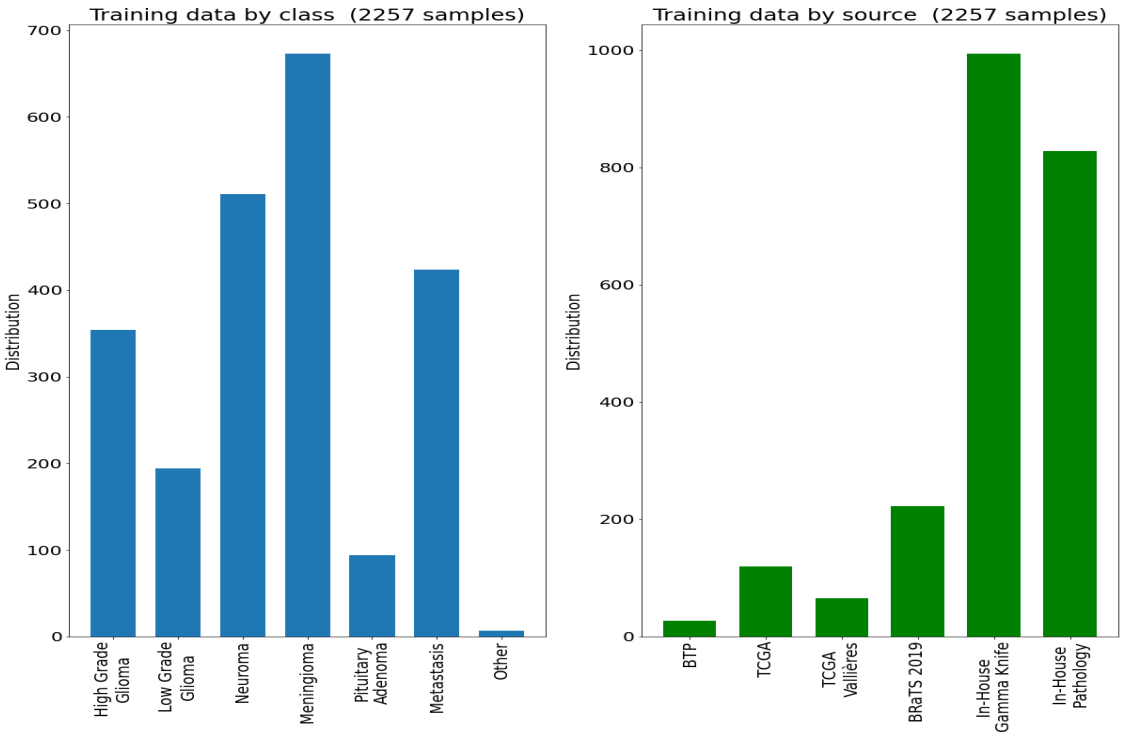


Figure 3.1: The used data is collected from six different sources and contains seven brain tumour sub types. The majority of the data is collected In-House at VGHTPE, either from a Gamma Knife radiosurgery program data set or from a pathological database. The rest of the data originates from different public databases. The relative abundance of tumour sub types varies by orders of magnitude, with Meningiomas being the most common and Pituitary Adenomas being the least occurring. The sub type referred to as 'Other' is the union of all cases that do not belong in either of the other six classes.

The largest such external data set is the already discussed BRaTS 2019 [147, 148, 149, 150, 151]. The BRaTS data contains a mix of High and Low Grade Glioma cases. The second largest external set was collected from TCGA [152, 153, 148, 150, 151] and also included both High and Low Grade Gliomas. The third data set, referred to as TCGA Vallieres [154, 153, 155], as well as the last data set, referred to as BTP [3], only contained Low Grade Gliomas. Besides these three data sets there is essentially only one more publicly available set for brain tumour segmentation, namely [156]. This data set is the only external set that contains other cases than Gliomas, which makes it especially valuable. However, it only contains 2D images which means that it had to be excluded from this study.

It is worth mentioning that integrating samples from this diverse set of sources is a very tedious and time consuming work. The reason for this is that different sources store data according to different protocols, provide annotations in various formats, contain different feature spaces etc. Note that this is despite the mentioned fact that hospitals world wide have largely agreed to follow the same DICOM standard for data storage [31]. Even within the standard there is room for significant inter-hospital variations. Hence, significant effort was put in to homogenise the full extent of the data set in order to convert it into a general format which could be interpreted by the model. However, the gain from performing this initial grunt work is that the resulting data set is more varied and extensive which will lead to better chances of building a more generalisable model.

The collected data was split into a training set, a validation set and a test set. An average of 12 previous studies suggest a 65%, 9% and 26% split across those three sets respectively [97, 99, 103, 112, 110, 87]. Based on this the used data was split randomly and uniformly according to 70%, 10% and 20% respectively. The slightly lower test proportion was motivated by the fact that an additional 540 healthy cases were kept for final test evaluation as well. This meant that the training set contained 2257, the validation set 282 and the test set 607 tumour samples respectively. The distribution by tumour sub type and data source in the training set is illustrated in Figure 3.1. Note that the distribution by virtue of random sampling is similar for the validation and test sets as well.

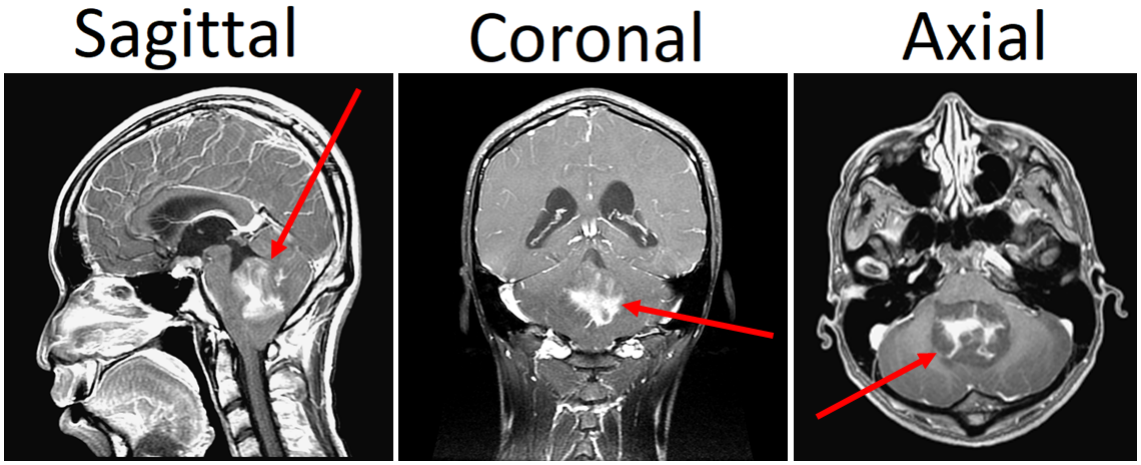


Figure 3.2: MRI scans of the brain is commonly performed in three different views. This allows practitioners to analyse suspected tumours from several angles, which simplifies diagnostics. Note the significant difference in visual appearance for the same tumour in the three views, as indicated by the arrows. However, this study only uses axial images for simplicity.

The study only considered a horizontal federated learning scenario, which meant that the all samples had the same feature descriptions. Practically this meant that only a sub set of the available MRI pulse sequences and views, see Figure 3.3 and Figure 3.2 respectively, were used. More precisely the study only considered Axial view images taken in the T1-weighted post contrast pulse sequence. The motivation

behind choosing this specific sequence is that it is commonly used by practitioners, since tumours show up bright and clear [17, 14]. A downside is that the so called peritumoral edema in Gliomas show up less bright in this pulse sequence than in for example T2-weighted of T2-FLAIR, as shown in 3.3. However, the choice is still well motivated and adequate for the purpose of this study.

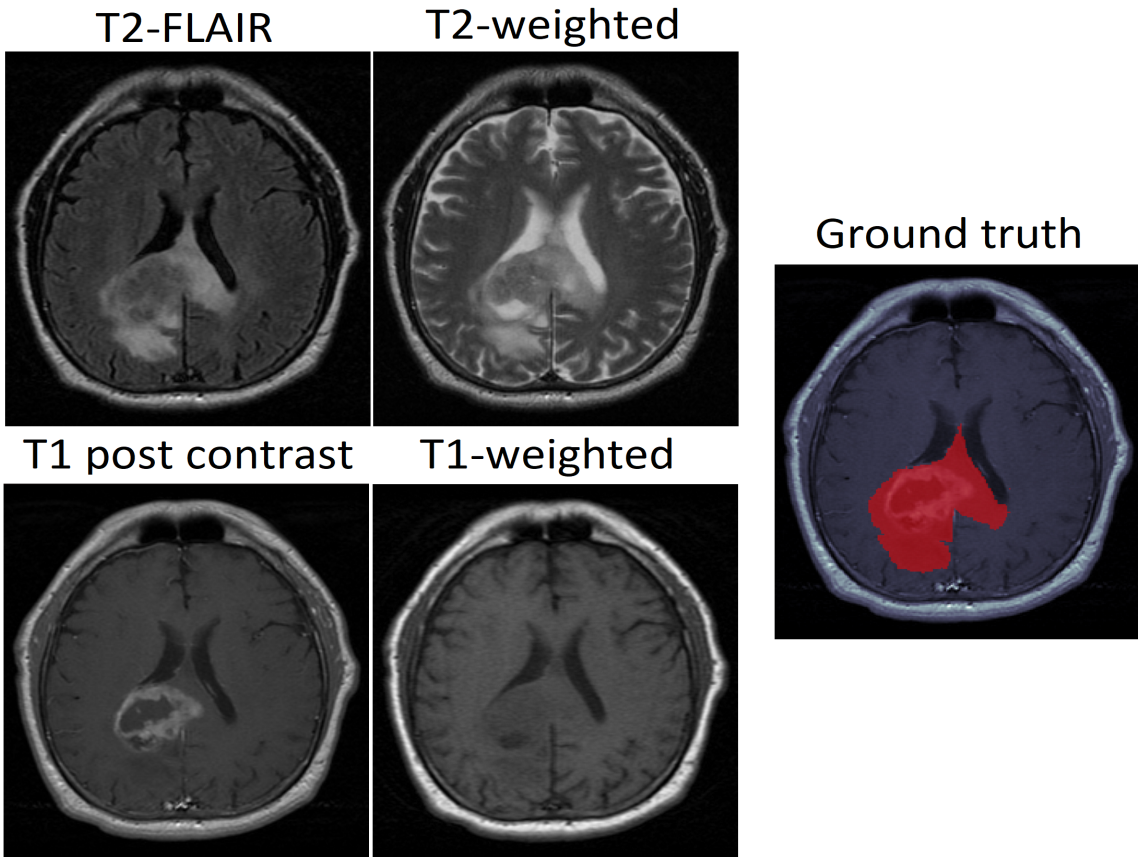


Figure 3.3: MRI scans of the brain are commonly performed in several different so called pulse sequences. The four most common are T1-weighted, T1-weighted post contrast, T2-weighted and FLAIR. The latter is an improved pulse sequence that is specifically designed to suppress strong signals from fluids, thus making the tumour more discernible. The effect of this is apparent by comparing the T2-FLAIR and T2-weighted cases in this illustration. Different types of tissues, including tumourous material, show up differently in each of the pulse sequences. This allows practitioners to analyse suspected tumours from several complementary images, which simplifies diagnostics. For example, peritumoral edema tend to show up brighter in T2-weighted and T2-FLAIR, whereas it is barely visible in T1-weighted or T1-weighted post contrast. Especially note the visual discrepancy in the perceived tumour extent when analysing each of the pulse sequences on its own. However, this study only uses T1-weighted post contrast images for simplicity which means that the model has a significant disadvantage. Nevertheless it should be mentioned that the entire tumour does appear in all pulse sequences, albeit not always to the extent that the human eye can perceive it. Lastly, although the illustration only shows T2-FLAIR it is also common to use T1-FLAIR.

Having described the origin and extent of the data set, it is relevant to discuss the matter of data quality and annotation ambiguity. As discussed in the theoretical review, tumour annotation is far from trivial even for experienced doctors [19, 35]. This leads to a great deal of subjectivity and inter-rater disagreement. In fact it has been shown that the overlapping annotation for Gliomas by two experienced medical professionals is a meager 0.74-0.85 in terms of dice score [93, 61]. As a side note, an important consequence of this is that human level performance can be considered to be in the same ball park.

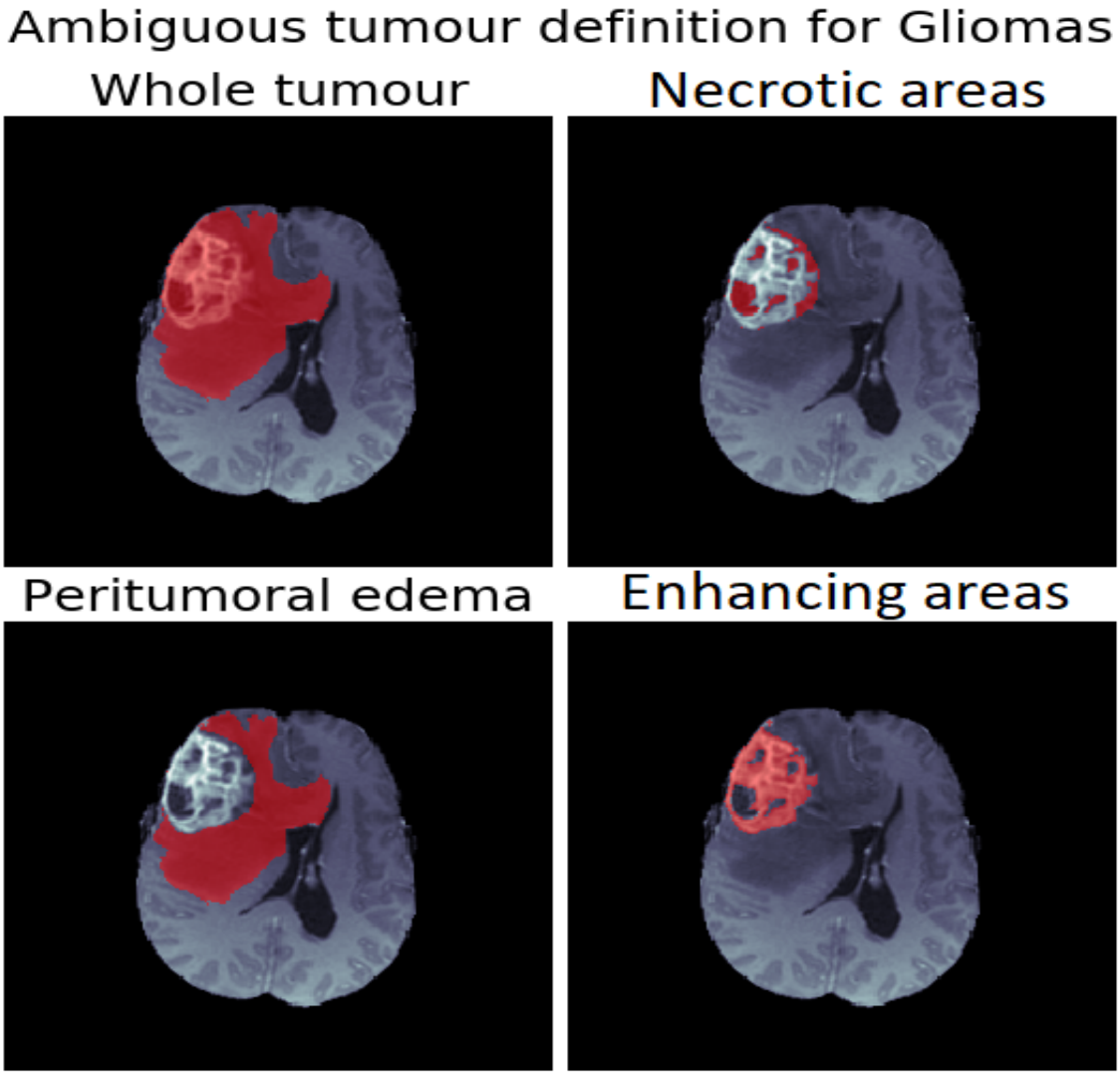


Figure 3.4: Glioma tumours can be divided into three subsections: necrotic areas, enhancing areas and peritumoral edema. Different institutions use different definitions of whether all or only some of these areas should be included in the tumour annotation. This implies that the so called ground truth for a sample will differ significantly depending on where the sample was collected from, leading to inter-source heterogeneity. In this study the vast majority of cases follow the 'Whole Tumour' definition. However, to emulate the real world ambiguity subsets of cases that follow the other definitions are also included. See Figure 3.6 for an example of this discrepancy.

This fallacy can be understood by looking at Figure 3.4. The reality is that different doctors and institutions use different definitions concerning what constitutes a tumour. Although this is especially true for Gliomas, some degree of ambiguity exist for all tumour sub types. Since the goal of the data collection was to gather a real world simulation, cases were included that follow different such definitions. This leads to a great deal of heterogeneity and ambiguity, which makes the model learning scenario significantly more difficult [36].

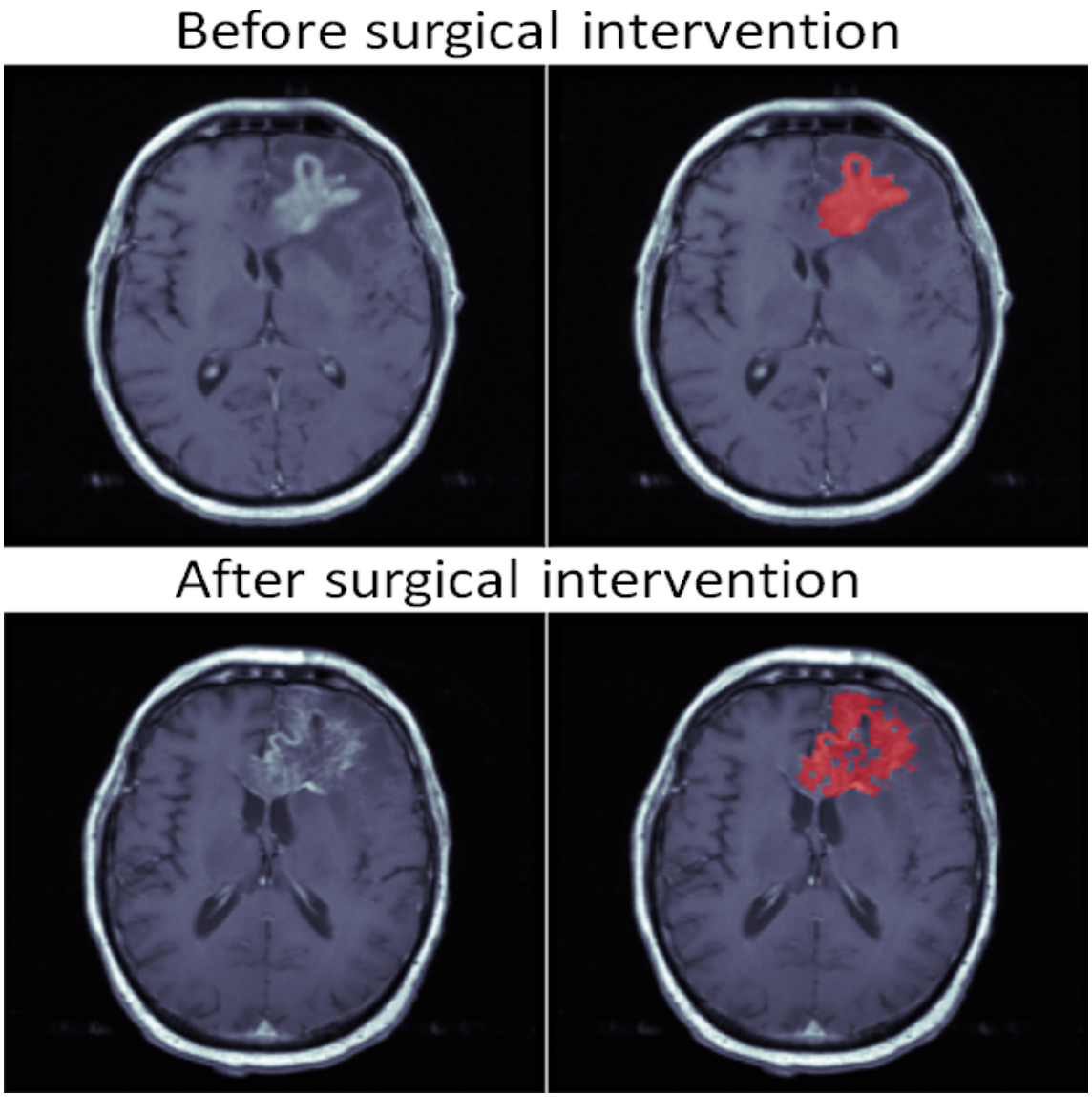


Figure 3.5: Brain tumours have different visual characteristics before and after surgery. Typically, the tumour boundary become more diffuse after the operation which complicates accurate, and objective, annotation of the tumour extent. For this reason a majority of public data sets, including the leading benchmark BRaTS, only consider pre-surgery cases. Likewise the vast majority of cases in this study are also pre-surgical. However, to emulate real world conditions a subset of post-operation samples are also included from the data set referred to as BTP [3].

Another source of ambiguity concerns the fact that the visual appearance of tumours before and after surgery is very different [18], as illustrated in Figure 3.5. In general the boundaries of post surgery tumours are less distinct, which increases the inter-annotation difference since the diagnosis becomes less objective. For this reason the people behind the leading benchmark data set BRaTS have decided to exclude all post-surgery cases in the most recent versions to alleviate the learning objective [87]. Similarly, the vast majority of cases in the data set used in this study were also pre-operation. However, to emulate real world conditions a small subset of post-operation patients were also included. They all originated from the data set referred to as BTP [3].

Before finally moving on to discuss the experiments and models, it is instructive to visually observe some representative cases from the data set. One such sample from each of the six main sub types are illustrated in Figure 3.6. The images are after the preprocessing described in section 3.3. Since the tumour sub type referred to as 'Other' is not a homogeneous subset no good class representative can be chosen and it is thus excluded from the Figure. However, to a layman the 'Other' cases in general visually look similar to cases from the other six classes. The cases in Figure 3.6 highlights several relevant points regarding the inter-class differences. For example the Gliomas are generally larger with much less distinct boundaries, compared to the other sub types that tend to be easier for a layman to identify. Furthermore, the Metastases tend to be significantly smaller than all other sub types. It is also the only sub type that patients tend to have multiple of, sometimes with as many as 20 tumours, although on rare occasions patients can have multiple cases of other sub types as well or potentially even a mix of different types. Furthermore, the statement that Gliomas in general are larger and that Metastases are smaller is supported by Figure 3.10.

Another important observation is that the Pituitary Adenomas by definition always occur at the distinct location illustrated in Figure 3.6. On the same note Meningiomas by definition occur close to the Meninges, meaning that they are located in between the brain and the inner surfaces of the skull. Similarly, the Neuromas tend to be located at its own unique location and it usually shows up with a more complex shape, as illustrated in Figure 3.6. The implication of all these observations is that although the learning objective is a binary problem, each of the seven sub types has its own unique traits. Importantly, this means that a model will behave differently depending on which sub types it was trained on. This point is going to be crucial in the federated setting, since each local model will be exposed to its unique distribution of tumour sub types.

Finally, the representative cases in Figure 3.6 also highlight two important quality deficits. The first one concerns the two Gliomas. In general these two sub types visually look similar to a layman, however in this Figure they are very different. This is a consequence of the annotation definition ambiguity discussed in Figure 3.4. The high grade Glioma has been annotated using the 'Whole tumour' protocol, whereas the low grade case is only annotated for enhancing regions. This illustrates the occurrence of large intra-class ambiguities in the used data set.

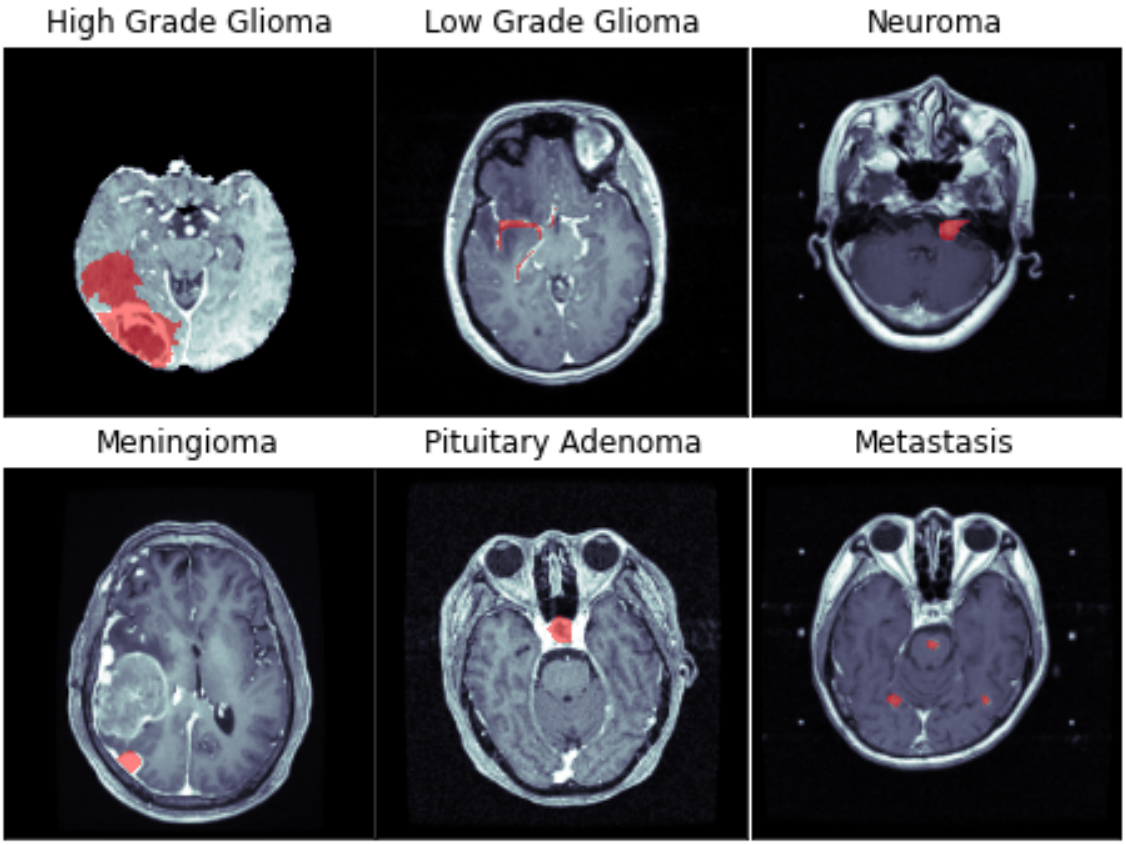


Figure 3.6: Representative images for the six main tumour sub types used in the study. The seventh class, referred to as 'Other', is excluded since no good class representative can be chosen due to the heterogeneity. The illustration highlight several important class dependent characteristics. Firstly, it shows the the tumour volume varies by orders of magnitude across sub types. Secondly, it illustrates that Metastases is the sub type that most often tend to result in multiple lesions. Thirdly, it presents that sub types tend to occur in distinct location. Pituitary Adenomas and Neuromas by definition occur around the specific locations illustrated. Meningiomas by definition occur in between the brain and the inner surfaces of the skull, connected to the Meninges. On the other hand, both Gliomas and Metastases can show up virtually anywhere. Finally, the images demonstrate the issue of annotation quality. Firstly, the Meningioma case shows that a large abnormal region is not annotated. This might either be because it is not actually a tumour, although it certainly looks like it to a layman, or that the annotator has simply failed. Either way, it poses a difficulty for the model learning scenario. Secondly, the low grade and high grade Glioma annotations are vastly different although they in theory should be very similar. This is because these specific cases are annotated using different definitions, as discussed in Figure 3.4. This discrepancy exists for both high and low grade Gliomas and is not class dependent.

The second quality deficit that Figure 3.6 reveals concerns the Meningioma case. The image clearly shows a large abnormal region just above the tumour, but this region has not been annotated by the medical professional. This illustrates that the

so called ground truth in the used data does not have perfect recall but instead contains significant imperfections. Although this is a relatively rare case, it accurately emulates the real world scenario in which the occasional human error is prone to exist. An adequate model must thus be robust to these imperfections.

3.3 Preprocessing standardises inputs and mitigates the class imbalance issue slightly

As discussed in the theoretical review preprocessing as a necessary initial step, since the model requires all inputs to be drawn from at least approximately the same distribution in order to ensure proper convergence and inference. The following analysis and discussion concerns only the training data in order to prevent information leakage to the validation and test sets, which would unfairly bias the evaluation performance. Instead, whatever process is found to be best practice for the training data is finally blindly applied to the two other sets as well.

As expected the raw training data did not exhibit the necessary IID properties, as seen in Figure 3.7. Importantly the Figure reveals that there exists two drastically different image storing protocols in the data set, in which the pixel values are spread around different centres. Besides this absolute pixel shift the images also present a large difference in standard deviation, or equivalently stated the pixel intensity resolution varies by orders of magnitude. Both these issues implies that sample normalisation is crucial, a conclusion that is in line with the recommendations from previous studies.

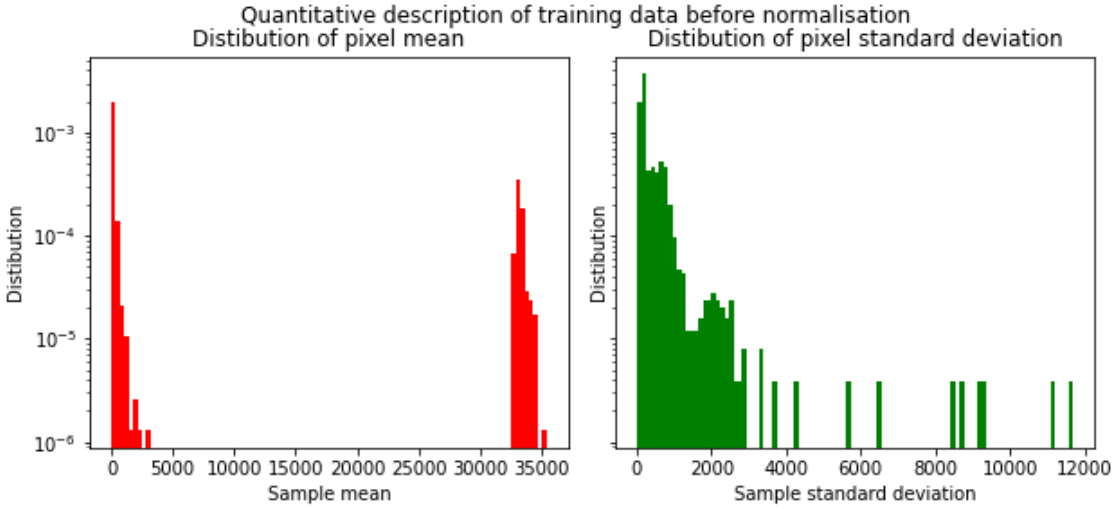


Figure 3.7: The original samples have very different pixel distributions. The two large peaks in the left hand image suggest that two different image protocols are present, in which the pixel values are shifted significantly. Furthermore, the pixel resolution varies by orders or magnitude across samples, as seen by the large difference in standard deviations in the right hand image. The consequence of this heterogeneity is that normalisation is a necessary preprocessing step.

The resulting distribution after sample wise normalisation is presented in Figure 3.8. It is evident that the normalisation, although it is relatively naive and simple method, helps significantly to produce a more homogeneous data set. As discussed in the theoretical section performance might be improved further if the outliers are removed before normalisation. However, different authors propose different interpretations of what constitutes an outlier. One paper proposes to clip the top and bottom 0.2% [109] and another propose the more restrictive 1% [99]. By empirical evaluation this study finally chose to follow the latter convention. As seen in Figure 3.8 this outlier cleaning indeed produced a tighter inter-sample spread. The resulting distribution was considered satisfactory and adequate for the purpose of this study which meant that the less commonly quoted usage of N4-ITK bias field correction or Wiener filtering was disregarded. According to the empirical findings of previous papers these additional mechanisms have a negligible effect.

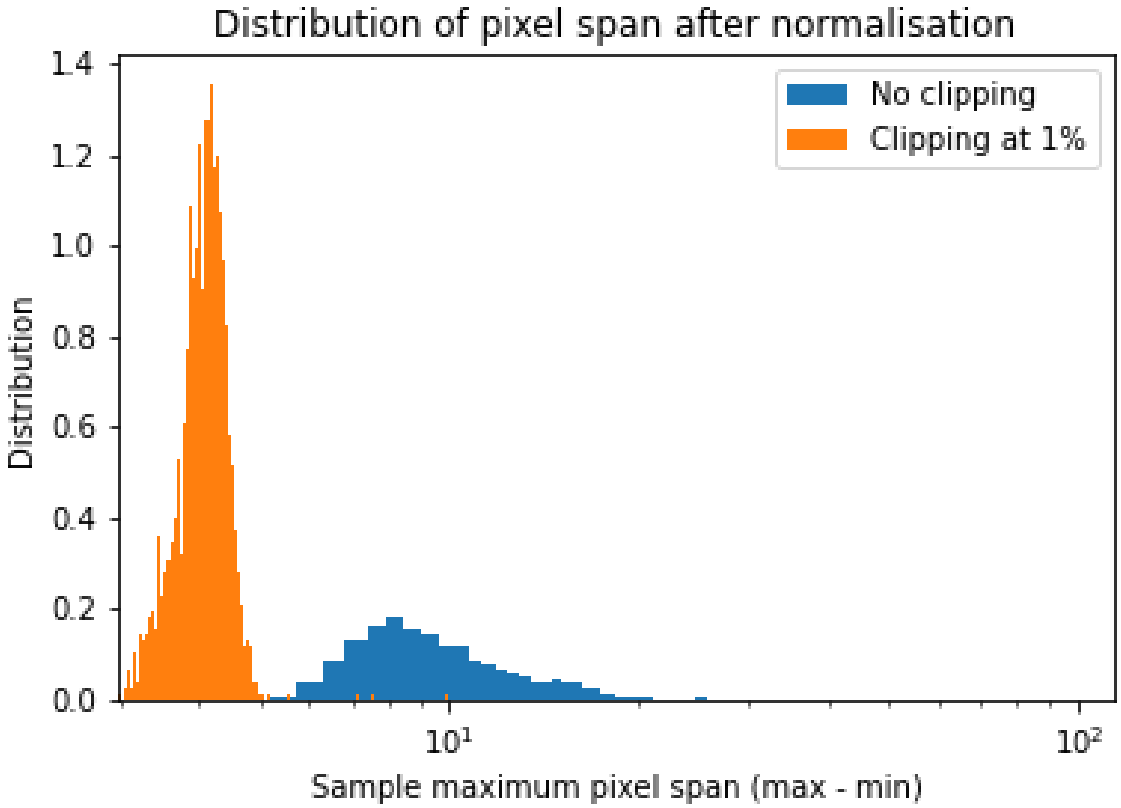


Figure 3.8: After normalisation all samples have zero mean and unit variance by definition, however pixel outliers might still exist. By clipping the pixel values at the 1st and 99th percentile this issue is mitigated. Comparing this final distribution to the original, illustrated in Figure 3.7, it is clear that the preprocessing produces a much more homogeneous data set which simplifies the model learning scenario.

The preceding discussion handles inter-sample issues, however it is necessary to consider statistical complications related to individual samples as well. As discussed in the theoretical review one of the main difficulties in medical AI is the large class imbalance. According to a study as much as 98.46% of the brain volume in

tumour patients is healthy tissue [99] which leads to a skewed learning scenario. This statement is supported by the findings concerning this data set as well, as illustrated in Figure 3.10 and Figure 3.11. In an attempt to mitigate this another preprocessing step was performed in which all edge slices of the brain volume that did not contain any tumourous materia was discarded. As seen in Figure 3.9 this process provided an immense 49.2% reduction of the imbalance problem. As an additional upside this also reduces the memory consumption and computational expense significantly which speeds up training. Note that this trimming process can only be performed on training data, since it requires access to the ground truth.

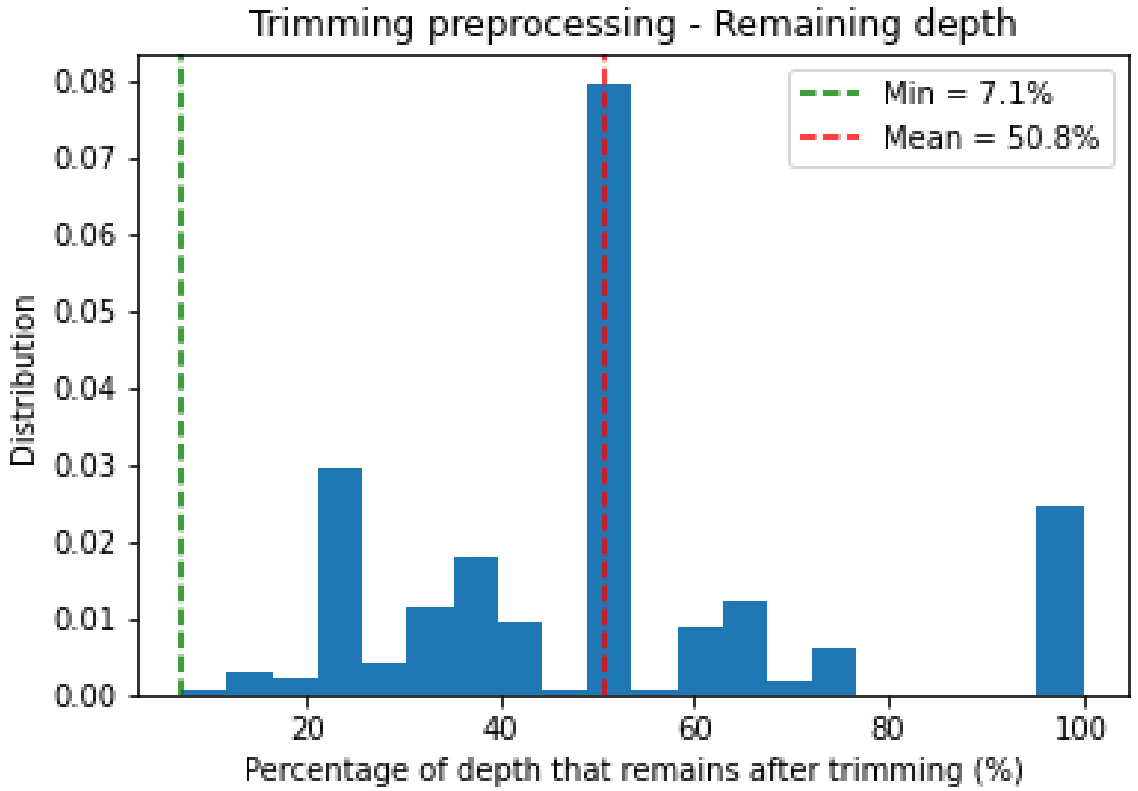


Figure 3.9: The vast majority of the brain is naturally non-tumourous, as illustrated in Figure 3.10 and Figure 3.11, which leads to a severe class imbalance issue. To mitigate this a preprocessing step was performed that removed all non-tumourous edge slices. This process resulted in a significant 49.2% depth reduction on average. Firstly this means that the magnitude of the class imbalance problem is cut in half, but it also has the upside that it speeds up training significantly due to the lower dimensionality.

It is worth analysing this class imbalance issue further before moving on to the final preprocessing steps in order to gain instructive insights on the learning scenario at hand. Even after the edge trimming the tumour volumes only occupy a tiny fraction of the skull, see Figure 3.10. Besides, it is evident that this class imbalance issue is note the same across tumour sub types. The box plot reveals that the magnitude of the issue is orders of magnitude worse for some, such as Neuromas and Metastases. This further enhances the non-IID issue already discussed.

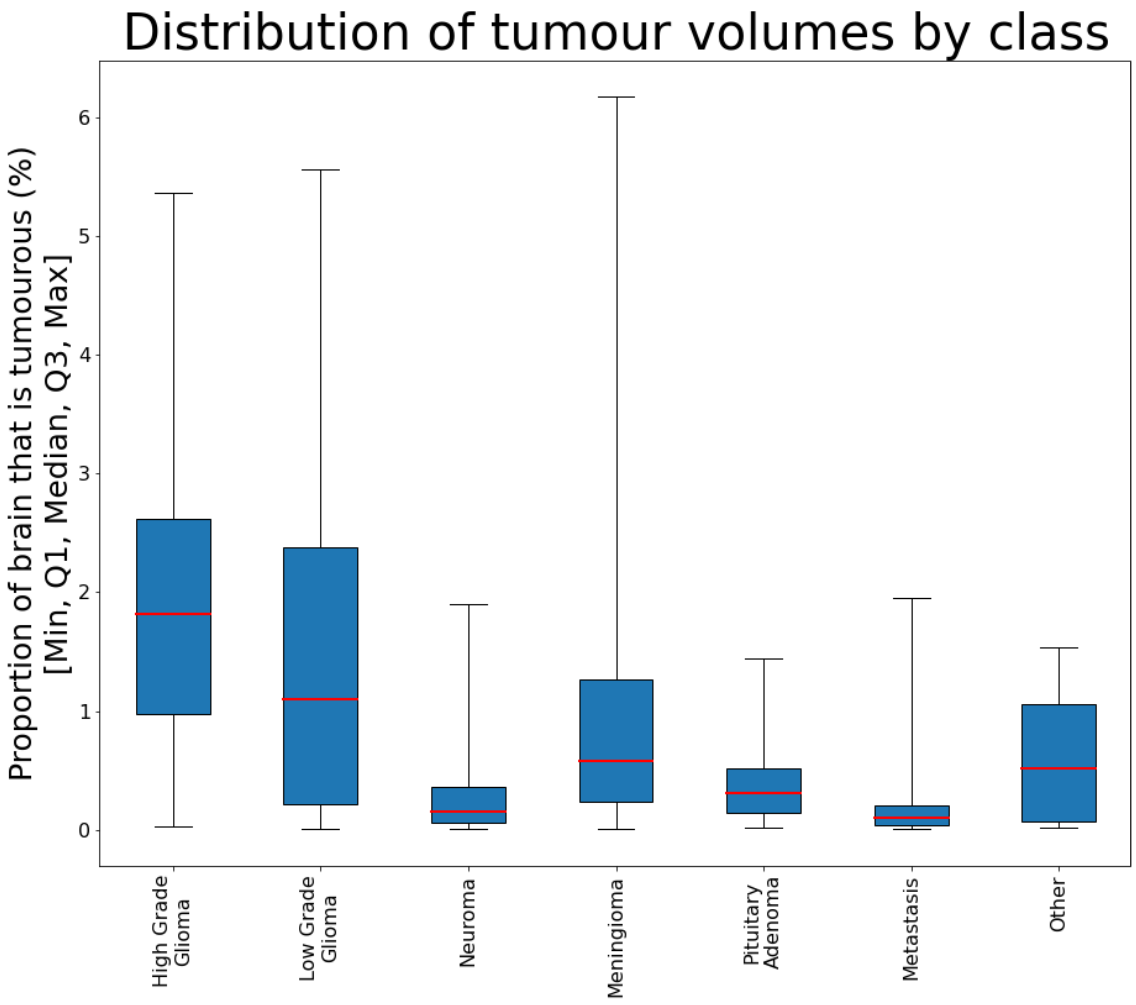
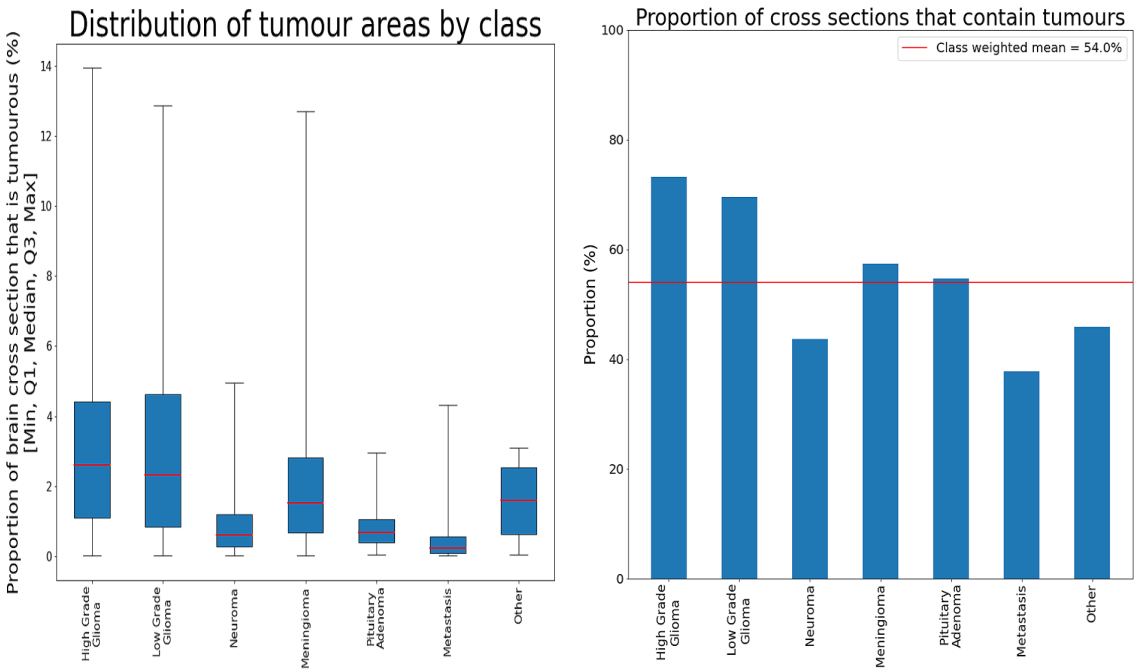


Figure 3.10: The binary class problem under consideration in this study is highly unbalanced since the tumour volume is only a few percent of the full brain volume. Note that the illustrated distribution is after preprocessing, meaning that the true imbalance that is present in the blind validation and test data is much more severe. Furthermore, the box plot reveals that different tumour sub types exhibit very different characteristics in terms of size. For example, the Glioma tumours are significantly larger than Metastases or Neuromas. Besides this inter-class difference there is also a large intra-class spread in this regard, especially for the Meningioma class. This implies that the model must be able to properly handle tumours with volumes that varies by orders of magnitude.

Furthermore, as shown in Figure 3.11b, even after trimming a significant proportion of slices are entirely without tumours. Besides, the slices that do contain tumours are still primarily made up of healthy tissue as illustrated by Figure 3.11a. In fact the largest cross sectional tumour area of all the 3146 cases is no larger than 14%. Again, this imbalance will likely cause difficulties in terms of convergence. However, there is a glass half full insight to be derived by comparing the two box plots in Figure 3.10 and 3.11a. Since the class wise distributions are very similar this suggests that the tumour depths are actually relatively independent of the tumour

type. Consequently this specific factor does not further complicate the non-IID issue, which is good news.



(a) The tumour cross sections tend to be very small but proportionally larger than in terms of volume, see Figure 3.10.

(b) Even after the depth trimming discussed in Figure 3.9 only 54% of the slices contain any tumourous materia at all.

Figure 3.11: In general tumours tend to be small both in terms of cross sectional area and volume. The latter is illustrated in Figure 3.10. The class wise distributions of areas and volumes are very similar, which suggests that the tumour depth is not highly correlated to the sub type. Furthermore, only half of the depth contain any tumours at all. All these findings support the statement that the model learning scenario is highly unbalanced, as expected.

Besides the statistical preprocessing already performed it is necessary to consider practical issues as well. Firstly the used model, see Figure 3.18, requires that all image dimensions are a multiple of 16. However, samples do not need to have the same dimensions since the model i fully convolutional. For this reason all image depths were padded to the nearest multiple of 16. Note that the images were normalised before padding, in order to maintain the same intra-skull distribution. To avoid edge artefacts the padded pixels were given the same intensity as the background, namely the lowest value after normalisation. A final note concerning the previously discussed edge slice trimming, see Figure 3.9, is that this process was performed down to the smallest multiple of 16 and not further for this practical reason.

Finally the width and height dimensions had to be processed. However, due to that the original data is very high dimensional, see Figure 3.12, padding was not the answer. The reason for this as discussed in the theoretical section is simply that these large volumes cause the GPU memory to overflow during training. Consequently

the images must be down sampled. Due to differing resource capabilities different researches uses a wide array of different widths and heights: 25 [96], 32 [96], 80 [92], 96 [92], 128 [109, 108], 144 [95], 160 [4, 98, 95], 192 [4, 98] and 224 [62]. Commonly the ones that propose the lower range use random patches, rather than down sampling the entire image. However the general consensus is that the best practice is to use the largest possible size that fits in memory. Furthermore, it has been empirically found that it is better to increase the image size to the point where a unitary batch size is used than the other way around [4]. Based on these arguments this study finally decided to use three squared input sizes in the mid to high portion of the proposed range, namely 96, 144 and 176. The latter was found to be the largest multiple of 16 that did not cause memory overflow. After down sampling this led to the creation of three different data sets. As will be discussed in section 3.6 individual models were trained on the respective sets in order to evaluate how the input size impacts performance.

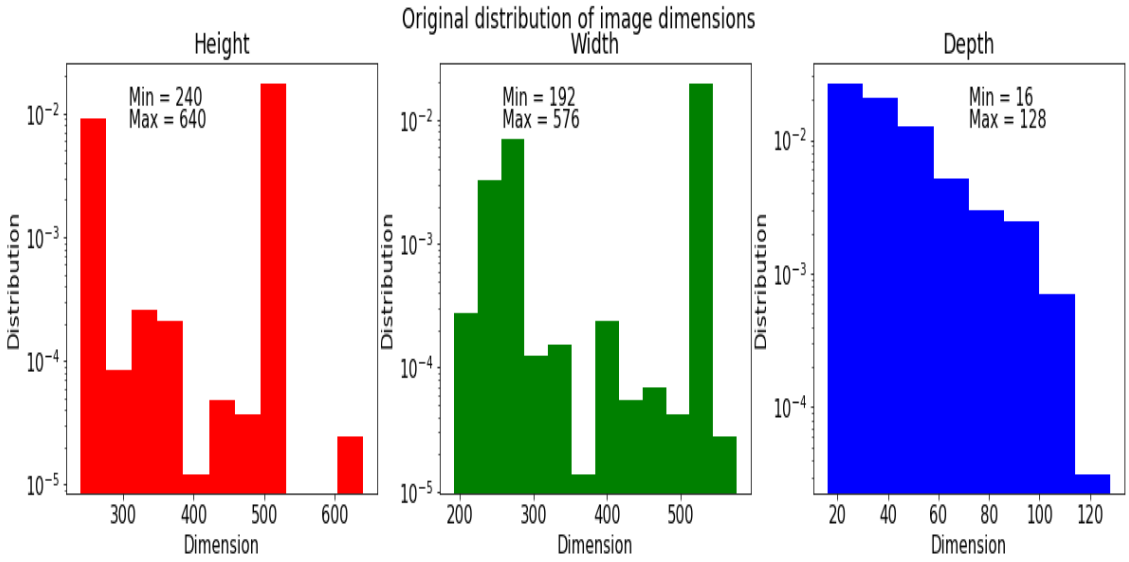


Figure 3.12: The original data is too high dimensional to fit in memory, with maximum dimensions of 640 by 576 by 128 pixels. Due to this the final preprocessing step in this study was to down sample the width and height. Three different down sampled, squared versions were created with dimension 96, 144 and 176 pixels respectively. Note that the illustrated distribution is after the initial depth trimming described in Figure 3.9.

3.4 Improved geometric and intensity invariance is achieved using extensive augmentation

Another common denominator between previous studies is that the vast majority advocate for the usage of data augmentation. Furthermore, one should perform both geometric and intensity based augmentation in order to improve model invariance to both phenomena. Relying on these empirical suggestions this study decided to employ both methods. This section describes and illustrates the used techniques.

In terms of geometric augmentation the most commonly quoted methods are mirroring and rotation. Some studies also suggests the use of cropping and up/down sampling. However, in order to avoid excessive levels of augmentation - which as discussed in the theoretical review degrades performance - it was decided to only use the mirroring and rotation techniques. More precisely the study used both vertical and horizontal mirroring, see the upper row of Figure 3.13. Different authors propose a variety of rotation angles, ranging from just 1 degree [92] up to 45 degrees [90] or even 90 degrees [108, 90, 95]. Since no general consensus exist and there are no counter arguments laid out for why arbitrary rotations are not valid, this is exactly what was decided for this study. More specifically it was decided to divide the rotation mechanism into three separate transformations: 90 degrees clockwise, counter clockwise or completely arbitrary rotations.

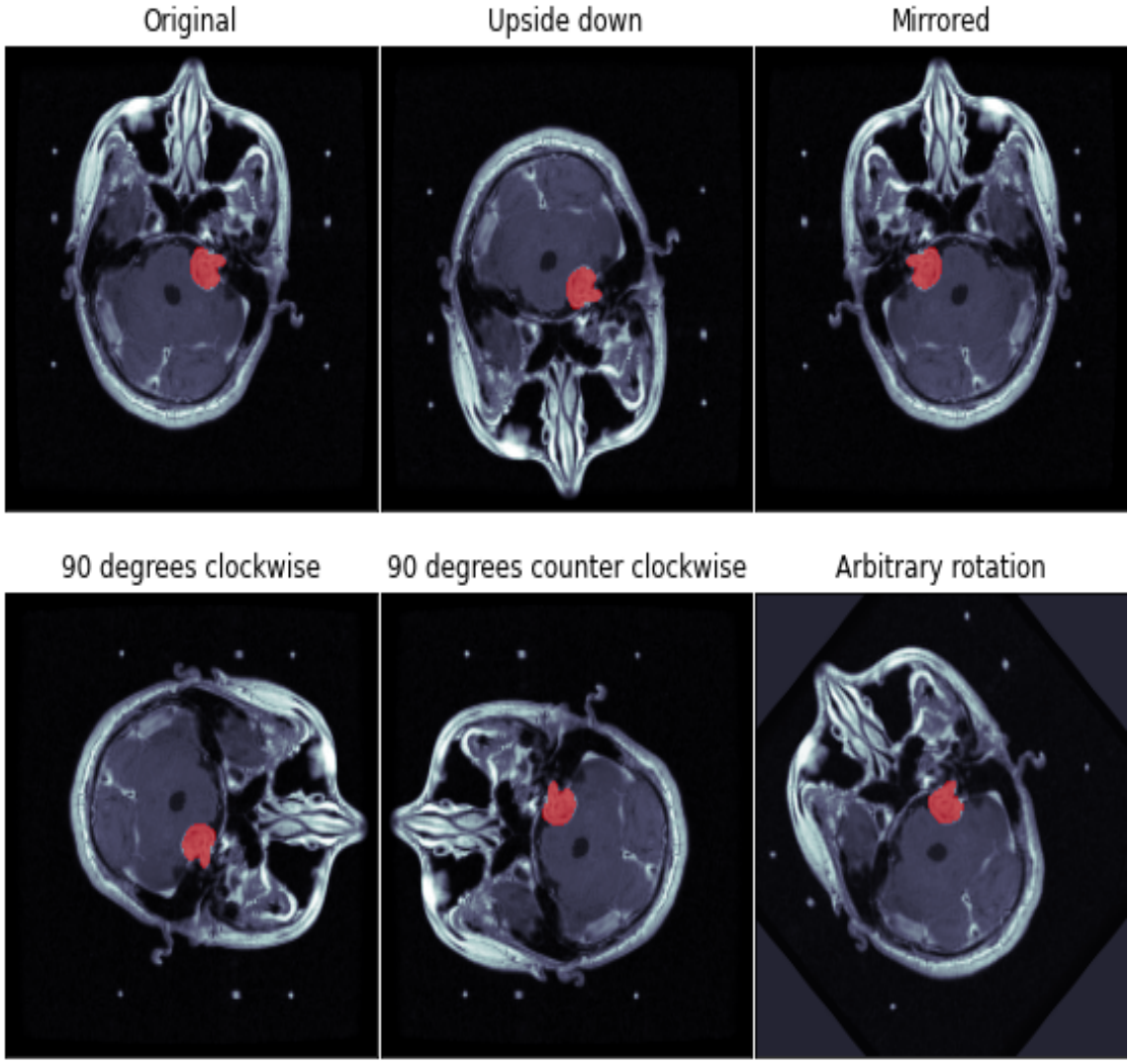


Figure 3.13: Visualisations of five of the six geometric augmentation techniques used in this study as a way to increase the variability of the training data. Four involve rotations and the fifth is a mirroring across the centre line. Besides these, a sixth method is illustrated in Figure 3.14.

As seen in Figure 3.13 all rotations and mirroring give rise to images that a human would be able to perform diagnostics on which suggests that it should pose a learnable scenario for the model as well. Although rotation offsets of more than a few percent is unlikely to occur in real life, meaning that such an invariance is not strictly necessary, it still serves the purpose of artificially increasing the data set which is motivation enough on its own. The same line of argument goes for the vertical, upside down case as well.

Besides the rotations and mirroring advocated by previous studies it was decided that it makes intuitive sense to introduce a novel augmentation technique which reverses the depth, see Figure 3.14. This is the analogue of allowing the model to view the brain from behind. Again this is motivated by the fact that a doctor would be able to diagnose a patient in this view as well. Importantly this technique should, in theory, make the model more robust to depth related heterogeneity. This is a very important property in inter-hospital AI since different MRI scanners and settings give rise to images of very different depth resolutions, as previously discussed.

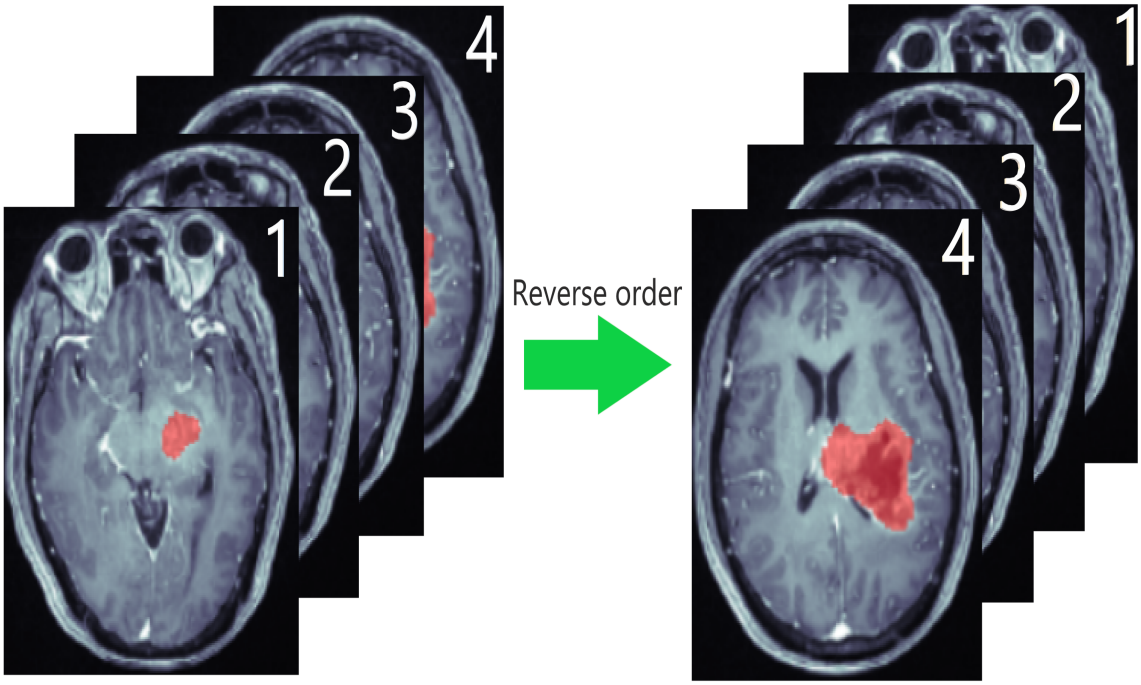


Figure 3.14: The last geometric augmentation technique, besides the five illustrated in 3.13, is to reverse the depth such that the model views the brain from behind.

All in all this gives rise to six geometric transformations. Besides these the study also incorporated the most common intensity based methods. From the theoretical review it is clear that the most common is to add Gaussian noise, thus providing some robustness to minor pixel imperfections. See Figure 3.15 for a representative illustration. It was empirically decided, based on visual appearance, to add zero-centred noise with 0.5 standard deviation. Note that the normalised images all have mean zero and unit standard deviation prior to augmentation.

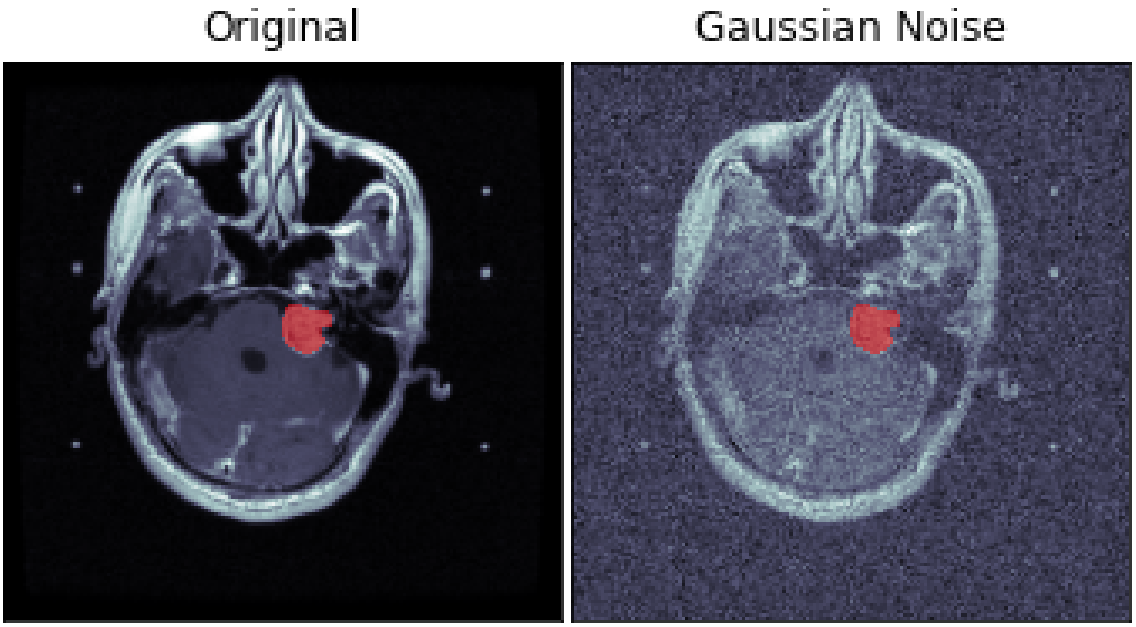


Figure 3.15: Besides the geometric augmentation techniques illustrated in Figure 3.13 and Figure 3.14 the study also uses three different intensity augmentation methods. The first one is to add random Gaussian noise, which hopefully makes the model more robust to pixel intensity variations. The original data was standard normal distributed and the added Gaussian noise was zero-centred with 0.5 standard deviation. The other two methods are illustrated in Figure 3.16.

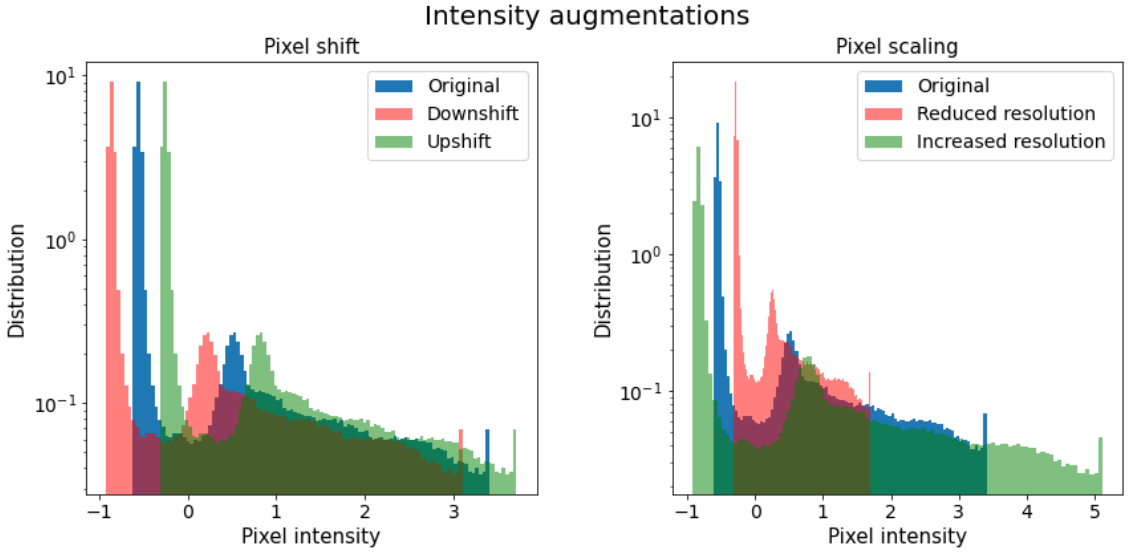


Figure 3.16: The last two intensity augmentation methods used in this study are pixel shifting and scaling. The former makes the model robust to minor pixel offsets and the latter provides increased ability to handle different pixel resolutions. For maximum variability the offset and scaling was chosen at random, up to a predefined threshold, for each samples during run time.

Lastly intensity shifting and scaling was implemented, see Figure 3.16. The motivation behind these methods are to make the model less susceptible to the minor inter-sample distribution differences that are still present even after preprocessing, as discussed in Figure 3.8. Again, different authors propose different levels of both shifting and scaling. For the absolute shift it ranges from 5% [89] to 10% [4, 98]. Concerning scaling some quote 5% [89], other 10% [4, 98] or even 20% [108]. Note that both operations can be either in the positive or negative direction. By empirical examination it was found that the used data set was actually robust to even larger shifts, up to 30%, and similarly scaling up to 40%. See Table 3.1 for a break down of used parameters.

Table 3.1: To investigate how different levels of augmentation impact performance three different setups were proposed. These used increasingly aggressive augmentation by allowing greater pixel shifting, scaling and image rotations. The likelihood that augmentation is applied to a samples was also increased. Note that the quoted probability is for each of the two mechanisms, meaning that the chance that a sample is not augmented at all is only $1 - P^2$. Furthermore, note that the scaling, shifting and rotation can be performed in either positive or negative direction up until the predefined level.

Reference name	Augmentation parameter			
	Probability P (%)	Scaling SC (%)	Shifting SH (%)	Rotation ROT (deg)
Minor	50	20	10	25
Moderate	90	40	30	45
Aggressive	90	40	30	180

As discussed by several previous studies a good practice is to perform online random augmentation, since this gives rise to a steady stream of slightly different cases which prevents overfitting. For this reason the following approach was used for each sample. First, by a given probability P geometric augmentation was applied. However for this end only one of the six described methods were chosen at random. After this, again with a probability P, one of the intensity based methods was chosen at random. Consequently, given a parameter P it is a $1 - P^2$ likelihood that no augmentation at all is applied and P^2 chance that both types are used.

Besides the parameter P, three other parameters were also included: SH, SC and ROT. SH is the absolute intensity shift, SC is the intensity scaling factor and ROT is the rotation used in the arbitrary rotation method. More specifically these parameters were set as the maximum. Practically a value was chosen at random uniformly in the range 1-SH to 1+SH, 1-SC to 1+SC and -ROT to +ROT for the three methods respectively each time it was applied. This introduces even more variability which prevents overfitting.

The remaining question is what these four parameters should be. It was decided to examine three different such settings to investigate how different levels of aug-

mentation impact performance. The three different setups referred to as 'Minor', 'Moderate' and 'Aggressive' levels of augmentation are presented in Table 3.1.

3.5 Naive postprocessing was introduced in an attempt to remove noisy predictions

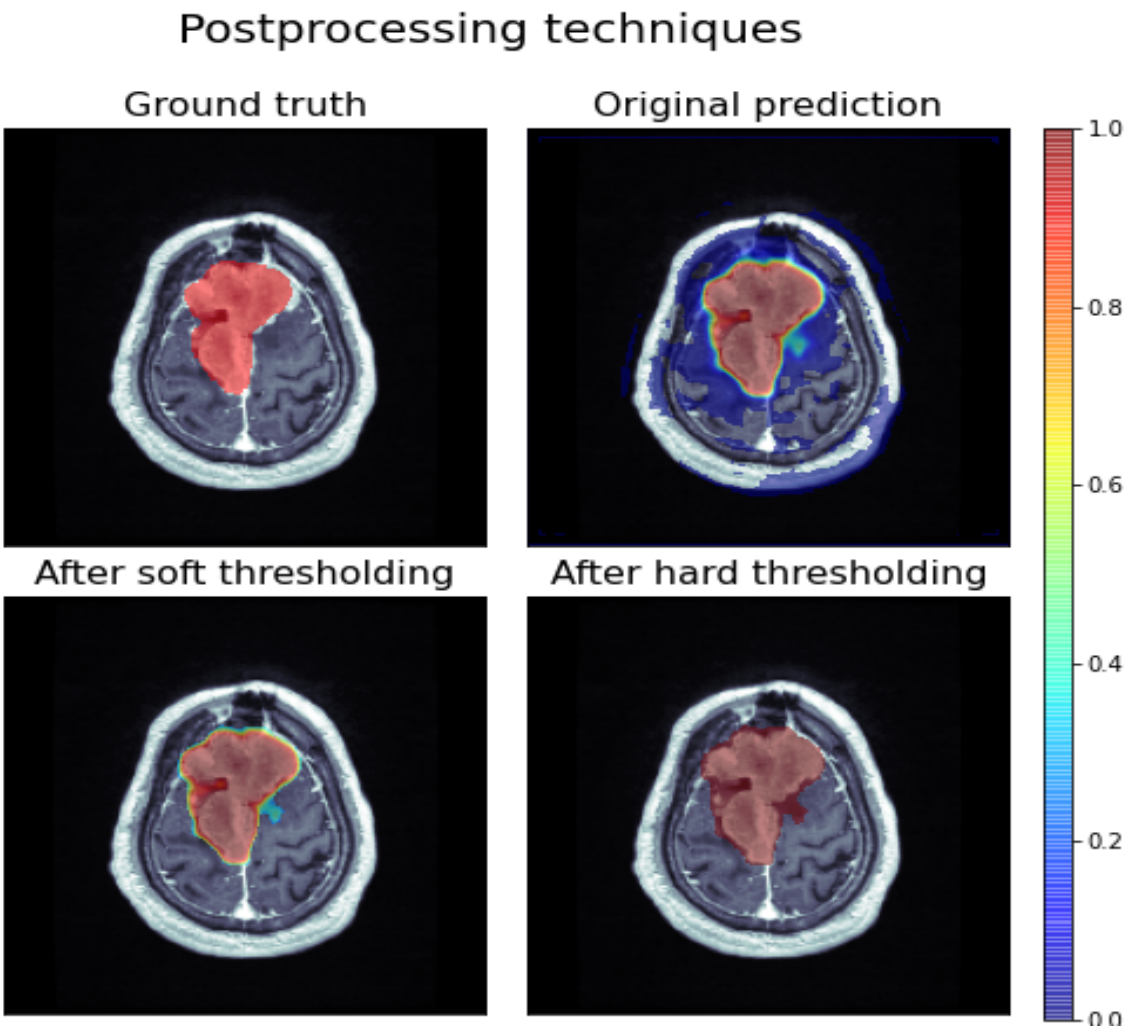


Figure 3.17: Two different types of simple postprocessing methods are examined in this study: soft and hard thresholding. They work as follows. First, the model predicts a value between 0 and 1 for each pixel that represented the likelihood that said pixel is a tumour. The postprocessing then rejects all predictions below a predefined threshold, the intuition being that low confidence predictions are likely incorrect as exemplified by the illustration. The subtle difference between the soft and hard thresholding is that the former keeps all remaining confidence values as is whereas the hard threshold turns the predictions into a binary classification. This difference is apparent by observing the two resulting heat maps in this illustration. The figure illustrates a Meningioma patient.

Preprocessing and augmentation are two methods that virtually all researchers use. However, some advocate for postprocessing the model predictions as well. Although the efficiency of this method is debatable, as discussed in the theoretical review, some initial evaluations in this study actually provided evidence to support that it would be useful. For this reason it was decided that the method would be included anyway. However, since the gain was expected to be relatively limited it was decided that the commonly advocated Conditional Random Field method was excessively complicated. Quite frankly, no adequate theoretical evidence was found to support that implementing this relatively complex technique would be worth the effort.

Instead of CRF a very simple and naive thresholding method was developed, as illustrated in Figure 3.17. The intuitive idea is to disregard all low confidence predictions, per pixel, since these are less likely to actually correspond to a true positive. More specifically all predictions below a predefined threshold would be considered as negative predictions. The hope is that such a postprocessing filter should remove all noisy predictions in order to create cleaner predictions that are easier to interpret. As the experiments will show this actually turned out to be true, a fact that is also illustrated in Figure 3.17. Furthermore, two slightly different kinds of postprocessing were used: hard and soft thresholding. The subtle difference is that the hard threshold outputs a binary prediction, whereas the soft method keeps the confidence score for all values above the threshold. The two methods are visually explained in Figure 3.17.

3.6 Experiments and model description in the conventional setting

According to the research presented in the preceding chapter the current state of the art model from brain tumour segmentation is a 3D U-NET with ResNet blocks. It should use the soft dice loss, Adam optimiser, drop out and either Instance or Group Normalisation. However, there is evidence to support that Group Normalisation is preferable in the specific case of Federated Learning since it is better at handling the non-IID case [37]. Lastly, as discussed in section 3.3, a unitary batch size should be used in favour of increased input dimension. Consequently the study has used a model that is made up of all these components, see Figure 3.18. This specific model backbone was used to win BRaTS 2018 [4] and finish fifth 2019 [5]. A series of initial rough grid searches showed that the model works well with default Adam parameters and a constant learning rate of 0.00001. No further tuning of these hyper parameters were deemed necessary.

The model description provided above consists of all clear cut conclusion that can be drawn from the theoretical review. Besides these there are some studies that suggest that the use of L2 weight regularisation, Focal Loss and ensembling would be beneficial as well. For completeness of argument it was decided to run experiments to investigate this also. The more experimental suggestions such as incorporating a VAE loss, a weighted loss, lesion priors or attention were disregarded due to a lack of empirical evidence to support the effort. These together with the immature

research areas of Curriculum Learning, Active Learning and Mixed Supervision are left for intriguing future work. Lastly, as already concluded in the theoretical review, no feasible solution for Transfer Learning is currently available for this application. Consequently all models were trained from scratch.

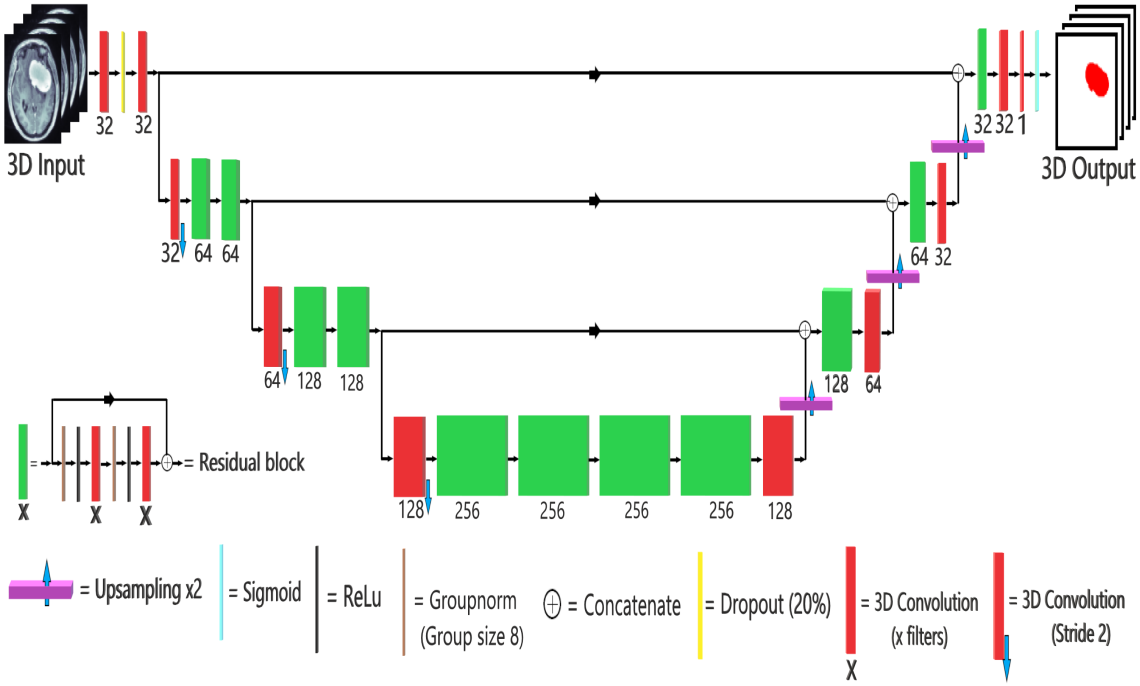


Figure 3.18: The study used this 3D ResNet U-Net model throughout for all experiment, both in the centralised and federated setting. It is the current state-of-the-art model architecture and this specific configuration won the BRaTS 2018 challenge [4] and finished fifth in 2019 [5]. The model takes a 3D brain volume as input and outputs an identical volume, with a prediction between 0 and 1 for each pixel that represented the likelihood that said pixel is a tumour. The model was trained using the default Adam optimiser, 0.00001 learning rate without decay and a soft dice loss. Experiments with a Focal loss and L2 weight loss on the kernels were also investigated.

All in all 28 different model configurations were examined in order to analyse how the input size, level of augmentation, Focal Loss, L2 weight regularisation, different postprocessing and ensembling impact performance in the conventional setting. A detailed breakdown of all experiments that were carried out is provided in Table 3.2. All models were evaluated on the validation set using the median, 25th and 75th percentile dice score to gain more detailed insight on how they perform for different subsets of the data. In order to protect the integrity of the blind test set no model was ever evaluated on this data until after all experiments had been carried out, including the ones in the federated setting.

Table 3.2: Several experiments were conducted in the conventional setting to investigate how different factors impact performance. A detailed breakdown of these are presented in this Table. All experiments use the model in Figure 3.18 and an explanation of the used augmentation levels is given in Table 3.1.

Model input size		
96	144	176

(a) Series to determine how input size impacts performance. No augmentation, L2 weight regularisation, Focal Loss, postprocessing or ensembling was used.

Level of Augmentation			
None	Minor	Moderate	Aggressive

(b) Series to determine how Augmentation impacts performance. All experiments were repeated for all three input sizes but no L2 weight regularisation, Focal Loss, postprocessing or ensembling was used.

Model L2 regularisation parameter μ								
0	0.001	0.01	0.1	1	10	100	1000	10000

(c) Series to determine how L2 weight regularisation impacts performance. The input size was consistently 176, the level of augmentation was 'Aggressive', no Focal Loss, postprocessing or ensembling was used.

Model Focal Loss parameter γ							
0	0.2	0.5	0.8	1.3	2	3	5

(d) Series to determine how Focal Loss impacts performance. The input size was consistently 176, the level of augmentation was 'Aggressive', no L2 weight regularisation, postprocessing or ensembling was used.

Postprocessing confidence threshold								
0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9

(e) Series to determine how postprocessing impacts performance. All experiments were repeated using both the hard and soft thresholding technique. The analysis was only performed on the best model, as determined from the initial four experiments described above. No ensembling was used.

Top-X model ensemble				
2	3	5	7	10

(f) Series to determine how ensembling impacts performance. The analysis used the Top-X best models, as determined from the initial four experiments described above. In addition all final ensemble predictions were postprocessed using the best such method, as determined from the experiment described above.

After the extensive experiments described in Table 3.2 were completed two model

configurations were extracted based on the results. The first one was the best stand alone model, excluding any postprocessing. The second was the best model in general, including ensembling and postprocessing. These two, and no others, were then evaluated on the blind test data. More extensive analysis of the stand-alone model were then conducted. Firstly it was evaluated on the 540 healthy patients to find the model's false positive rate. It was also examined how its predictive performance correlated to the tumour sub type, the source of the data and the tumour volume. Lastly it was examined how well calibrated the model is in the sense of how the predicted confidence score correlates to how often the prediction is actually correct. This is useful in order to understand how trustworthy the model is and if the confidence level actually corresponds to something tangible. Ideally we want a model that does not only predict correctly, but also with some assertion of how likely it is that the prediction is true. Finally the model performance was analysed qualitatively by visual scrutiny in an attempt to better understand when and where it fails.

From the description of this detailed performance analysis it is important to note that it was conducted on the stand-alone model rather than on the one which included ensembling and postprocessing, although the latter turned out to be slightly superior. The reason for this is because the experiments in the federated setting, which will be described in the sections to come, did not include any ensembling or postprocessing for simplicity. For the sake of comparison it is thus more fair to only consider the stand-alone model in the conventional setting as well.

3.7 Federated Model and experiments

This section handles the second research question concerning the feasibility of Federated Learning. It starts of by a description of the system design and the underlying assumptions, before moving on to the research design and a detailed breakdown of all conducted experiments.

3.7.1 System design and assumptions

As discussed in the theoretical review there are a number of practical assumptions and system design choices that must be decided on in order to create the federation. Concerning system design it is primarily two factors that must be examined, discussed in section 2.7. Firstly the topology can be either a star or completely decentralised using for example Blockchain technologies. This is discussed in depth in section 2.7.1. However, the latter is much more complex and no currently feasible option exist for large scale FL. For this reason this study decided to use the star topology, in line with a majority of previous studies. The second and related design choice concerns system management. As discussed in section 2.7.2 the federation can be either synchronous or asynchronous. However, the gain of using the latter is primarily if a very large number of clients is involved which is not the case in a hospital application. For this reason the decision was to only consider synchronous algorithms, again in line with the majority proposal. Furthermore, a star topology

with a central server naturally facilitates a synchronous approach.

These design choices comes with several practical difficulties due to the fact that clients are autonomous, as discussed in section 2.4.2. This lack of complete control for the central server implies that the synchronisation algorithms must be able to cope with unavailable clients, stragglers and stochastic drop outs. However, compared to if clients are mobile this issue is mitigated in the hospital application case since all clients can be assumed to have a stable and constant infrastructure. Practically this translates to that hospitals have uninterrupted access to power and high bandwidth networking, which are deemed to be a relatively reasonable assumption. Furthermore, the synchronisation issues and costs are naturally proportional to the number of clients. Again, this implies that the current application is less impacted. Motivated by all these arguments it was decided to limit the study to only consider the idealised scenario where these issue are non-existent.

Another implication of the assumption that hospitals have adequate and stable infrastructure is that the commonly researched FL issues concerning communication and computational overhead can be disregarded. Again, these issues are also less severe due to the fact that the number of clients is relatively small. See section 2.4.2 for a discussion on these limitations. Practically this implies that all proposed algorithms designed to limit communication, such as the praised Deep Gradient Compression discussed in section 2.6.2, can be disregarded. Note that the models in this study are only 70 MB, which is not an issue for any decent internet connection.

The perhaps most daunting practical difficulty in FL is the fact that client autonomy implies that no assumption on the local data quality can be made, as discussed in section 2.6.4. In contrast to the above discarded issues, this fact can not be overlooked in the hospital case. See section 2.5 for a detailed discussion on this. Consequently this means that any real world implementation must be robust to arbitrarily low data quality at select clients. However, this is beyond the scope of this study and is left for future work.

Finally, a word on privacy. This is really a part of the study's results, discussed and synthesised in section 4.3.2, but it is worth briefly mentioning the conclusions here to set the scene. In short the study follows the majority decision in previous work and assumes that all participants are both honest-but-curious and non-colluding. Importantly this implies that a (partially) *trusted aggregator* exists, which in the star topology translates to the central server. Furthermore it does not assume secure channels. A consequence of these assumptions is that although medical AI in general requires significant security guarantees, it is excessive to consider the most extreme vulnerabilities. In practice the assumed system has to be be protected against arbitrary attacks from third parties, such as the MITM attack illustrated Figure 2.1. However, it does not need absolute protection from internal participants since the assumption is that hospitals have at least some degree of mutual trust.

3.7.2 The training data was distributed both non-IID and IID across five virtual hospitals

After having established the setting for the federated scenario, the first order of business was to create the virtual hospitals that were going to collaborate in the federation. Some authors suggest that as many as 100 [85, 51, 23] or even 200 [123, 38] such clients should be used to properly emulate the real world conditions. However, others show that if suffices to use just three [40] or a handful [52] collaborators. Motivated by this, it was decided to use five clients for simplicity. This should be adequate for a proof of concept implementation.

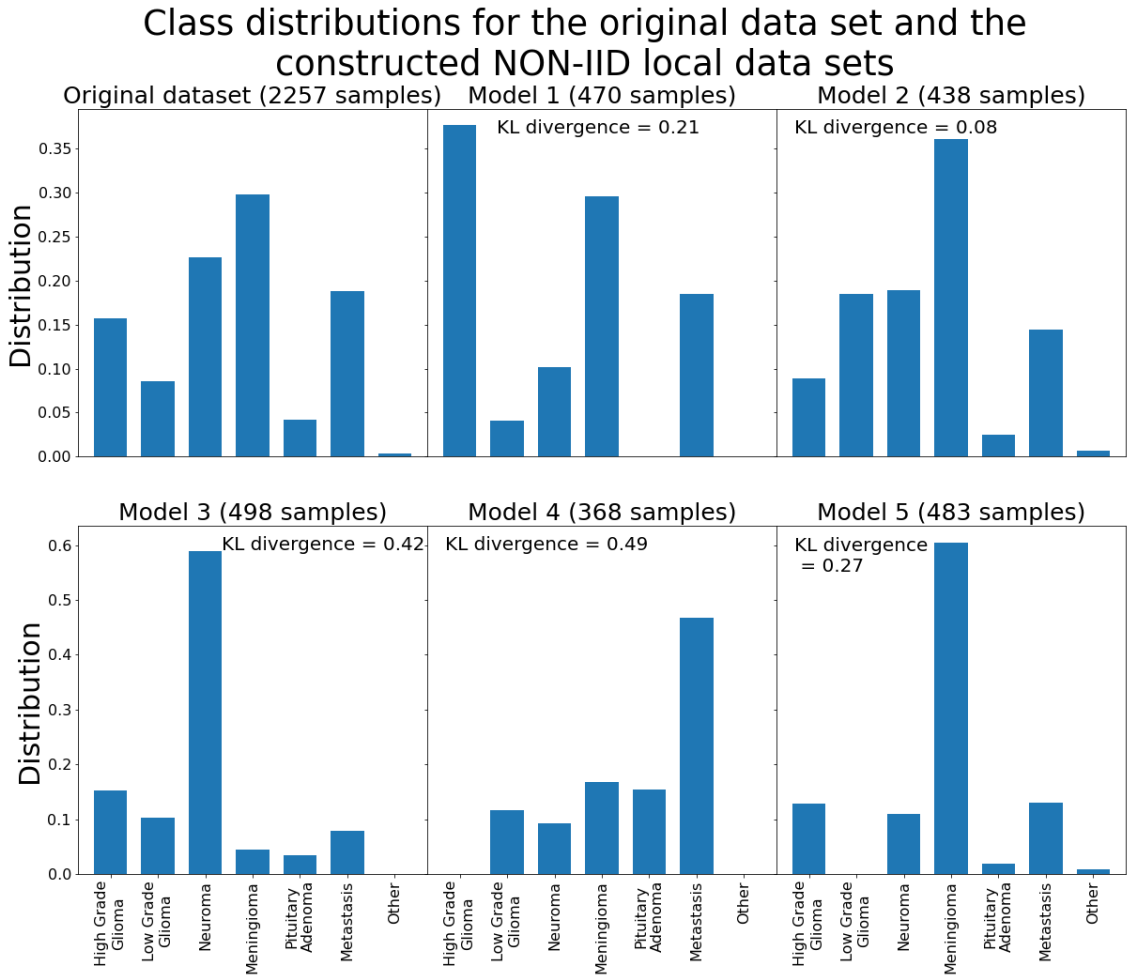


Figure 3.19: The class distributions of the original full training set and the five local data sets used in the federated learning experiments. The local data sets are by design NON-IID to mirror real world conditions. The Kullback–Leibler divergence is relative to the original data set and it shows that model 2 has the most similar distribution and that model 4 is the least similar. It is apparent from the histograms that the sites have very different distributions, however they are all approximately the same size. Future experiments should analyse how the performance is impacted if the relative scales are not the same. Finally, it should be noted that the common validation and test data follow the same distribution as the original training data.

The next step was to split the training data across the five clients. In order to emulate the sought after real world non-IID case the data was distributed randomly, but stratified such that the local distributions would differ significantly. The resulting five data sets are illustrated in Figure 3.19. Note that all local sets are approximately the same size. This is obviously not the case in real world conditions, as already discussed in preceding sections, since hospitals differ by orders of magnitude in scale and number of patients. However, the data was distributed in this way in order to isolate the non-IID effect. It is left for intriguing future work to study how the inter-hospital scale differences would impact the federated performance.

Concerning validation one can choose to either have local validation sets [85] or to use a global validation set located at the server [85, 34]. The latter is a more practical scenario which according to [34] is also a reasonable assumption to make. Another argument for using the global version is that it allows for a more fair comparison to the conventional setting, since they both can use the exact same validation set - which is exactly what was finally decided. It is important to note that the validation data and the original training data is IID, which implies that the newly created local sets are non-IID in relation to the validation set as intended.

Finally, in order to investigate how the non-IID condition impacts performance it was decided to create a second set of local training data, but this time distributed IID.

3.7.3 The federated model is evaluated relative to the conventional model and the five local models

As discussed in the theoretical review the common approach when evaluating a federated model is to compare it to the conventional model. One possible evaluation metric is to use the algorithm's δ-accuracy loss as defined in equation 2.4. However, this is an absolute measure. Instead, the common approach is to evaluate the model in relative terms. Plainly this means that the federated model's performance is measured as a percentage of the equivalent centralised model. This metric provides a way to analyse how much performance is lost relative to a theoretical scenario where the data was gathered in the central, conventional way. This implies that it carries academic interest but it is less relevant for real world conditions. As a complement to this metric it is thus common to compare the federated model's performance to models that were trained at each hospital on their own. The latter presents the real world best practice if no federation is used. This second metric thus tells us how much the federated collaboration would have improved performance in real life.

Motivated by this discussion the experimental approach used in this study was thus the following. Pick a high performing model configuration from the conventional setting and record its validation performance. Then train five local models on their own local data only. Evaluate them on the same common validation set and record their respective performance. This was repeated for both the local IID and non-IID case. These performances serve as a baseline for all federated experiments described

in the sections to come. In the conventional setting all models were evaluated in terms of three metrics - the median, the 25th and 75th percentile dice score - for increased detail. However, for increased readability all federated experiments will only be evaluated in terms of the mean dice score. This goes for the baseline stand-alone local models as well.

Having established a baseline it is time to move to the federated models. It was decided to only use the non-IID data for these experiments, since that is the most relevant case. It is left for future work to repeat all federated experiments on the IID data as well in order to fully appreciate how the non-IID condition impacts performance. This study only partially answers this question by being able to compare whether a federated model trained on non-IID data performs better than local, stand-alone models under the IID assumption. Importantly, if this is found to be true that means that federated learning is superior to the local stand-alone case regardless if the IID condition holds or not.

3.7.4 Federated experiments include using FedAvg, FedProx and a novel technique termed adaptive momentum

The outcome of the former subsection is that we have a baseline performance for both the local and conventional setting. The only thing left is to start training the federated models and evaluate them relative to these baselines in order to find the best configuration. Again, only the non-IID local data sets in Figure 3.19 are used for this purpose. The same model configuration that was used to train the baseline local models is used throughout all these federated experiments.

The experiments all relate to different model aggregation techniques, since this is the only thing that is new in the federated setting. From the theoretical review it is clear that the current benchmark algorithm is the FedAvg, see equation 2.5. It is the most commonly used and presents reasonable performance, which is why it will also serve as a benchmark in this study. As discussed in the theoretical review the convergence of FedAvg naturally depends on the number of local epochs during each round of federation. The first experiment to be carried out was thus decided to investigate this phenomena. Furthermore the previous studies suggest that FedProx is the current state-of-the-art algorithm, see equation 2.6, which motivates an examination of it as well in this study. More precisely it was decided to examine how the performance depended on the μ parameter, which determines the extent of the proximal loss.

Furthermore, several authors report that the aggregation improves if momentum is used, see equation 2.7. The amount of momentum is governed by the η parameter. However, all previous studies have used a constant η throughout the entire training process. Intuitively this seems sub optimal for two reasons. Firstly, the choice of η is completely arbitrary which might lead to a sub optimal choice. Secondly and more importantly, since momentum is only relevant if the previous generation model was relatively well behaving this implies that early stages of training should use low amounts of momentum whereas the last stages, in which the model has started to converge, should use more momentum. A consequence of this is that the ideal

η parameter should be decreasing, not constant. Motivated by this intuition this study proposed a novel technique termed adaptive momentum.

The underlying idea of adaptive momentum is to allow the model to choose the η parameter freely at each aggregation instance. Practically the implementation was as follows. Given the new local model updates, aggregate them with an initial choice of η. Evaluate the new joint model on the validation set and record the performance. Repeat this process for a number of different η and finally choose the one which caused the joint model to perform the best. For this study it was decided to try $\eta \in 0.2, 0.4, 0.6, 0.8, 1$. This implies that a minor computation overhead is incurred, since the model must be evaluate on the validation set five times as opposed to one. However, this is a negligible cost compared to the time each round of training takes. Besides, if the constant η momentum approach is used then it would be necessary to to a grid search to find the optimal η anyway - a process that is much more time consuming. It is worth noticing that the adaptive momentum is a strict extension of ordinary momentum, since the model is able to choose a constant η level. This implies that it is also a strict extension of FedAvg since it includes the option of choosing $\eta = 1$. This guarantees that the novel method performs at least as good as both previous techniques.

3.7.5 Differential Privacy merits an empirical examination concerning how it impacts convergence

Besides choosing an adequate model aggregation technique it is also necessary to consider the privacy constraints in the federated setting. Based on the extensive discussion on privacy in the theoretical review it is evident that any Deep Learning model that it trained on sensitive data must employ some mechanism to protect the system from model leakage. Furthermore it was found that the best technique to achieve this was to make the model differentially private using the Gaussian Mechanism. Importantly, this conclusion is independent on the training setting and includes the assumed honest-but-curious scenario. In fact, as long as there exists even the slightest chance that the model could be apprehended by an outsider party it must be trained using differential privacy. Consequently this implies that it is true for both the conventional and the federated setting.

To conclude, differential privacy is the only technique that provides protection without having to worry about who gets their hands on said model. In theory the model could be freely distributed to any arbitrary attacker without risking the integrity of the patients, as long as the $\epsilon - \delta$ guarantees are adequate. The higher the risk of model apprehension, the higher the required guarantees. This implies that the federated setting requires more DP than the conventional - but neither case can do without it. However, since excessive DP levels unavoidably leads to model deterioration this in practice implies that a federated model would be useless, due to the inherent vulnerability of a decentralised setting. Luckily, there are several complementary privacy mechanisms that mitigate the risk that an attacker can even apprehend the model. The premier such methods are Homomorphic Encryption, channel encryption and secure secrete sharing techniques. By adding these addi-

tional levels of protection the overall risk is decreased, which implies that the $\epsilon - \delta$ guarantees can be loosened. As discussed in section 2.10.1 no consensus exist on what exact $\epsilon - \delta$ guarantees are sufficient in general. Furthermore, no one has attempted to quantitatively determine to what extent the guarantees can be lowered as a function of what additional layers are added. This is also beyond the scope of this study and left for future work. However, the qualitative conclusion still stands: DP is necessary in federated learning, at some level or another.

The second best to knowing what $\epsilon - \delta$ guarantees are theoretically sufficient is to turn the question around and instead examine what levels are acceptable in terms of model performance. All that can be done is then to implement that maximum amount of DP that still outputs adequate performance. In practice this translates to finding the lowest $\epsilon - \delta$ that leads to convergence. If this level is then deemed to be excessive one can keep adding additional layers of protection until it becomes adequate in a privacy sense. Since DP is the only mechanism that has an impact on model performance it is the only one that requires to be empirically examined. The others are left as theoretical constructs and are not considered further in this implementation.

Lastly, as discussed in section 2.10.2 the Gaussian Mechanism can in principle be added either to each local gradient update or to the final model update. This study decided to do the latter. This provides an even stronger privacy guarantee since it makes the attacker unable to identify if an entire hospital was or was not part of the training. As discussed in section 2.10.2 this protection naturally implies that all patients at said hospital are also protected. Following the proposal by [23] the model updates were bounded by the median norm in the aggregation. The noise was added locally, in line with the (partially) *trusted aggregator* assumption used in this study.

3.7.6 Description of the experimental series

To summaries the discussions in the preceding sections it was found that four concepts need to be empirically examined. An illustration of the system design that was used to perform this analysis is presented in Figure 3.20. A breakdown of all the conducted experiments is provided in Table 3.3. All in all 20 global models were trained. The remainder of this section is dedicated to explain and motivate these experiments.

Firstly, the study intended to investigate how the number of local epochs in FedAvg impact performance. It was decided to consider the range of 1-100 local epochs, where a unitary value actually corresponds to the obsolete FedSGD model. See Table 3.3a for a detailed series description. The rational behind the experiment is that excessive numbers of local epochs might lead to that the local models diverge relative to each other which would hinder global convergence. However on the other hand it is preferred to use the largest acceptable number of local epochs, since that requires the least overhead in terms of communication and synchronisation, which as discussed in section 2.4.2 is a significant issue in practice.

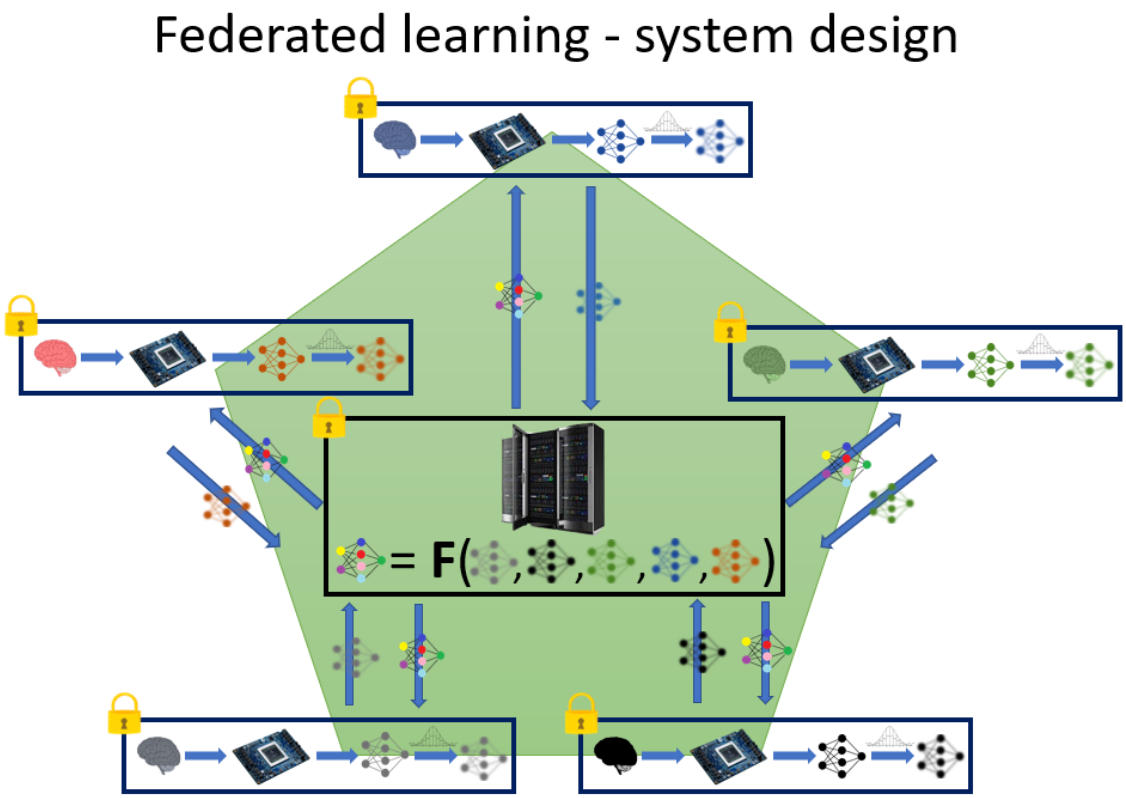


Figure 3.20: An illustration of the federated learning scenario under consideration in this study. Five local clients are used, each with its own unique data set as presented in Figure 3.19. All clients and the server are complete black box, meaning that the only thing that they share is the model updates. Each client trains a local model on its own data set, adds noise to make the model update differentially private and then shares the update with the server. The server aggregates the local updates and redistributes the new global model. This process is repeated for a predefined number of rounds. The different experiments involved FedAvg, FedProx, a novel adaptive momentum and differing levels of DP. See Table 3.3 for a detailed breakdown of the conducted examination.

The second experiment was to examine how different levels of proximal loss, measured in terms of the regularisation parameter μ, impacts FedProx performance. Furthermore, the study would reveal if FedProx is indeed superior to FedAvg as suggested by previous studies. The rational here is that if μ is too high then global convergence will never be achieved, since local updates are not allowed to nudge it towards the optima. If the experiments suggests that very low levels are ideal this implies that FedProx is not actually an improvement, since $\mu = 0$ corresponds to ordinary FedAvg. Motivated by that the inventors of FedProx [60] examined $\mu \in 0.001, 0.01, 0.1, 1$ but found that a majority of tests showed that the maximum $\mu = 1$ value was ideal it was decided in this study to extend this range to also include $\mu \in 10, 100, 1000, 10000$ for completeness of argument. See Table 3.3b for a detailed series description.

Table 3.3: Several experiments were conducted in the federated setting to investigate how different factors impact performance. A detailed breakdown of these are presented in this Table. All experiments use the model in Figure 3.18 as the backbone. Based on the results from the experiments in the conventional setting a high performing model configuration was decided and used throughout all federated experiments, including the local baseline models discussed in section 3.7.3. All experiments used the non-IID local distribution illustrated in Figure 3.19. Since the novel adaptive momentum is a strict extension of both FedAvg and ordinary momentum it was used throughout all experiments.

Number of local epochs					
1	5	10	20	50	100

(a) Series to determine how the number of local epochs during each round of federation impacts FedAvg performance. No FedProx or DP was used ($\mu = 0, \epsilon = \infty$).

FedProx regularisation parameter μ								
0	0.001	0.01	0.1	1	10	100	1000	10000

(b) Series to determine how different regularisation levels of FedProx impacts performance. Based on the results from the FedAvg series it was decided to use 50 local epochs during each FL round. No DP was used ($\epsilon = \infty$).

Degree of ϵ-DP privacy						
1	5	10	20	50	100	∞

(c) Series to determine how different levels of Differential Privacy impacts performance. $\delta = 0.01$-DP was used throughout all experiments. Based on the results from the FedAvg series it was decided to use 50 local epochs during each FL round. No FedProx was used ($\mu = 0$).

Thirdly, the study intended to investigate what levels of $\epsilon - \delta$-DP still leads to acceptable performance losses. The consensus rule of thumb is to set $\delta << \frac{1}{N}$ [23, 55], where N in this case is five. For this reason it was decided to set $\delta = 0.01$ throughout all experiments and only examine the effect of the more important parameter ϵ. One paper suggests that acceptable ϵ levels are 0.1-3 [145], another claims that $\epsilon \sim 1$ correlates to "very strong privacy" [55] and yet another study allows ϵ to be as high as 8 [143]. However, these proposals are when DP is used on its own without additional layers of protection. For this reason it was decided to extend the examined range for completeness of argument. See Table 3.3c for a detailed series description.

The last remaining question is how the novel adaptive momentum impacts performance. However, since it is a strict extension of both FedAvg and ordinary momentum is was decided to use it throughout all other experiments. Instead, the analysis of the method will be conducted in hindsight by recording which η was indeed found to be ideal at each round of aggregation. If it is found that it is indeed non-constant this proves that the novel model is superior to ordinary momentum.

If it is further found that it is not constantly $\eta = 1$, this proves that momentum in general is superior to bare bone FedAvg. Lastly, if it is found that η tends to decrease as training progress this supports the motivation behind the introduction of adaptive momentum, namely that momentum is mostly useful in the last stages of federation.

3.8 Software and Hardware setup

Before turning to present and discuss the results it is worth taking a few moments to explain what the computational environment looked like throughout the study. Firstly, concerning the software side of things. The entire project was conducted using Python, with an emphasise on the Deep Learning framework Keras which in turn used a Tensorflow backend. The core code for the conventional model, illustrated in Figure 3.18, was taken from [157], which is a direct implementation of the model used in [4]. This code base was extensively modified to include all additional features, such as image processing, augmentation, Focal Loss, L2 regularisation and ensembling.

Concerning the federated setting there are a number of open source platforms to choose from. The major ones are PySyft [82], TensorFlow Federated [75], NVIDIA Clara [77], CrypTen [76] and FATE [36]. However, as discussed by [36] these are all in early stages of development, since the field is young, which means that they are both computationally inefficient and offer limited features. For example it was shown by the people behind PySyft that it is 46 times slower than PyTorch [82]. Furthermore, with the exception of NVIDIA Clara they are all only capable of managing a virtual setting. Since this study also only considers this virtualised environment this is not an issue per se, but it serves as a good exemplification of how immature the frameworks are. Motivated by these drawbacks it was decided to instead implement the entire federated environment on my own using Python and Keras. Due to the relatively limited required features this actually turned out to be relatively simple. The gain was that essentially no computational overhead was incurred, relative to the conventional setting, which proves that it was indeed the right decision.

Finally, concerning hardware. A already discussed the models under consideration in this study require substantial GPU computational power and especially a lot of memory. The latter is a consequence of that 3D MRI scans require 50 times more memory than what is common in the natural imaging domain [30]. Evidence of the fact that strong GPUs are required is that training a 3D U-NET took two days on a NVIDIA Tesla V100 according to [4] and half a day on a NVIDIA 1080 Ti according to [109]. Furthermore, it should be noted that these models were only trained on BRaTS, which is almost 10 times smaller than the data set used in this study. As a consequence training a single model in this study on even these relatively strong GPUs would take on the order of days. Since the study examined 48 different global models and 110 local models, excluding all preliminary attempts in order to fine tune the learning rate and Adam parameters, it is easy to see that this would not be feasible.

Thankfully this study was supported by Taiwan Computing Cloud (TWCC). This meant that it had virtually unlimited access to Taiwania 2 [158], the world's 21st highest performing supercomputer [159] with in excess of 170000 cores, close to 200 TB of RAM and a peak performance at over 15000 Teraflops/s [158, 159]. Each node consists of one NVIDIA Tesla V100 SXM2 GPU and four Xeon Gold 6154 18C 3GHz CPUs [158, 159]. In short, computational power was not a limiting factor in this study. The used approach was to create a virtual container, comprising one such node and 90 GB of RAM, and to train the model on this. It turned out to take between one and three days for the model to converge in this environment, depending on the specific model configuration. In order to speed up the process all the examined models were trained on one such container each, completely in parallel. This meant that a years worth of experiments could be completed in a matter of just weeks.

Finally, a word on inference times. Although the training process is tedious, the final model presents a significant clinical speed up relative to human diagnostics. The two previously mentioned benchmark studies showed that inference took 0.4 seconds on a NVIDIA Tesla V100 [4] and 1.5 seconds on NVIDIA 1080 Ti [109]. This saving means that it is well motivated to spend all that training time.

4

Results

This chapter is largely divided into three sections, each dedicated to one of the three posed research questions. Firstly all results from the experiments in the conventional, centralised setting is laid out. It then moves on to present the findings concerning the federated setting in order to evaluate its feasibility for inter-hospital diagnostics. The chapter is finally concluded with a synthesis of the mechanisms that are necessary to ensure adequate protection of patients' privacy.

4.1 Conventional Deep Learning

This section presents the results from the experiments described in Table 3.2. It consists of 28 different model configurations that examine six different factors. The goal of the series is to analyse how these different factors aid, or potentially deteriorate, the model performance in a conventional, centralised setting. All these findings boil down to one final, best practice model. This model is evaluated on the 607 test cases and on the 540 healthy patients in order to establish its performance and false positive rate. These scores symbolise whether or not it is feasible to get high performing Deep Learning for brain tumour diagnostics in a heterogeneous, mixed quality setting. The goal of the study was to achieve super human level performance. It could be argued that this goal is met when the dice score supersedes 0.85, since this is the inter-rater agreement [93, 61]. The test performance will thus finally be compared to this metric.

4.1.1 Increased image size aids performance

The first experiment, described in Table 3.2a, was conducted to evaluate how the input size impact performance. The result of the experiment is summarised in Table 4.1. Just as expected and in line with the results from previous studies the experiment clearly show a positive correlation between performance and image size. However, the results are far from from human level performance and the experiment does not suggest that this could ever be achieved by simply increasing the image size.

Table 4.1: The first experiment in the conventional setting was to analyse how the image input size effect performance. The results conclusively show that a larger image size increases performance significantly for all three metrics.

Dice score	Model input size		
	96	144	176
25th percentile	0.10	0.15	**0.19**
Median	0.41	0.49	**0.52**
75th percentile	0.67	0.75	**0.77**

4.1.2 Moderate levels of augmentation boosts performance

The second experiment, described in Table 3.2a, was conducted to evaluate if different levels of augmentation could increase performance. The result of the experiment is summarised in Table 4.2. Just as in the first experiment the results are conclusive and in line with theoretical prediction. It is shown that even limited levels of augmentation provide a major performance increase. This is likely a consequence of the fact that the data set is relatively small, which implies that overfitting will become problematic if excessive epochs are used. However, this issue is naturally mitigated by the online augmentation. It was indeed seen by observing the training progress (not shown) that the gap between training and validation dice was drastically reduced when augmentation was introduced, which is evidence of this statement.

Another insight from the results is that the marginal gain of augmentation is diminishing rapidly. Importantly this implies that it would likely not be possible to increase the final performance significantly more by simply adding even more augmentation. This is in line with what previous studies have also shown, namely that excessive augmentation is not beneficial.

Table 4.2: The second experiment in the conventional setting was to determine how different levels of augmentation impact performance. The experiment were carried out for all three input sizes (96, 144 and 176 respectively) and the results are conclusive. The experiments clearly show that augmentation helps to improve performance significantly, especially in terms of the 25th percentile dice scores. The three different levels of augmentation are explained in detail in section 3.4. However, the experiments show that the primary factor is whether augmentation is used or not - the exact level of augmentation is of secondary import. This might be because the so called 'Minor' level is already relatively extensive, meaning that the model keeps being fed unseen samples regardless of the level. Furthermore, the augmentation was shown to be the most beneficial if the input size is small. This suggests that if memory or computation constraints are present which forces down sampling, then it is necessary to focus on sophisticated augmentation methods.

Dice score	Level of Augmentation		
	None	Minor	Moderate
25th percentile	0.10	**0.28**	**0.28**
Median	0.41	**0.58**	**0.58**
75th percentile	0.67	0.79	**0.81**

(a) Augmentation boosts performance when the input image size is 96.

Dice score	Level of Augmentation		
	None	Minor	Moderate
25th percentile	0.15	**0.34**	0.29
Median	0.49	**0.60**	0.59
75th percentile	0.75	**0.81**	0.80

(b) Augmentation boosts performance when the input image size is 144.

Dice score	Level of Augmentation			
	None	Minor	Moderate	Aggressive
25th percentile	0.19	0.31	0.31	**0.32**
Median	0.52	0.61	**0.62**	0.61
75th percentile	0.77	0.82	**0.83**	0.80

(c) Augmentation boosts performance when the input image size is 176.

4.1.3 Focal Loss only improves the worst case performance

Previous studies had shown that Focal Loss was beneficial, which motivated the third experiment presented in Table 3.2d. However, the results from this analysis were not very significant. See Table 4.3 for a full break down. Both the median and 75th percentile dice scores were only increased by a few percent.

More interestingly, however, is that the 25th percentile performance was improved by a lot for a carefully tuned amount of focus. This is in line with the theoretical

intuition that the Focal Loss should force the model to be better at the most difficult cases. This implies that although the aggregate performance is not impacted significantly, the Focal Loss actually makes the worst case performance less severe. This is a very desired property for a medical application.

Table 4.3: The third experiment in the conventional setting was to examine if the incorporation of a Focal Loss would increase performance. Surprisingly, it was shown to have a very limited effect but at moderate focal levels the performance rose slightly. If excessive focal levels were applied then the model started to deteriorate. Note that results should be compared to the case $\gamma = 0$, which corresponds to a model with no focus.

Dice score	Model Focal Loss parameter γ							
	0	0.2	0.5	0.8	1.3	2	3	5
25th percentile	0.32	0.32	**0.4**	0.34	0.32	0.31	0.29	0.3
Median	0.61	0.61	0.63	**0.64**	0.59	0.57	0.56	0.52
75th percentile	0.80	**0.82**	**0.82**	0.80	0.77	0.75	0.73	0.70

4.1.4 L2 regularisation increases performance by 33%

Table 4.4 present the perhaps most surprising result in this study. It is the outcome of the experiment described in 3.2c which shows that adding L2 weight regularisation has a dramatic effect on model performance. It caused the median dice score to jump by a remarkable 33%. This is contrary to the results in most previous studies who have reported only minor, if at all, improvements. In fact, the best model with L2 regularisation has about the same 25th percentile performance as the median performance of the best non-L2 model.

Table 4.4: The fourth experiment in the conventional setting was to incorporate a L2 regularisation term that punished the model weight norm. This was shown to have a dramatic effect on performance and increased the median dice score by a remarkable 33% for the highest performing model. The usefulness of L2 regularisation is in line with the theoretical predictions since it reduces overfitting. However, as expected excessive regularisation with $\mu \geq 10$ causes the model to diverge and become useless.

Dice score	Model L2 regularisation parameter μ								
	0	0.001	0.01	0.1	1	10	100	1000	10000
25th percentile	0.32	0.38	0.48	0.57	**0.60**	0.31	0.01	0.01	0.00
Median	0.61	0.64	0.72	0.78	**0.81**	0.55	0.02	0.02	0.01
75th percentile	0.80	0.84	0.87	**0.90**	**0.90**	0.71	0.08	0.07	0.05

4.1.5 Post processing improves performance by disregarding low confidence predictions

The four experiments already discussed conclude the analysis of different model configurations. The only thing left to consider is how to use these stand-alone models intelligently. The first such step was to incorporate two different kinds of post processing, which resulted in a series of experiments described in Table 3.2e. The results from this analysis is presented in Figure 4.1 which shows that the technique was indeed beneficial in terms of all three metrics. The increase was not dramatic, but still noteworthy. This is in line with the referenced theoretical predictions. Furthermore, it turns out that the soft post processing technique was superior to the hard version. Additionally the soft method proved to be very robust to the selected threshold and actually outperformed the stand-alone model for all such values. This is reassuring since it mitigates the risks associated with choosing the wrong hyper parameter.

An important lesson from the experiment is that the conclusive results suggests that the model exhibits a relatively strong correlation between its stated confidence and the likelihood of it actually being correct. This would be a desired trait since it provides better interpretability, which slightly counters the commonly quoted arguments against Deep Learning as being a black box approach.

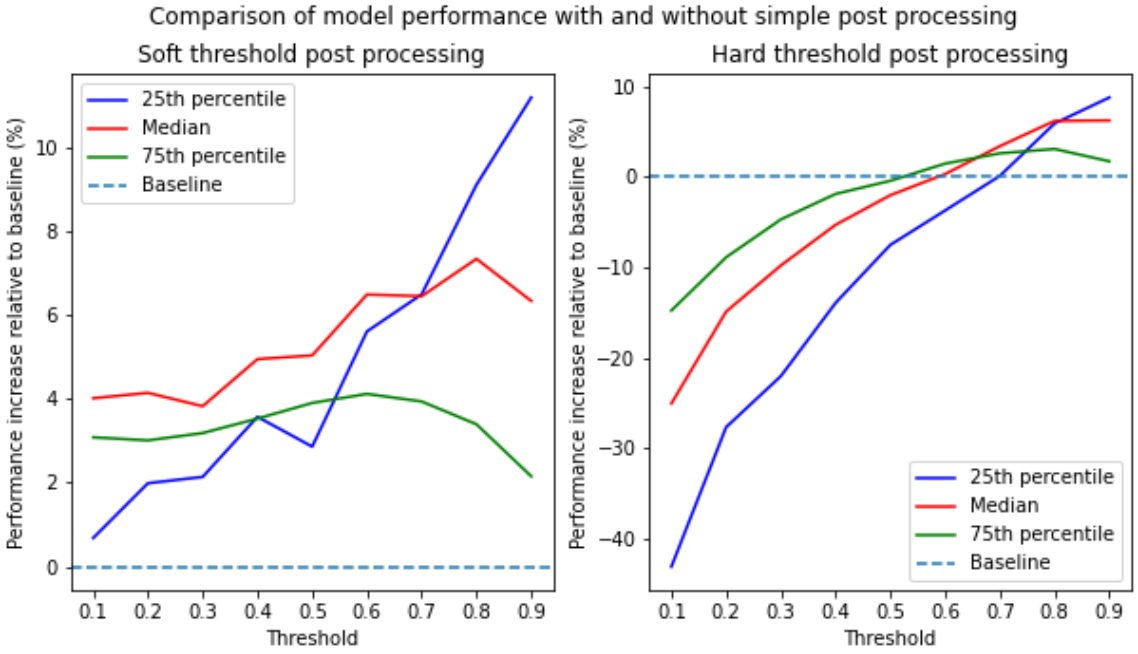


Figure 4.1: Post processing with soft thresholding outperforms the initial model for all examined thresholds. The hard thresholding method degrades performance if the threshold is set too low, but provides an improvement for high confidence thresholding. In general the soft method outperforms the hard. Interestingly the post processing is superior when only very high confidence predictions are considered, peaking at a 80% threshold. This suggests that whenever the model is correct it is also very confident.

4.1.6 Ensembling provides a modest improvement

As can be seen from Figure 4.2 it is indeed possible to improve the final model performance slightly by ensembling. These results are based on the experiments described in Table 3.2f. However, the improvement is not very significant for any of the three metrics. It is also shown that the ensembling only works for a set of the two or three best models, if any more are added then the performances goes down again.

It should be noted that the top three performing models achieve 0.81, 0.78 and 0.72 median dice - see Table 4.4 - whereas the top 4-10 models only perform between 0.64 and 0.61. The large performance drop from the third to the fourth model might be the reason behind why ensembling stops working at this point. The rational behind ensembling is that the models should be complementary, which they likely are, but this gain is overshadowed if some parties are significantly worse in absolute terms. It is possible that the ensemble performance would be further improved if all the models were identical in setup, but different in initialisation. This would ensure that they are both complementary and equally good. This is left for future work to analyse.

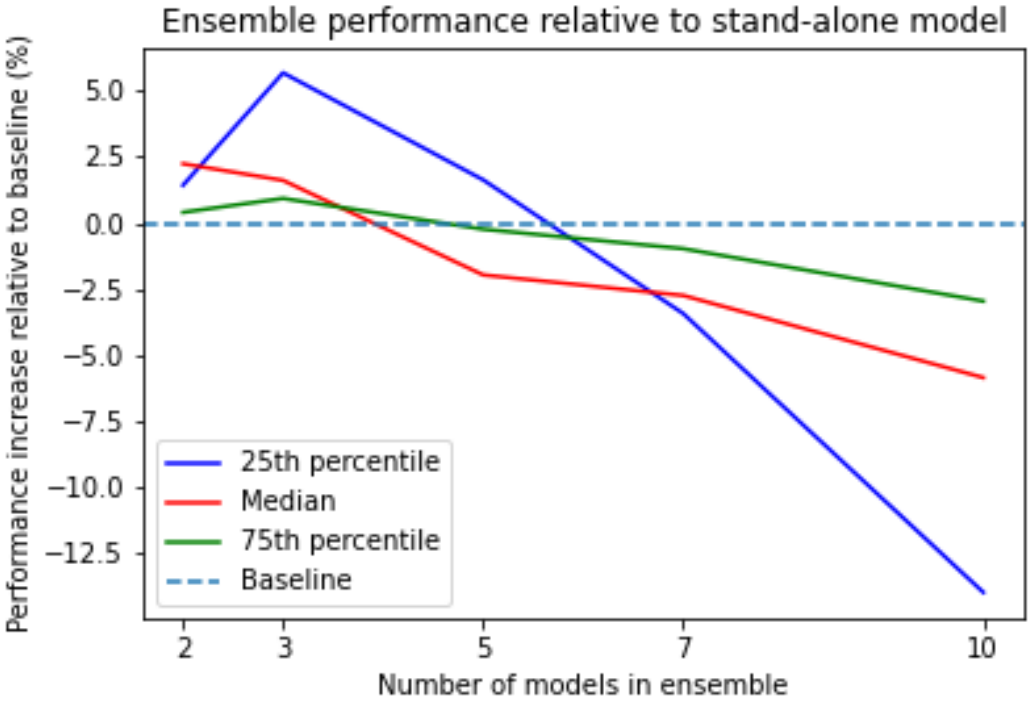


Figure 4.2: An ensemble of the top 2 or 3 stand-alone models manages to outperform the best stand-alone model in the conventional, centralised setting. However, the performance increase is relatively modest. Furthermore, adding more models to the ensemble degrades performance.

4.1.7 Summary: Efficiency analysis of the used tools to improve the conventional model

This section is provided for increased readability and is intended to summaries how the different tools have had an impact on the model performance. The aim is to understand which ones are the most efficient and possibly how they complement or substitute each other. See Figure 4.3 for a detailed breakdown.

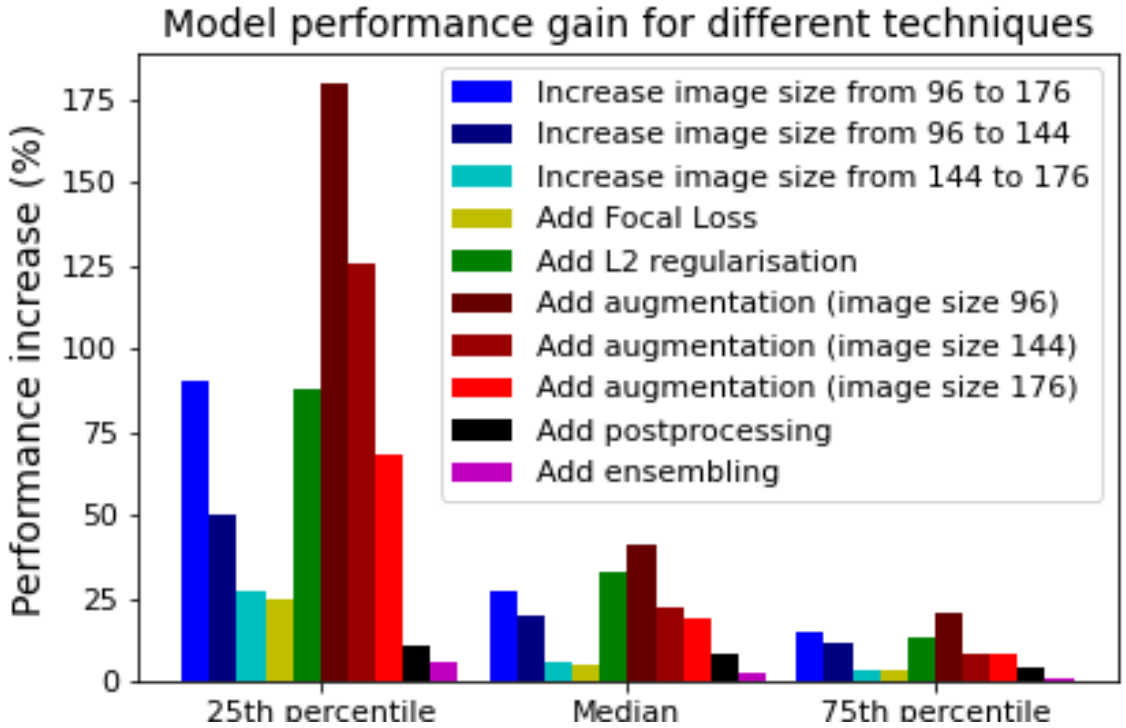


Figure 4.3: From the bar graphs it is clear that the most invaluable tool is to introduce augmentation. However, the gain from augmentation diminishes if the image size increases - which turns out to be the second most efficient tool. This has the effect that the image size becomes less relevant if sufficient augmentation is used. Adding L2 regularisation is shown to provide a big performance boost for all three metrics, whereas the Focal Loss only significantly improves performance on the hardest samples, illustrated by the lower percentile. Postprocessing and ensembling produce minor gains. In general the methods are most efficient at increasing the performance on the most difficult samples, as seen from the large increase in terms of the 25th percentile.

Interestingly it shows that augmentation has a diminishing return on large images, which has the effect that the initial large spread between model performance after training on large or small images respectively is lessened. If no augmentation is used the model trained on images of size 176 outperforms the 96 size equivalent by 90, 27 and 15% in terms of 25th, median and 75th percentile dice score respectively. After augmentation however, this difference is decreased to just 14%, 7% and 2% respectively. This suggests that augmentation and increased image size are two

efficient, but somewhat supplementary tools. Since increased image size requires more RAM, this suggests that if memory is a bottle neck the best thing to do is to focus on augmentation techniques.

4.1.8 The final model achieves super human performance, but the false positive rate is relatively high

Finally it is time to wrap up everything that has been learnt from the experiments in the conventional setting. Based on the already presented results it is concluded that the best model configuration uses an image size of 176, the most aggressive type of augmentation and L2 weight regularisation with $\mu = 1$. No experiments with both Focal Loss and L2 regularisation was performed, but the results suggests that this combination would lead to a slight additional improvement. This is left for future work to analyse.

The described final stand-alone model was lastly evaluates on the 607 blind test cases. Furthermore the results from the postprocessing and ensembling experiments suggest that a Top-3 ensemble together with a soft threshold at 80% confidence would be ideal. This ensemble version was also evaluated on the test set. The results are presented in Table 4.5. Just as the previous results suggest the ensemble version does indeed slightly outperform the stand-alone version. Importantly, the ensemble version manages to achieve super human performance. This serves as a strong indicator that it is indeed possible to create a clinically relevant Deep Learning CAD system for brain tumour diagnostic, even in a heterogeneous and mixed quality setting.

Table 4.5: Summary of the test performance of the final conventional, centralised model. It uses an image input size of 176, the most aggressive augmentation as well as L2 regularisation with $\mu = 1$. No Focal Loss was incorporated. The blind test set consists of 607 samples, IID across all seven tumour sub types and all six data sites. The stand-alone model is outperformed by the top-3 ensemble with soft postprocessing at 80% threshold. Importantly the ensemble version exhibits super human performance, which in this study is defined at 0.85 dice score. These results should be compared to the current leading 0.92 dice score on the related task BRaTS 2019 challenge [6]. It should be noted, however, that the latter represents a much less heterogeneous scenario.

Dice score	Stand-Alone	Top 3-Ensemble + Post Processing
25th percentile	0.58	**0.60**
Median	0.83	**0.87**
75th percentile	0.90	**0.92**

The final step of the study on the conventional setting involved analysing the false positive rate of the final model. For this reason the ensemble version discussed above was used to inference on the set of 540 healthy patients. The model correctly identifies 76.9% of patients as entirely healthy, in the sense that not a single pixel

was mistaken for tumourous. In other words, it incorrectly diagnoses 125 patients (23.1%) with brain tumours. This fallacy is discussed in the next chapter.

4.2 Federated Learning

This section presents the results from the experiments described in Table 3.3, with the exception of the analysis of Differential Privacy which is left for the succeeding section concerning results related to privacy. The aim is to unravel how different federated aggregation techniques work and if it is indeed possible to get adequate global performance using FL, despite the heterogeneous and non-IID scenario. Furthermore it will be investigated if the proposed novel adaptive momentum technique represents an improvement relative to the current state-of-the-art. The federated model will be compared to both the conventional analogue and the stand-alone, local models.

4.2.1 Collaboration is necessary independent of whether the IID assumption holds or not

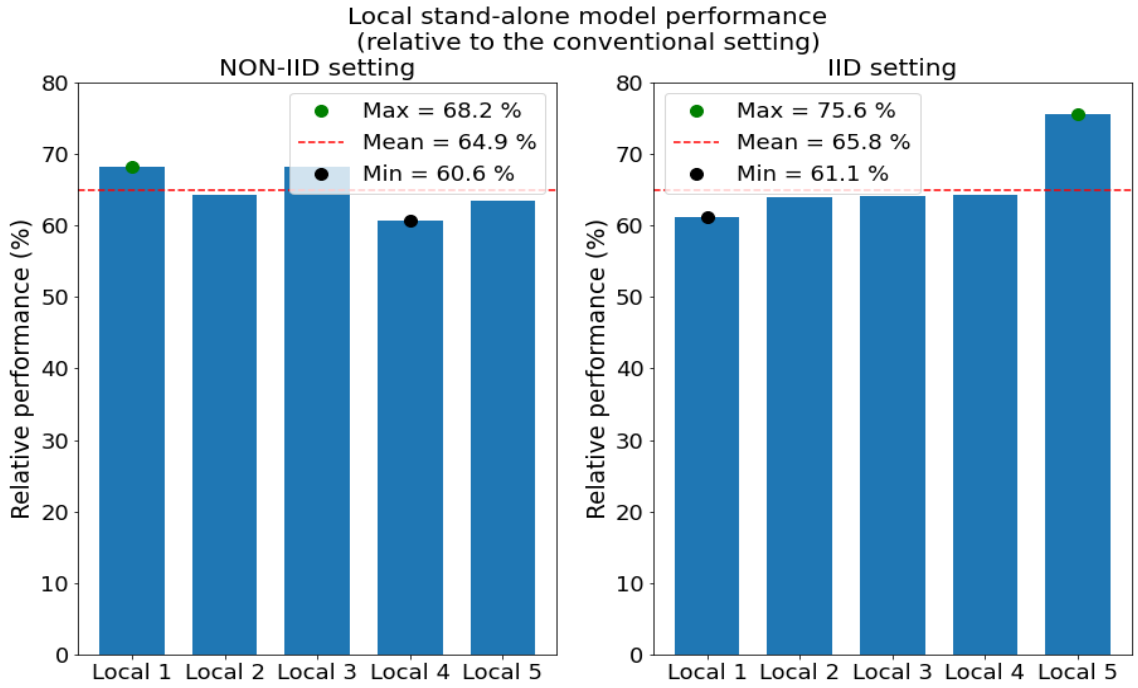


Figure 4.4: The baseline performance in the federated case is to train five local, individual models on their unique data. This was repeated on both a IID and a non-IID data split across the five clients. As seen from the two bar plots it is clear that the local models perform much worse than the conventional analogue, as expected. Importantly this is true regardless if the IID assumption holds. Although the average local performance is relatively similar in both scenarios, one of the local IID models perform significantly better than all the rest. In general the result suggest that the non-IID scenario is more problematic, as expected.

As discussed in section 3.7.3 the federated experiments was initialised by training local, independent models on each of the five local data sets. This was repeated for both the IID and the non-IID case for comparison and the results are visualised in Figure 4.4. It is clear that the local models are inadequate, which implies that some form of collaboration is necessary in order to get performance that is clinically relevant. Importantly, this statement holds regardless of whether the IID assumption holds or not. The latter is essentially empirical proof of a known fact from Deep Learning, namely that more data is always better. Since each local client only has a fifth of the conventional model's training set it is reasonable that their performance should be inferior.

Another insight from the two bar plots in Figure 4.4 is that the general local performance is slightly higher under the IID assumption. This is in line with theoretical predictions, since training on data that is IID in relation to the validation set should improve generalisation and hence performance. However, this also provides assurance that the performed non-IID split is indeed actually non-IID.

4.2.2 Federated Models under non-IID conditions

Based on the local analysis it is concluded that they perform much worse than the conventional setting, as expected. The best of those local models perform 68.2% of the conventional performance, under non-IID conditions. This section will analyse if the FedAvg or FedProx models can outperform this, hence proving that FL is a feasible solution.

4.2.2.1 FedAvg is robust to the number of local epochs

As can be seen from Figure 4.5 the global joint model performance does not show any strong dependence on how many local epochs are run. These results are from the experiments described in Table 3.3a. Surprisingly, the trend even suggests that running more local epochs before aggregation is better. This is positive since it means that the issue of communication and synchronisation overhead can be mitigated. Moreover, the results are significantly better in the federated setting than in the local, stand-alone setting. This provides strong evidence that Federated Learning is indeed a feasible solution for the application at hand. As it turns out the best model from this specific experiment, scoring 88.6% of the equivalent conventional model, was the best of all examined federated models in this study. This is thus the quantitative results to symbolise the feasibility.

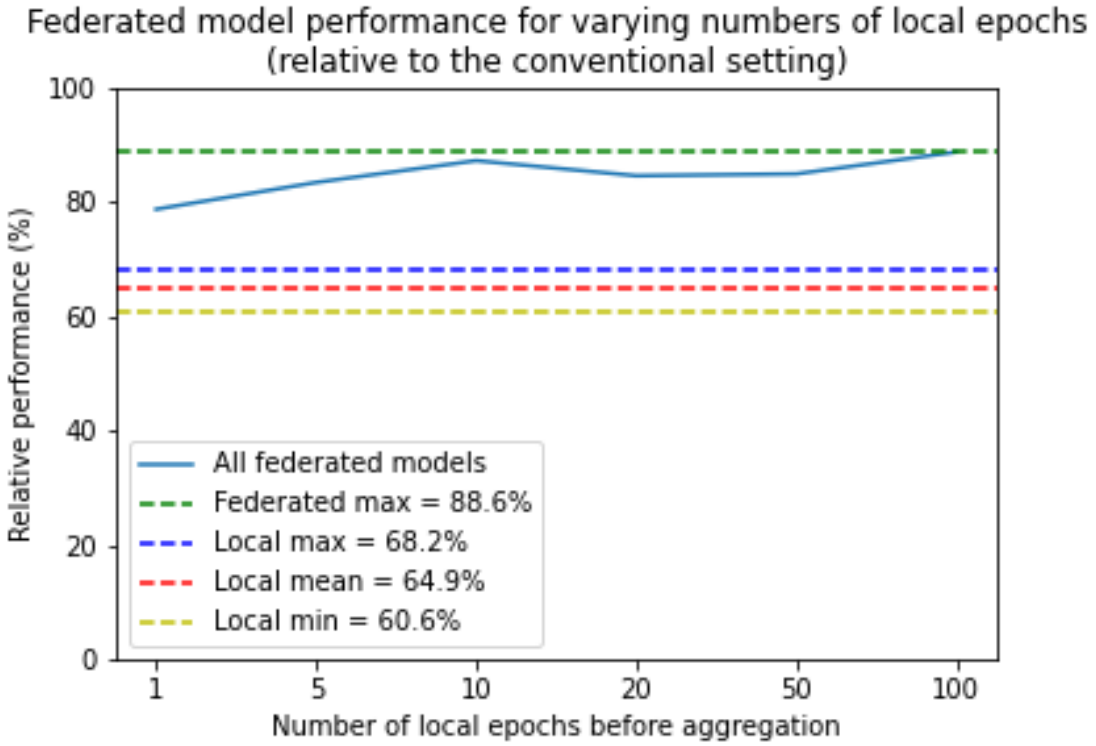


Figure 4.5: The first experiment in the federated setting was to analyse how the number of local epochs during each FL round impacts performance. The result suggest that the FL algorithm is relatively robust to this. Importantly, the experiment showed that a federated model can achieve 88.6% of the equivalent conventional model performance, even though the local data sets were non-IID. Furthermore, this is a 30% increase relative to the best individual, local model which proves that FL is better than attempting to train the models without collaboration.

4.2.2.2 The FedProx algorithm does not impact performance at all

The second federated experiment was to consider how FedProx might improve convergence, see Table 3.3b for a description of the analysis. However, contrary to all previous studies the FedProx did not impact performance at all. In fact, the results suggest that the convergence was completely independent of the FedProx parameter μ. This might be due to one of two reasons. Either the FedProx algorithm just does not work for this specific data set, or a sufficient span of possible μ values was not examined. However, since the examined span was significantly broader than the proposed range by the original authors this seems unlikely.

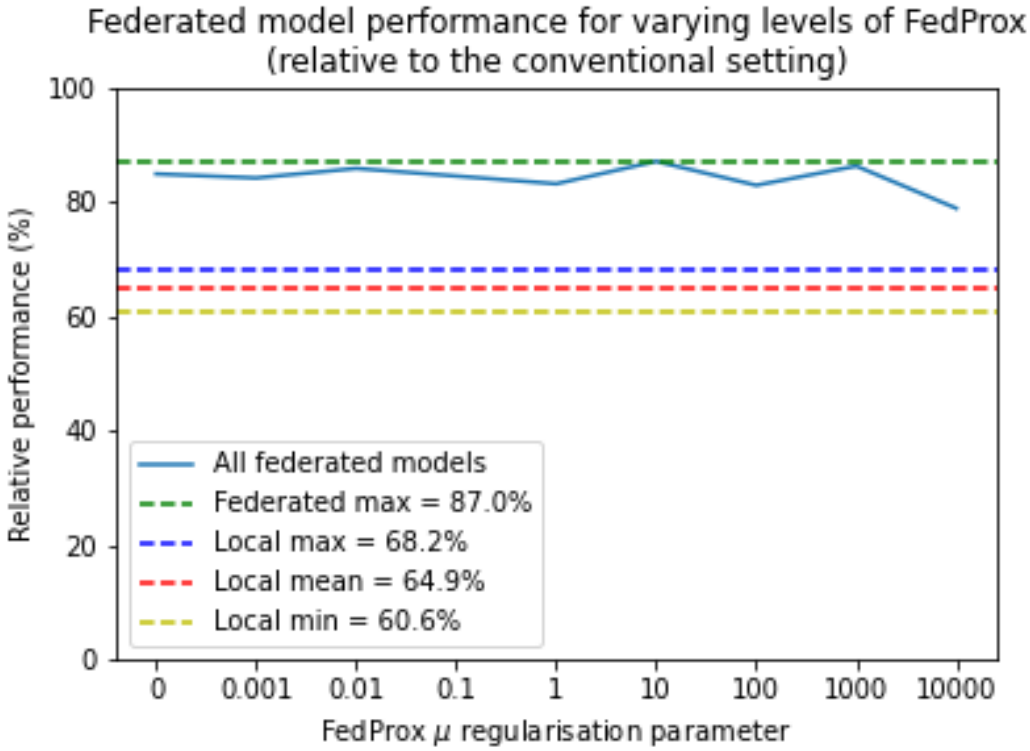


Figure 4.6: The second experiment in the federated setting was to analyse if FedProx outperforms FedAvg. However, contrary to the theoretical predictions no such evidence was found. In fact, the level of FedProx regularisation was not found to have any significant impact on performance at all.

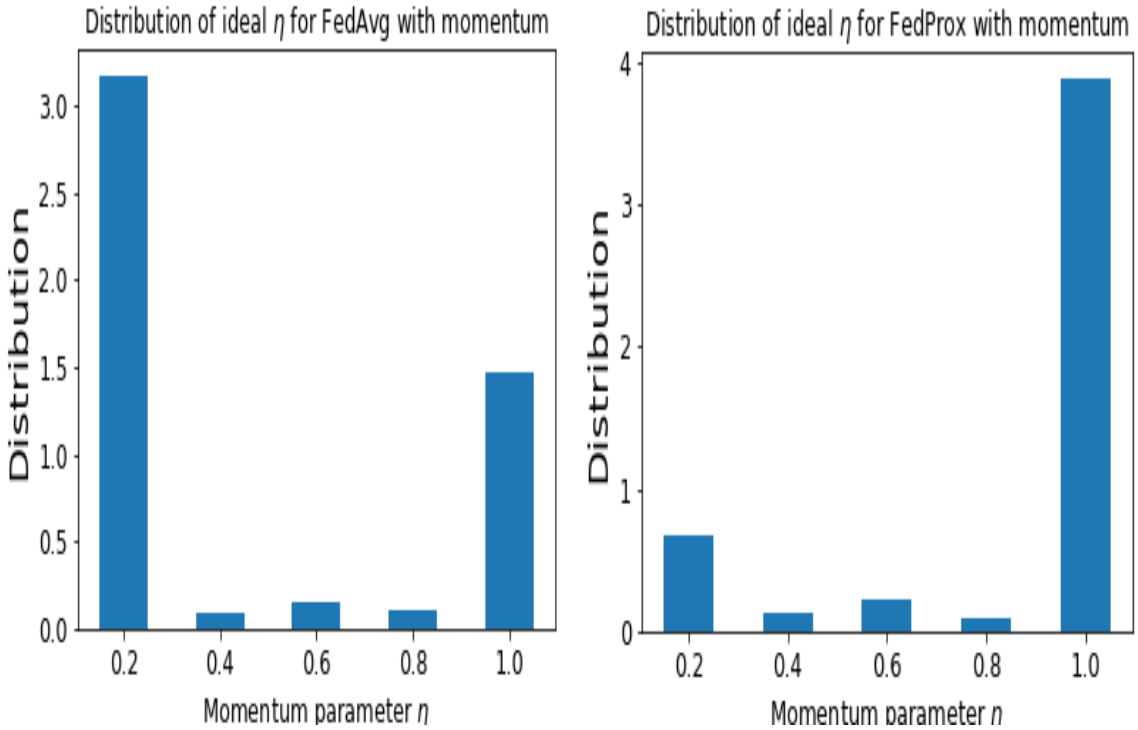
4.2.2.3 The proposed novel adaptive momentum algorithm is very promising and improves the current state-of-the-art

By design all federated models in this study were aggregated using the proposed novel algorithm termed adaptive momentum. It was motivated by the fact that ordinary momentum requires a constant, predefined level of momentum throughout all rounds of federation which intuitively seems excessively restrictive. The first hypothesise was thus that if the model was able to choose the level independently at each step, then the ideal level would be non-constant. It was further hypothesised that the level of momentum should increase as training progresses, since the previous generation model becomes more and more mature. See section 3.7.4 for a detailed discussion and motivation of the technique. Besides these hypotheses directly related to the novel technique, there is also a more general momentum hypothesis to consider. More specifically it is empirically proven that momentum is superior to bare bone FedAvg/FedProx if, at any stage during training, the model chooses to employ momentum.

To conclude, this section is dedicated to investigate the following three hypotheses. Firstly, is $\eta! = 1$ ever chosen by the model? If so, momentum in general is superior to FedAvg/FedProx. Secondly, is η not constant throughout training? If so, adaptive momentum is superior to momentum. Thirdly, does η tend to decrease during

training? If so, momentum is more beneficial in later stages of training. These three questions will be answered based on the results from all three experiments described in Table 3.3, including the one concerning DP.

The first two hypotheses are conclusively confirmed based on the results illustrated in Figure 4.7 and Figure 4.8. They show that regardless of whether FedAvg, FedProx or FedAvg with DP is used for aggregation it is always beneficial to employ momentum in at least some stages of training. However, this benefit is larger for FedAvg than for FedProx, as seen in Figure 4.7. One possible reason for this could be that the proximal loss in FedProx leads to smaller local model updates, which in turns implies that it is redundant to use momentum.



(a) Using FedAvg it is common that the ideal model aggregation uses momentum.

(b) Using FedProx it is much less common that momentum is beneficial.

Figure 4.7: All federated models were trained using an adaptive momentum technique. This meant that the level of momentum η used to combine the next generation model was chosen based on which value maximised the validation score at each round. Every time that $\eta < 1$ was chosen this signifies that momentum was indeed beneficial. As seen from the histograms, this was indeed the case for both FedAvg and FedProx. However, the FedProx model used much less momentum. An interpretation of this could be that the Proximal loss in FedProx leads to smaller model update changes, which means that momentum is redundant. Surprisingly, the ideal aggregation seems to be either with maximum or no momentum - rarely in between. An interpretation of this is that momentum is only beneficial if the model updates are degraded, in which case maximum momentum is preferred.

Furthermore, Figure 4.8 suggests that if the privacy constraint is higher then more is to be gained from employing momentum. This is also reasonable, since higher privacy in DP equates to more noise which in turn implies that local updates will be less reliable. Frankly, the risk that the local updates have deteriorated rather than improved during the latest round is higher when more noise is present.

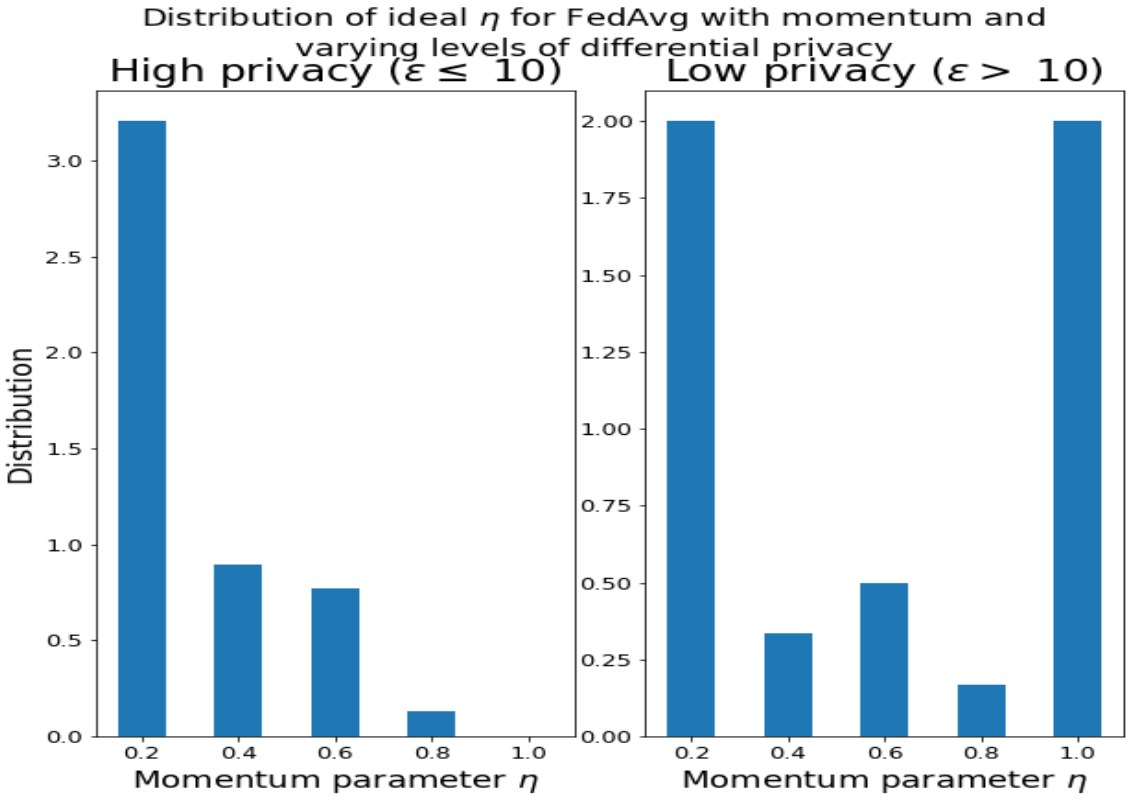
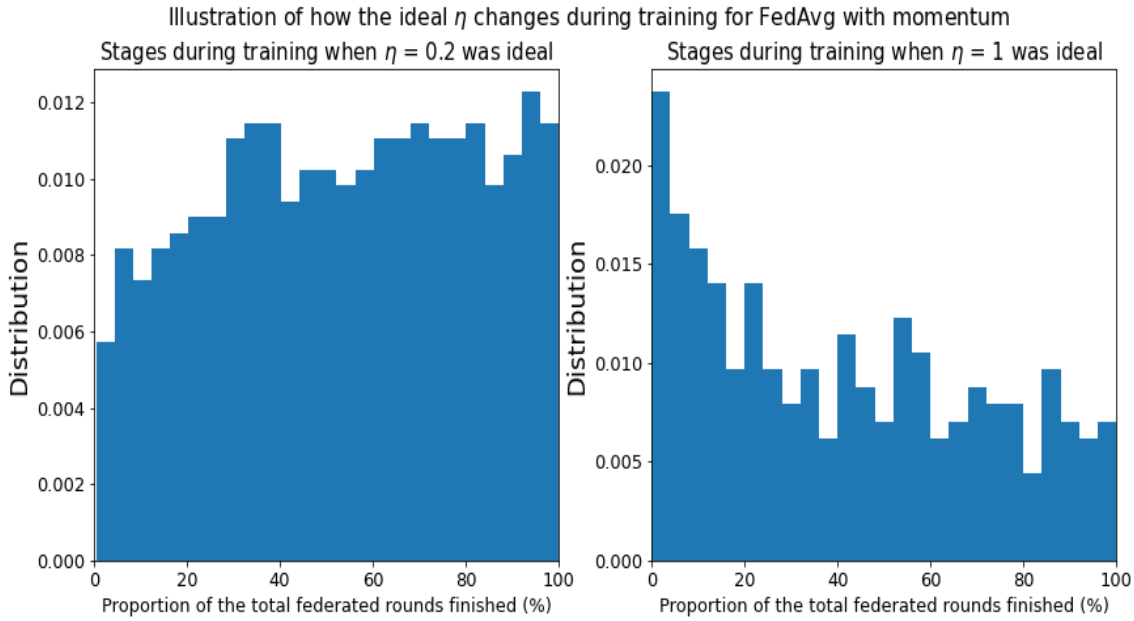
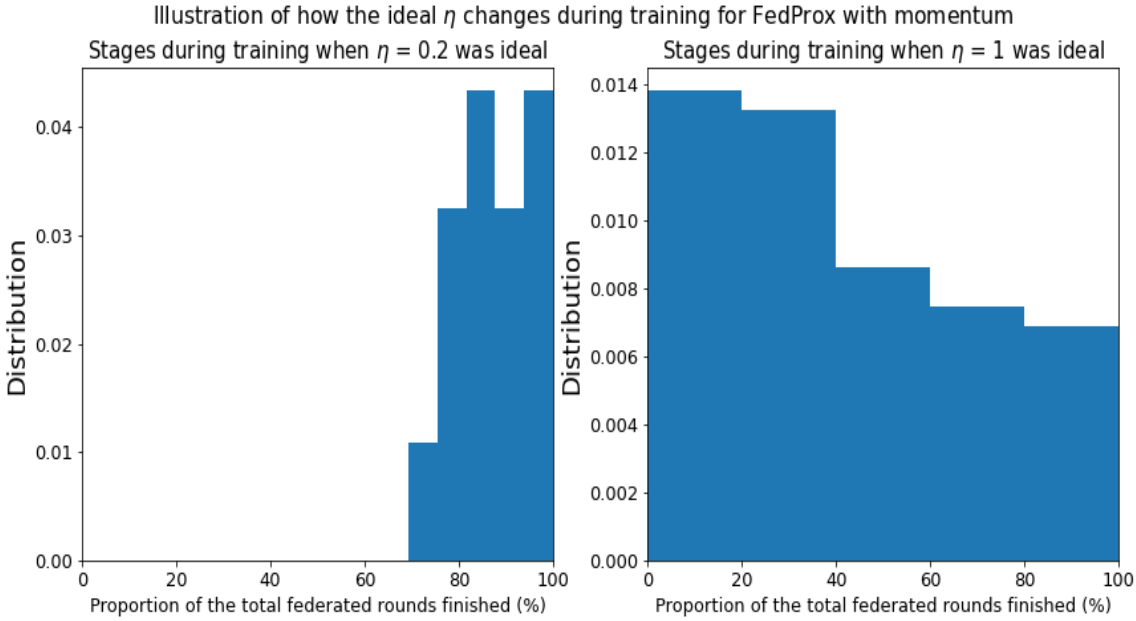


Figure 4.8: Illustration of how momentum impacts models with different levels of $(\epsilon - \delta)$-DP. Similar to Figure 4.7 it is clear that all models benefit from the adaptive momentum technique. Furthermore, the graph suggests that momentum is more beneficial for high privacy settings. Actually, it turns out that in this setting momentum was applied at every aggregation. The reason behind why the high privacy case uses more momentum than the low privacy could be because the latter contains significant noise, which means that the new model updates are more likely to be worse than the previous round in which case a low η would mitigate the global deterioration.

Lastly, the final hypothesis can also be verified by examination of the Figures in 4.9. It is clear that the general level of momentum tends to increase during training. This can be concluded because of the fact that $\eta = 0.2$ is more common in the end stages then in the begin, and conversely that $\eta = 1$ is more frequently occurring in the beginning.



(a) The amount of momentum continuously increases during training when using FedAvg, however it is present at all stages of training.



(b) Momentum is never employed during the initial 70% of federation rounds when using FedProx, but becomes increasingly common in the final stages.

Figure 4.9: The illustrated distributions reveal that momentum tends to increase during training, especially for FedProx. It is intuitive why momentum should increase, since it becomes more useful to retain parts of the old generation model if said model is more mature. As seen from the training progresses in Figure 5.1, 5.3, 5.5 and 5.4 the models tend to improve rapidly in the initial stages, which means that their is no use in keeping parts of the old model early on.

4.3 Protecting the federation

So far the study has shown that both the conventional and federated settings are capable of providing adequate performance. However, these experiments have not considered the privacy dimension. This last section is intended to wrap up the study by carefully analysing whether these conclusions still hold if this additional dimension is taken into consideration.

4.3.1 Federated differentially private models perform better than non-private local models

The study only considers one empirical aspect of protection, namely differential privacy. The experimental approach is discussed and described in Table 3.3c and the results are presented in Figure 4.10. The analysis shows that if excessive DP is imposed then the model becomes useless, as expected. However if the constraints are relaxed to the point where $\epsilon \geq 20$ is allowed, then the model is not impacted at all. Importantly this implies that a moderate level of DP can be employed at no cost at all. Furthermore, a relatively high level of DP with $\epsilon = 10$ comes at the cost of a moderate 11% performance reduction.

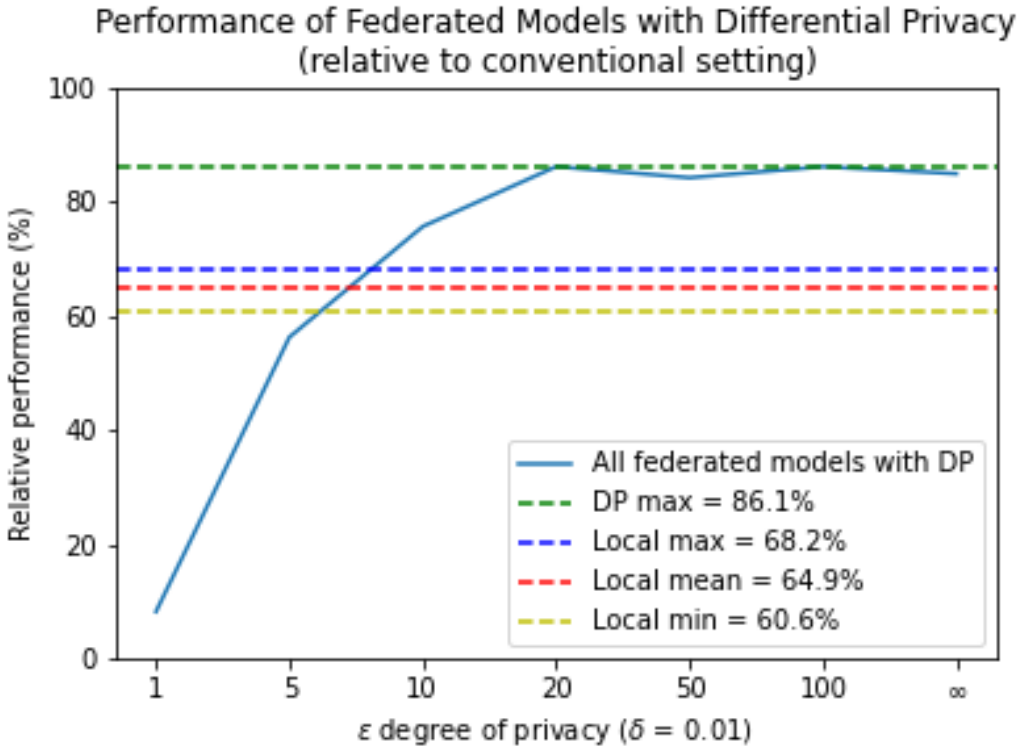


Figure 4.10: The final federated experiment analysed how different level of Differential Privacy impact the federated performance. All models are $\delta = 0.01$ DP, with differing ϵ. The experiment showed that as long as $\epsilon \geq 20$ performance is not impacted at all. Acceptable levels of performance seem to require that ϵ is not lower than 10 and if it is decreased to 8 then the federated model performs on par with the local stand-alone models, thus rendering it useless.

An important realisation is that the local stand-alone models do not provide any privacy guarantees at all to its local patients, since they were trained without DP. This leaves them vulnerable to attacks by outsider parties. The importance of this statement can be explained using the results in Figure 4.10. The experiment show that even a relatively well DP protected federation performs significantly better than the unprotected local models. In fact, the performance is superior to the best local model all the way down to $\epsilon \approx 8$. This means that federated learning can in fact improve both privacy and performance, relative to a non-collaboration setting. This statement provides strong empirical evidence that private FL is indeed a good option for inter-hospital AI systems.

4.3.2 Additional layers of protection are necessary to ensure adequate levels of privacy

The preceding discussion show that some levels of DP can be applied without incurring unacceptable performance loss. In fact, with $\epsilon = 10$ the model still performs 76% of the conventional model and with $\epsilon = 20$ the number is 88.6%. However, ideally this 11% performance cost should be avoided if possible. The question then becomes what level of DP is sufficient. As discussed in section 2.10.1 this is still an open question, but the general consensus is that ϵ should be somewhere in the single digit range. However, it is also discussed that higher ϵ levels can be accepted if additional layers of protection are included. Sadly, none of the referenced studies have provided any quantitative notion to this which makes it highly subjective. This section is intended to analyse the threat level of the application to untangle whether the ideal $\epsilon = 20$ level is indeed acceptable or not.

Firstly, it is necessary to discuss and argue for the used privacy assumptions that are presented in section 3.7.1. The major assumption is that participants are honest-but-curious. This is the setting advocated for by most previous studies which provides some assurance, see section 2.7.3. In practice it means that all parties follow the commonly agreed upon rules. Since it seems unlikely that a hospital would go rouge, this assumption is deemed adequate. Frankly, it is assumed that hospitals have at least some degree of mutual trust. An important consequence of this honest-but-curious assumption is that the system does only need to be protected against model leakage, if third party attackers are excluded. This setting is commonly referred to as *global privacy*. This means that the *backdoor attacks* discussed in section 2.8.3, such as *data poisoning attacks* and *model poisoning attacks*, can be disregarded. Finally, as discussed in section 2.9 the model leakage issue can only be ensured using differential privacy, which again motivates why it is absolutely necessary.

Having established that DP is the only mechanism that is needed to protect the federation against itself it is time to turn to the more pressing issue of outside attackers. As discussed in section 3.7.1 the study does not assume secure channels which means that it must be protected against arbitrary third party attacks. This includes the threat of MITM attacks illustrated Figure 2.1. This implies that the system must be protected using some form of encryption, as discussed in section 2.9.

A word of clarification is that although hospitals are deemed trustworthy, the studied scenario does not blindly trust the central server. It is assumed that the server is honest-but-curious in the sense that it does the aggregation and synchronisation according to the rules, but it is unreasonable to assume that it is a perfect so called *trusted aggregator*. The reason for this is that the server has a potentially larger gain from cheating, since it has access to everything as opposed to hospitals that are limited in their reach. For this reasons the hospitals must be provided additional protection against the server. According to the discussion in section 2.9 this is achieved using some form of Secure Multiparty Computation technique, commonly either Homomorphic encryption or Shamir's secrete sharing. Both these techniques provide adequate protection against the server and since it is assumed, see section 3.7.1, that networking is not a bottle neck they are both viable options. However, the current state-of-the-art solution suggests using both, for additional protection. Furthermore they can both be modified such that they provide the additional benefit of protection against some hospital collusion. Although the honest-but-curious assumption implies non-collusion, making this redundant it still can not hurt.

To summaries the application requires DP to protect against inherent model leakage, channel encryption to protect against third party attackers and some form of SMC to limit the server's power. Channel encryption using traditional AES is not strictly necessary if HE is applied, but it can not hurt. In the same way the commonly quoted authorisation mechanisms SHA-256 should also be added on top, in order to provide additional protection against third party attackers posing as insiders. Importantly this protects the system from the previously discarded *backdoor attacks*, which are a threat in this scenario since the third party would likely not be no honest-but-curious. A final notion is that although anonymisation was shown to be insufficient on its own, it still serves as yet another layer of protection. Conclusively, all of these additional layers should provide ample protection against the discussed vulnerabilities. Whether they are legally sufficient or not is left for lawyers to determine. This is beyond the scope of this study and so is the in section 2.9 discussed additional issue that arises if the application crosses multiple legal jurisdictions, which is fairly likely in a real world implementation.

Finally, now that all these additional protective layers are in place it is time to go back to discuss whether the empirical $\epsilon = 20$ level of DP is sufficient. A more detailed analysis of DP in section 2.10.1 reveals that the required level depends on a number of factors. Firstly, it is generally higher if both so called white and black box attacks are possible, which as presented in section 2.8.2 they are in this scenario. Furthermore the leakage is proportional to the number of parameters in the model, which in study is relatively large (17 million). This together with the fact that medical applications are above average in terms of sensitivity suggest that an adequate ϵ in this case is likely in the lower range of the previously mentioned single digit rule of thumb.

However, as also discussed in section 2.10.1, the DP constraint can be relaxed if the setting is non-IID and if the model exhibits minor overfitting which is the case in this study. More importantly, the socially acceptable level depends on the perceived

risk of an attack occurring. In a small scale system as this one there is more control and all clients share the same objective, meaning that there is limited incentive for any insider to attack the system. The perceived risk is thus lower, meaning that a higher ϵ can be allowed [145]. This together with the extensive additional layers suggests that a significantly higher ϵ should be allowed. Furthermore, the level of DP can be increased seamlessly by introducing a random sampling of clients with a probability q at each round, see section 2.10.2. By adding this additional feature in the real world implementation the new goal to meet is $\epsilon = 20q$.

All things considering this suggests that it is at least not impossible that this scenario is adequate from a privacy stand point. Rather, it seems pretty safe to say that it is indeed possible to achieve full protection without loosing any performance. Furthermore, $\epsilon = 10$ is definitely adequate which implies that complete protection can be achieved with maximum 11% performance loss.

5

Discussion

This chapter turns to analyse the system in more depth. The initial section examines if additional performance could be gained if additional experiments were carried out. After this the high false positive rate is analysed in order to understand this fallacy. This is followed by an investigation into the interpretability of the model's predictions. The chapter is concluded with a number of visual examples to qualitatively present how the model behaves.

5.1 The training progress reveals useful insights

The conducted experiments show that super human performance is achieved in the conventional setting. Furthermore, it shows that even after federation and privacy concerns the model performs at a clinically relevant level. This section intends to analyse the training progress of the best models to investigate if it is possible to gain even better performance.

5.2 Additional epochs, increased image size and Focal Loss would improve performance

The first insight is gained from analysing Figure 5.1. This clearly shows that the model has not actually fully converged. This suggests that the current best standalone test performance of 0.83 dice would likely improve if additional epochs were simply added.

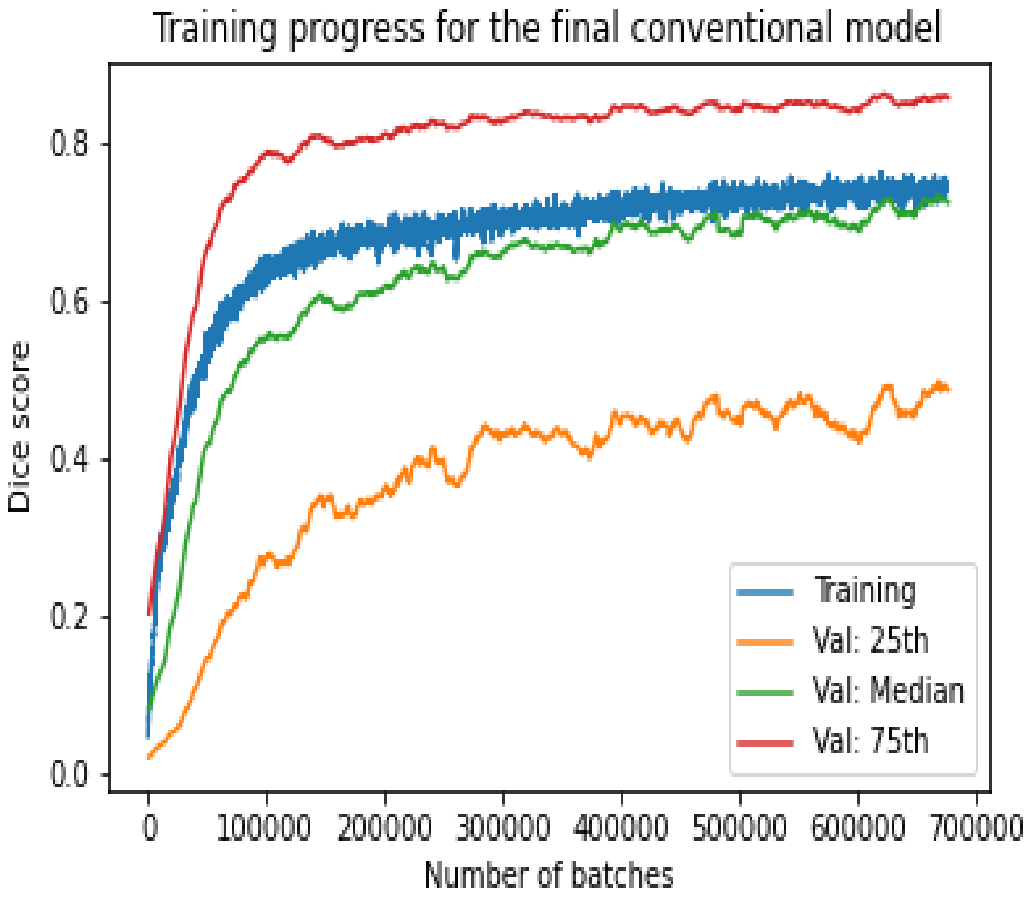


Figure 5.1: The graph illustrates the training progress for the best stand-alone model in the centralised setting, without postprocessing. The training progress is illustrated as a 1000-point moving average and the validation progress as a 20-point moving average to unveil the underlying trends. The model exhibits stable and controlled convergence, with a rapid improvement in the initial stage. Although the model was trained for 300 epochs there is still a slight trend that the performance, both on the training and validation set, is still increasing. This suggests that even better performance might be gained by adding additional epochs. Interestingly the difference in mean training dice and median validation dice is relatively small, which suggests that the model is not overfitting. This together with the fact that performance has not plateaued even after this many epochs is likely a consequence of the high level of augmentation used.

A second potential performance gain is suggested by the results in Figure 5.2. It provides evidence to support that the gain from increasing the image size has not yet been saturated. In other words, if a GPU with more RAM had been used such that the image size could be further increased then it is likely that the final performance would be better than it is now. Considering the large observed gains from increasing the image size, see Table 4.1, this might actually be rather substantial.

Finally, since Focal Loss was indeed shown to provide a small gain especially for the hardest samples it is likely that the performance of the final model would increase, had it been added on top of the L2 regularisation. Note that these two techniques are not mutually exclusive, but this joint model was never examined in this study.

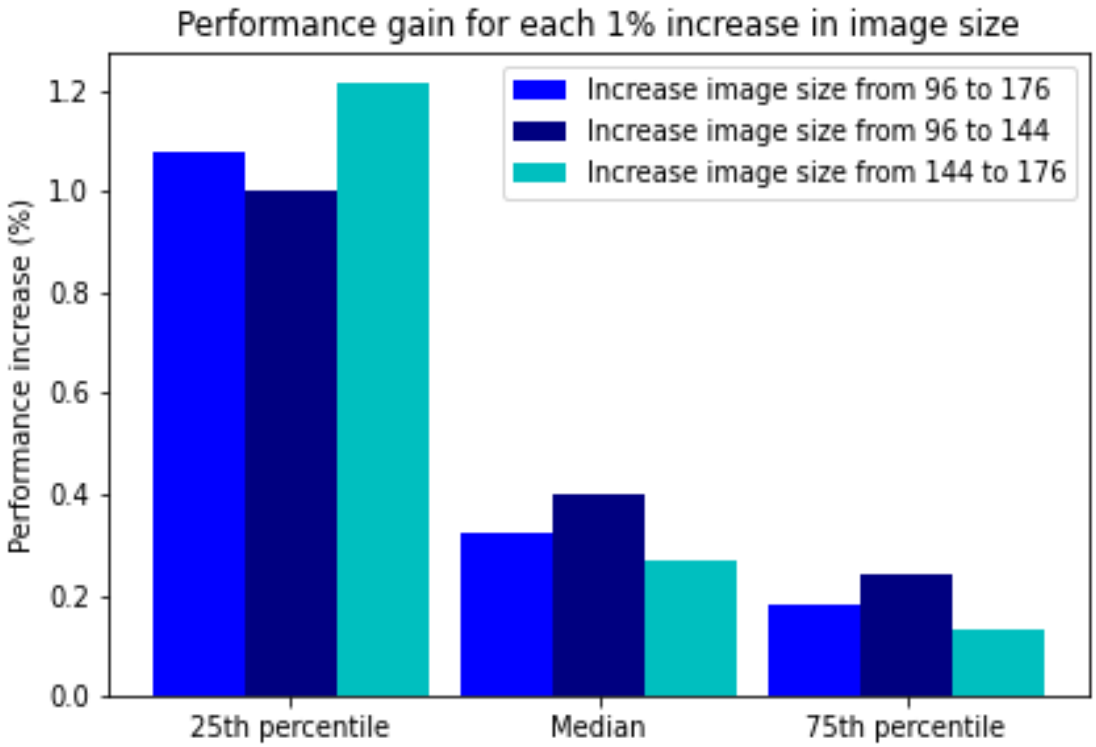


Figure 5.2: From Table 4.1 it is clear that increasing the image size has a positive effect on model performance. The three bar graphs illustrates how many percent the model performance increased as the image size was increased by one percent. It turns out that the return is only marginally diminishing as the input size is further increased, which suggests that the model performance might have been even better if the size had been increased to beyond 176. However, this becomes rather speculative since it relies on only three data points. It is thus hard to quantitatively say how much, if at all, this trend is true beyond the examined domain.

5.2.1 The federated model continues to improve even after the local models have converged

A similar conclusion can be drawn when analysing the training progress of the best federate model, see Figure 5.3. It shows that this model has not fully converged either. In fact, the trend is more significant in this case than in the previously discussed conventional case. This suggests that the final federation might actually be able to perform better than the current 88.6% if it was allowed to run for more rounds.

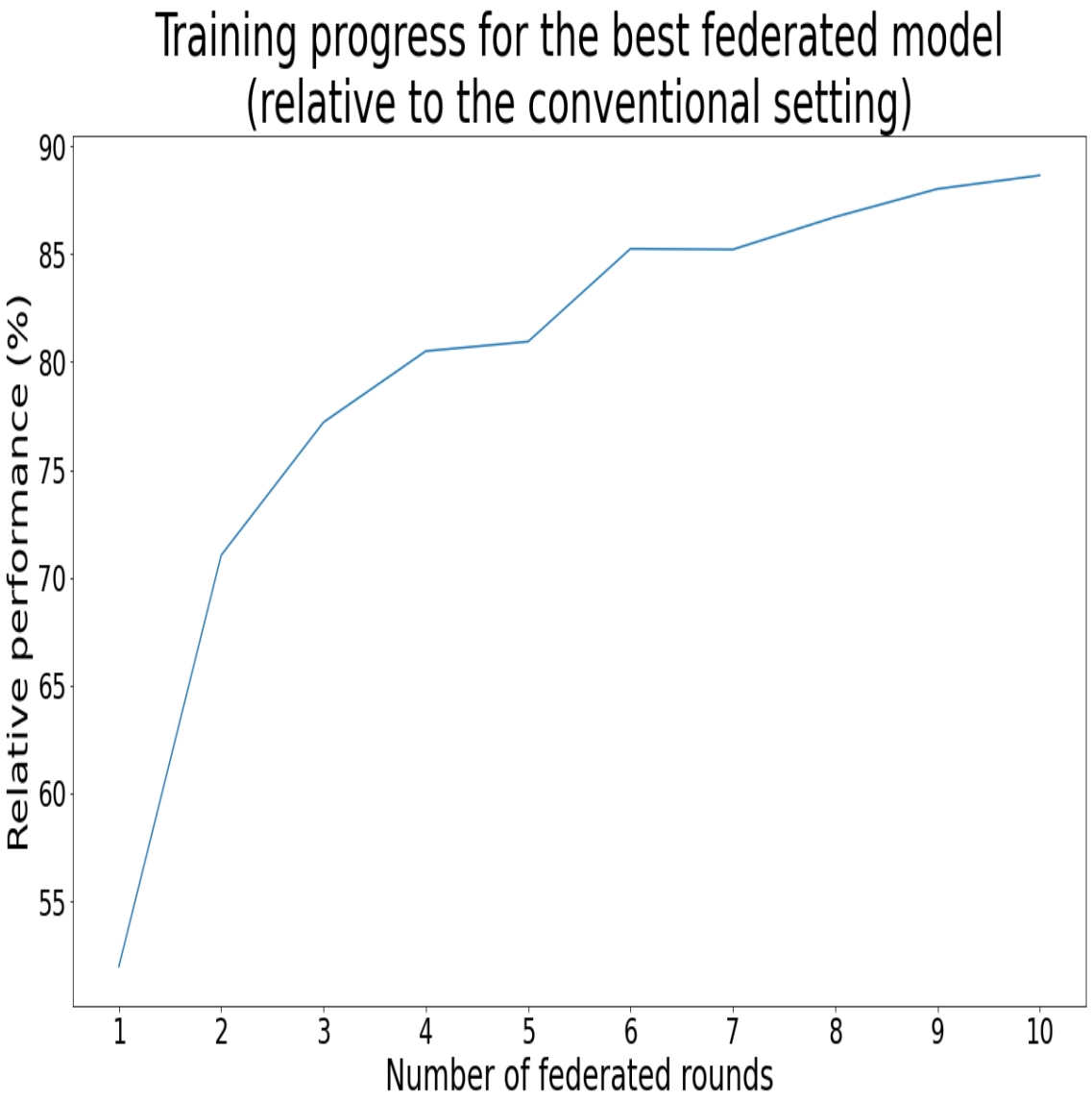


Figure 5.3: The best federated model achieves 88.6% of the equivalent conventional model performance, see Figure 4.5. However, as can be seen from the training progress the model has not fully converged yet. This suggest that if more round of federation had been used, even better performance might have been achieved.

An interesting insight concerning the federation is learnt from analysing the complementary local training progresses, see Figure 5.4, for the best federated model. It shows that even though the five local models seem to have converged in terms of validation the joint model continues to improve. This is quite extraordinary and suggests that the local models must contain complementary information that is somehow useful for the global distribution.

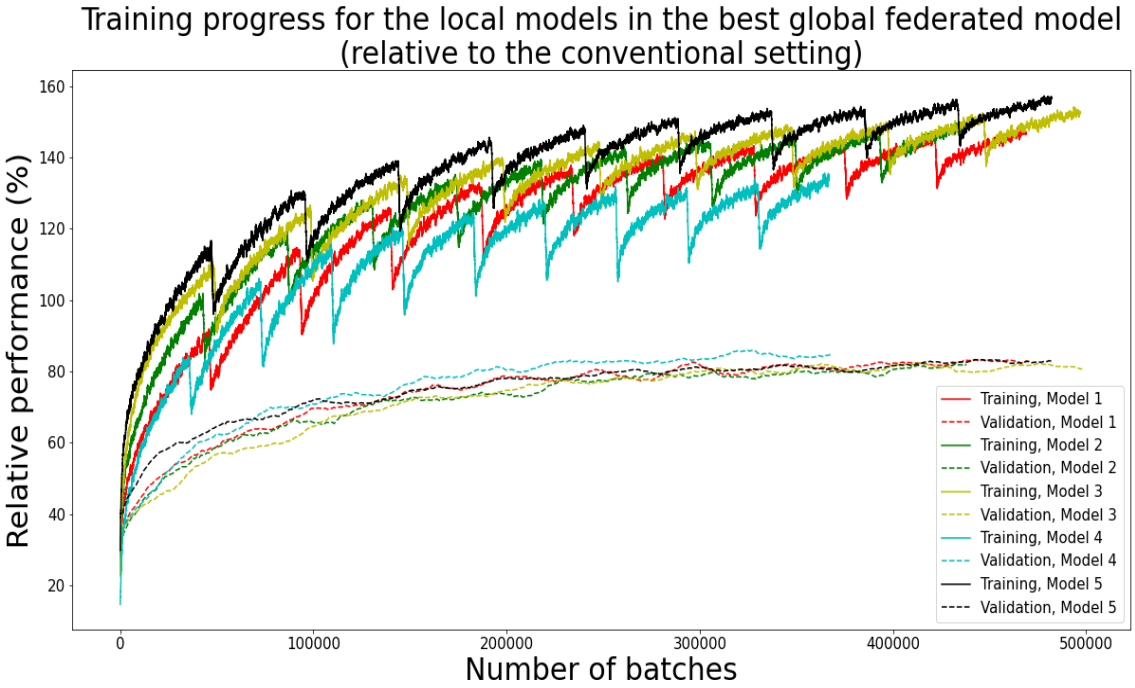
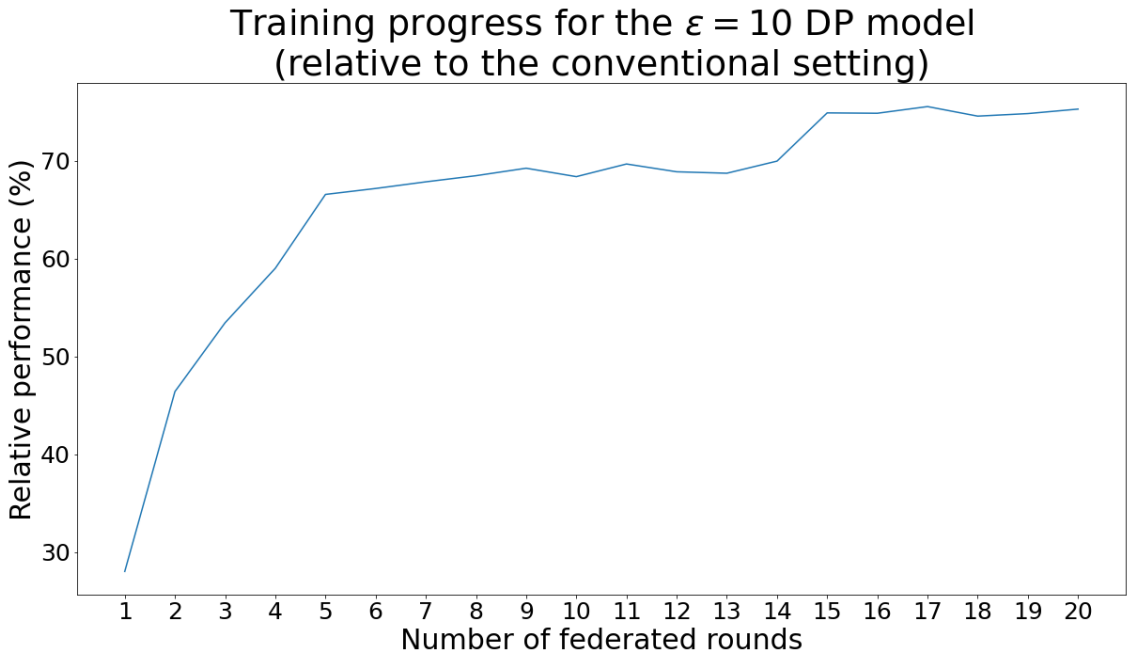


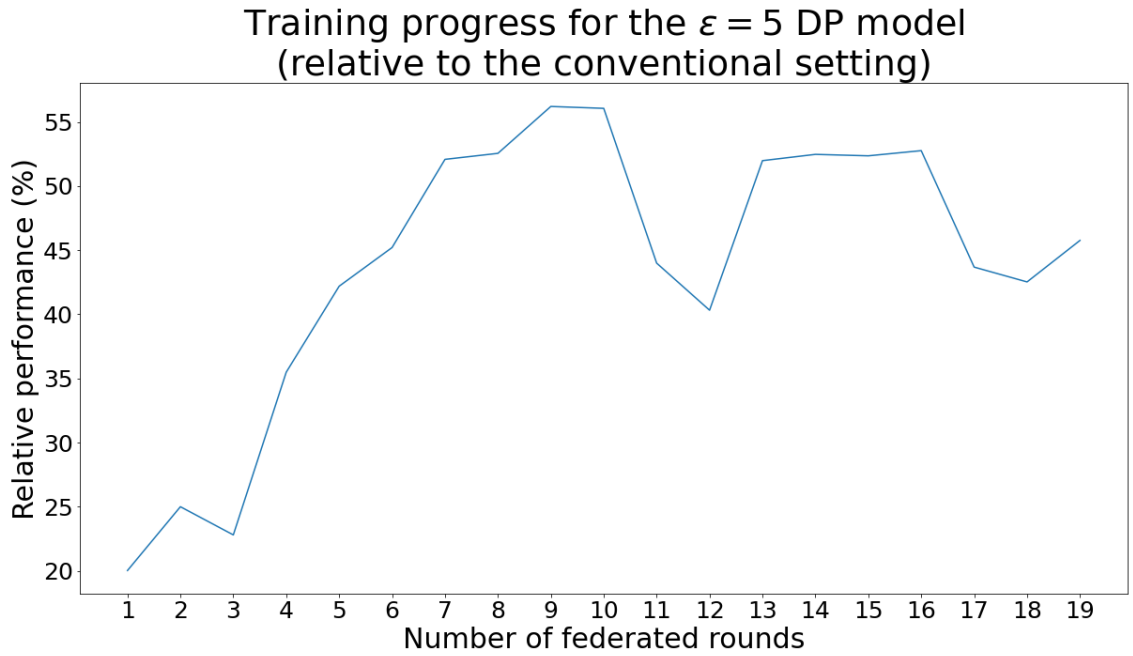
Figure 5.4: The five local models that combine to become the best federated model, see Figure 5.3, all exhibit similar training traits. Firstly, the local training loss dramatically reduces after each round of global aggregation, which is to expect since the next generation model is optimised for another distribution. However, the local models quickly recover which creates the characteristic jagged pattern. Noteworthy is that although the local models seem to converge in terms of validation performance, the global model still shows an increasing trend, see Figure 5.3. This suggests that the models contain complementary information that is useful for the joint, global distribution even though they have been exhausted on their respective local distributions. Note that both the training and validation performance is measured relative to the final validation performance of the equivalent conventional model, hence the illustrated training progress in excess of 100%.

5.3 Excessive Differential Privacy leads to poor performance due to unstable convergence

In the preceding chapter it was shown that federated models with high ϵ-DP perform on par with non-DP models, but that the performance deteriorates drastically if ϵ is decreased. See Figure 4.10. It is possible to gain some understanding as to why this is the case by examining the respective training progresses, illustrated in Figure 5.5. The experiments show that federated models trained with a high degree of differential privacy converge in a more unstable manner, whereas the low privacy models all converged monotonically. This tells us that the excessive noise levels incurred by the high privacy setting leads to very noisy and inadequate model updates, as expected.



(a) The global performance grows monotonically and in a stable manner.



(b) The global model shows a much more unstable progress.

Figure 5.5: In Figure 4.10 is was shown that models with high levels of differential privacy perform much worse. Additional insights are gained by analysing the respective training progresses. As seen the global model performance grows monotonically for the differentially private model with $\epsilon = 10$. The same is true for the models with higher ϵ as well, not shown. However, when increasing the level of privacy to $\epsilon = 5$ or below then the training progress becomes much less stable due to the excessive noise.

5.4 The conventional model performs well in general and poorly on very distinct subsets

From Figure 5.6 it is clear that the model in general performs well on most samples in the test set. However, it also highlight the fact that the model is completely useless for a small subset of samples. Luckily, this troubling behaviour can be explained by examining the distribution in more detail as in Figure 5.7a and Figure 5.7b. With this detailed break down of performance by site and sub type it is plain to see that most bad apples originate from very specific subsets of the data set. Importantly it support the posed hypothesis that ambiguous tumour annotation definitions is troubling. A real world implementation of inter-hospital collaboration would thus have to focus on how to ensure high and consistent annotation quality across institutions.

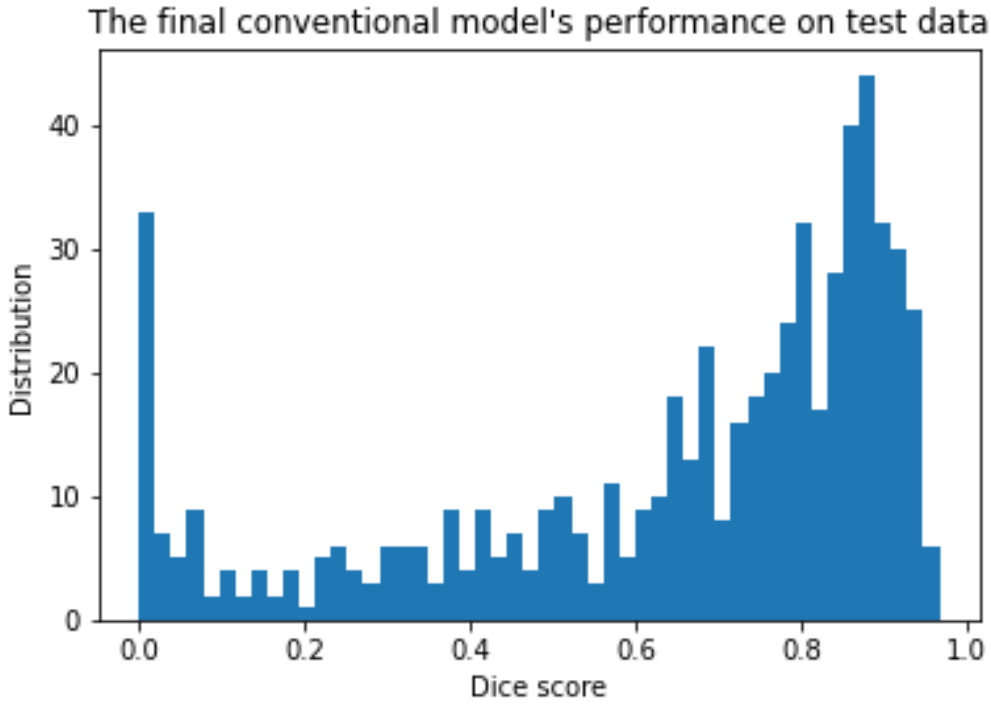
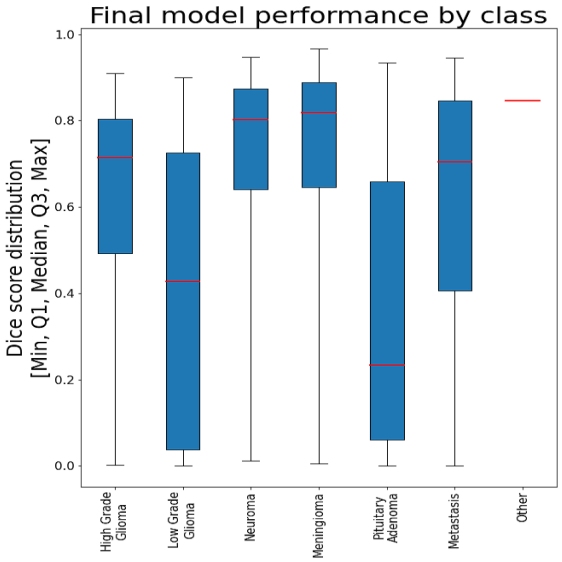


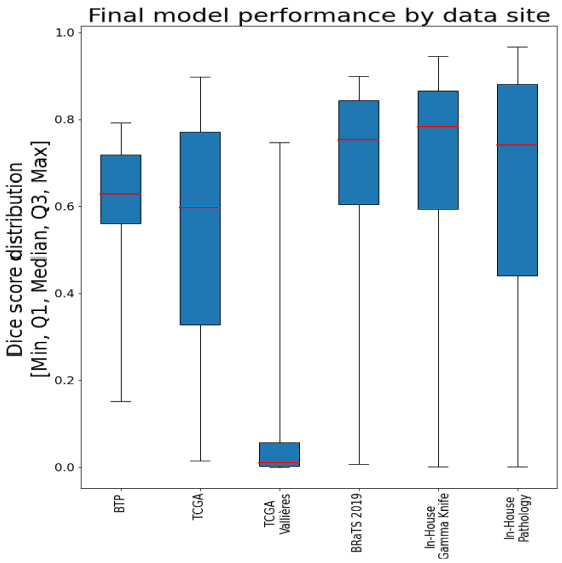
Figure 5.6: The histogram shows the performance of the best stand-alone model in the centralised setting, without postprocessing, for each sample in the blind test set. The distribution shows that the model performs well for most cases. However, it has an abnormal peak at a worthless dice score close to zero. This suggests that there are about 30 samples in the test set that are drastically different from those found in the training set. This and other in-depth details are discussed in Figure 5.7.

Another insight from Figure 5.7a is that the different tumour sub types are indeed perceived differently by the model. This is further supported by the fact that the local models were superior to the non-IID equivalents, as discussed in Figure 4.4,

since they were better at generalising to the global distribution. Conclusively this implies that the demographic differences concerning the abundance of tumour sub types indeed lead to a problematic non-IID situation for a federation. This supports the methodology used in this study.



(a) The model performs on average much worse on low grade Gliomas and Pituitary Adenomas. The performance on the other sub types is relatively consistent. Note that the test set only consists of a single sample for the 'Other' class.



(b) The model performs much worse on data that originates from the set referred to as TCGA Vallieres. The performance on the other sites is relatively consistent, with a slight decrease on TCGA and BTP data.

Figure 5.7: The two box plots present a detailed breakdown of the test performance for the best stand-alone model in the centralised setting, without postprocessing. Most significant is the consistently low performance on the data from the set referred to as TCGA Vallieres, which explains the abnormal peak at 0 dice score shown in figure 5.6. This is likely a consequence of the fact that this data set only highlights the enhancing region, compared to the other sites that mostly follow the 'whole tumour' convention - see Figure 3.4. Furthermore it is in line with theoretical predictions that the model performs worse on low grade Gliomas, since these are the most difficult sub types to accurately segment due to its diffuse boundaries. The reasons behind the low performance on Pituitary Adenomas is likely because this is the least occurring sub type, see Figure 3.1. On the same note it is also not surprising that the model performs best on the In-House and BRaTS 2019 data, since these data sets are the largest. Furthermore, regarding the In-House data it is clear that the model performs better and more consistently on the Gamma Knife samples, which is to expect since their annotation quality is expected to be higher. Finally, the two figures show that there are good and bad apples at all sites and all sub types, as seen from the consistently high and low maximas and minimas respectively.

Another interesting insight can be gained from Figure 5.8 which clearly shows that the model tends to perform better on larger tumours. This is in line with the intuition that small tumours are harder to spot. This correlation also provides additional evidence to support that adding the Focal Loss would be very beneficial.

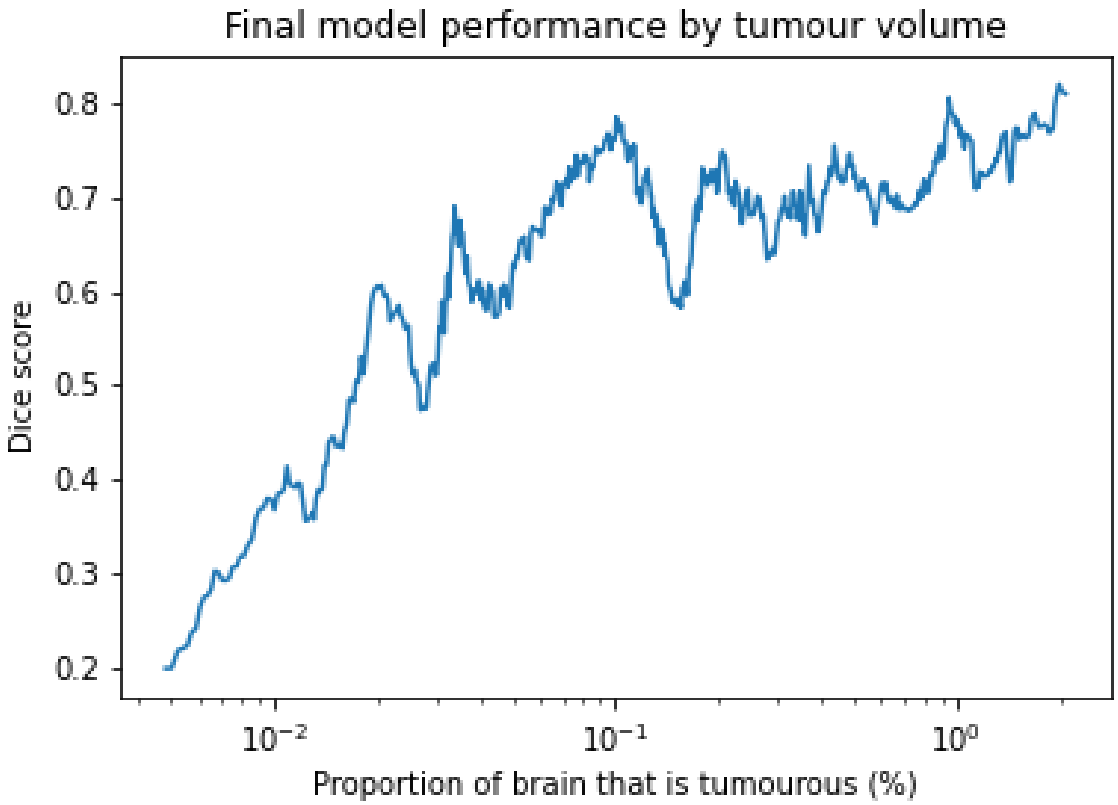


Figure 5.8: Illustrated it the test performance for the best stand-alone model in the centralised setting, without postprocessing. The graph visualises a 30-point moving average to reveal the underlying trend. It clearly shows that the model performs much better on larger tumours, as expected.

5.5 The false positive rate is relatively high, but it might be overestimated

The perhaps most problematic part from the conclusion of the conventional experiments is that the false positive rate is as high as 23.1%. This is of course very troubling if the model is to be used in a clinical setting. However, by analysing each individual false positive case a pattern starts to emerge that makes this less severe. Firstly, as seen in Figure 5.9 the false positive predictions tend to be for very small tumour volumes. This means that it is relatively easy to manually examine and amend the false diagnosis in practice. Conversely, if the false predictions had been all over the place instead it would be a tedious and laborious work to correct this.

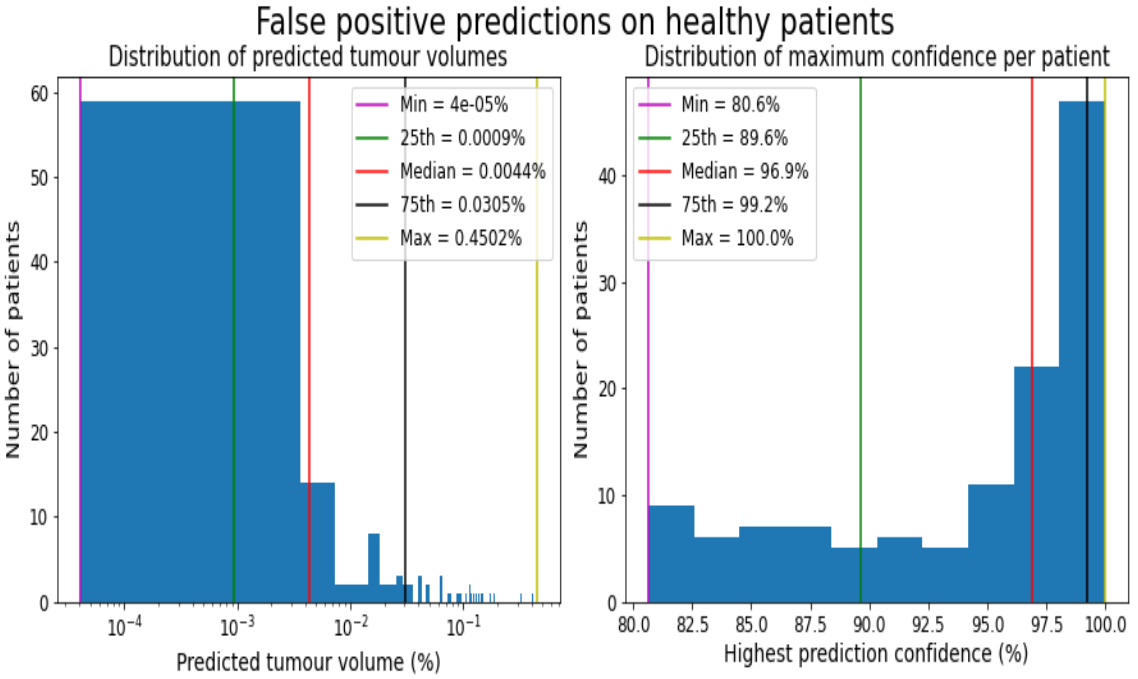


Figure 5.9: The final conventional model, with top-3 ensemble and post processing, incorrectly diagnoses 23.1% of the healthy patients with tumours. However, all the predicted tumours are very small and would thus be relatively easy to manually verify and reject. See Figure 5.10 for a qualitative analysis of these cases. What is worse is that the model does these false predictions with relatively high confidence. The latter might be explained by the fact that the model confidence level is not well calibrated, as discussed in Figure 5.13, which means that the confidence estimate should not be taken too literally.

Furthermore, a qualitative visual scrutiny of these fallacies reveal more reassurance of that the model is actually well behaving. See Figure 5.10 for representative examples. The examination show that whenever the model incorrectly diagnoses a healthy patient with brain tumours it usually does so based on some de facto abnormality in the MRI scan. One possible reason for the high false positive rate is thus that these abnormalities are actually tumours that the annotator missed. Another perhaps more likely explanation is that the mislabelling occurred during the data acquisition process. Bottom line is that, as discussed, since significant amounts of noise exist in the used data set this is a very real possibility. If this is not the case and the model is in fact incorrect in a medical sense, it is still reassuring because it shows that the model tends to identify abnormal regions that at least to a layman look like a tumour. This just means that the model has a high recall and it is up to the doctor to weed out false cases based on subtle medical differences.

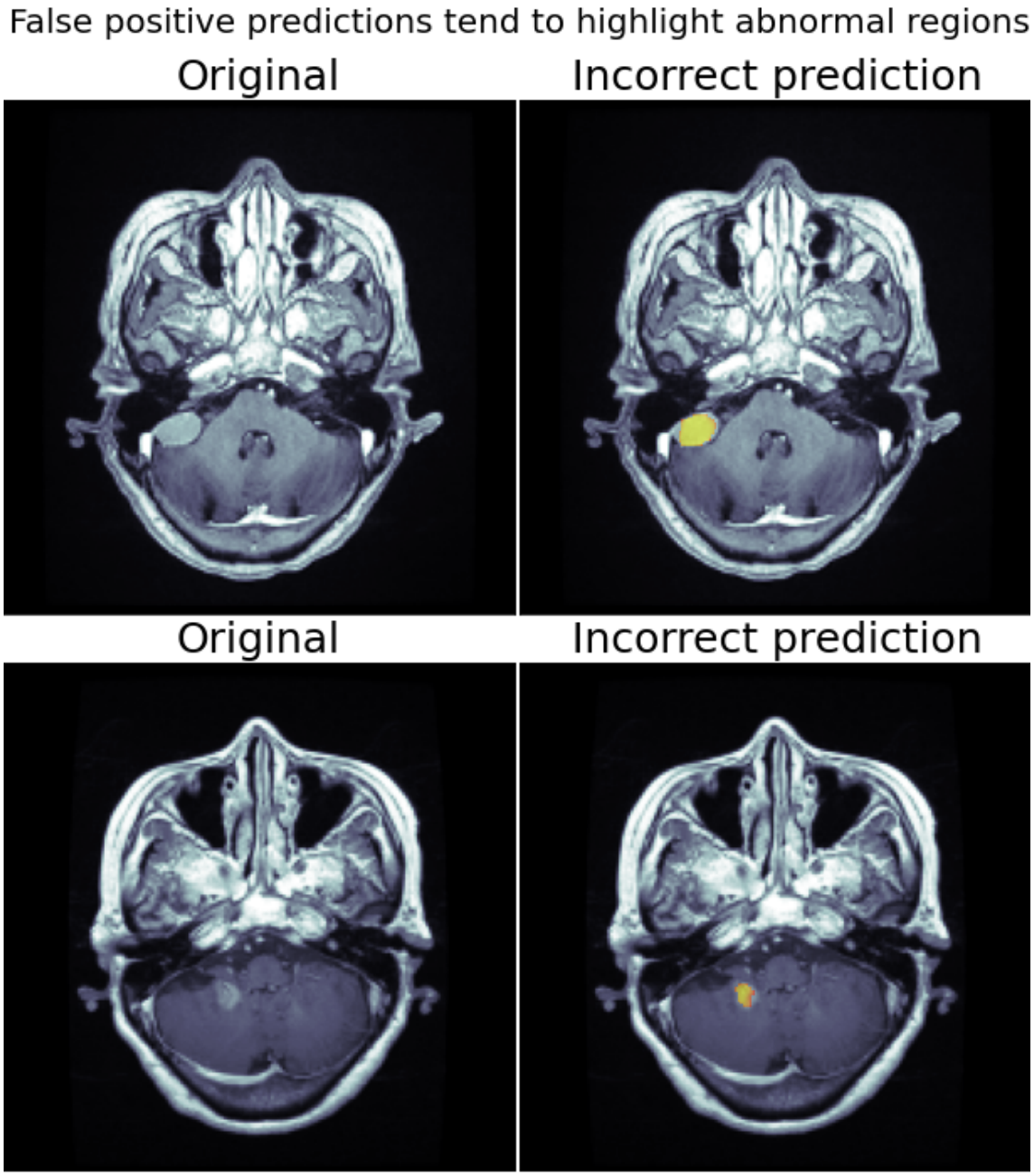


Figure 5.10: The final conventional model, including top-3 ensembling and post-processing, incorrectly diagnoses 23.1% of the healthy patients with tumours, as discussed in Figure 5.9. However, these instances of false positives tend to actually highlight abnormal regions - albeit not diagnosed as tumourous by the professional ground truth, but for a layman it is difficult to make out the subtle difference. Although this does not make up for the relatively high false positive rate, it at least provides some reassurance that the model has a good capacity at identifying abnormal regions. A final note is that after manual verification of the used so called healthy samples, the conclusion is that for several of these erroneous cases it is actually likely that the model is correct and that the ground truth is flawed due to an error in the data collection process. The reported false positive rate is thus likely slightly too pessimistic.

5.6 Confidence and predictive ability is correlated, but the model is excessively pessimistic

This section analyses the final model's behaviour. More specifically, it investigates how confident it is in its predictions. From Figure 5.11 it is clear that a majority of predictions are actually made with a low confidence. Another interpretation of this is that it is highly confident that it is not a tumour, which in this case are two equivalent statements. Only on rare occasions does the model proclaim to be entirely sure that it has found a tumour. The conclusion from the analysed histograms in Figure 5.11 is thus that whenever the model predicts it does so being either very sure that it is, or is not, a tumour. There is no middle ground.

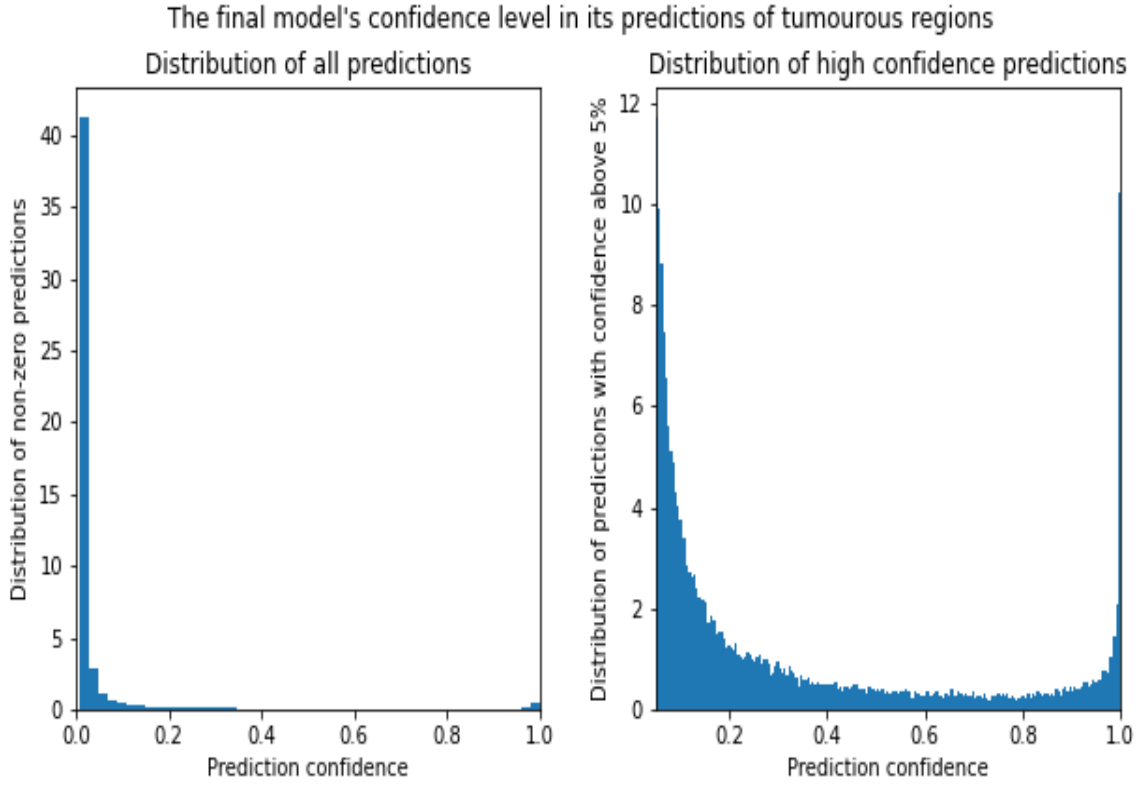


Figure 5.11: From the left hand figure it is clear that a great majority of predictions are actually made with a confidence close to zero. The right hand figure is shown for clarity to highlight the high confidence behaviour. Again it is clear that most model predictions are low confidence. However, as can be seen from the abnormal peak at 100% confidence level the model is indeed highly confident at times.

Another even more interesting insight is provided by Figure 5.12. This analysis shows that whenever the model is correct in its prediction, it was also very confident. Similarly, whenever it was incorrect it was also very hesitant. This strong correlation helps to explain why the postprocessing was shown to have such a positive effect.

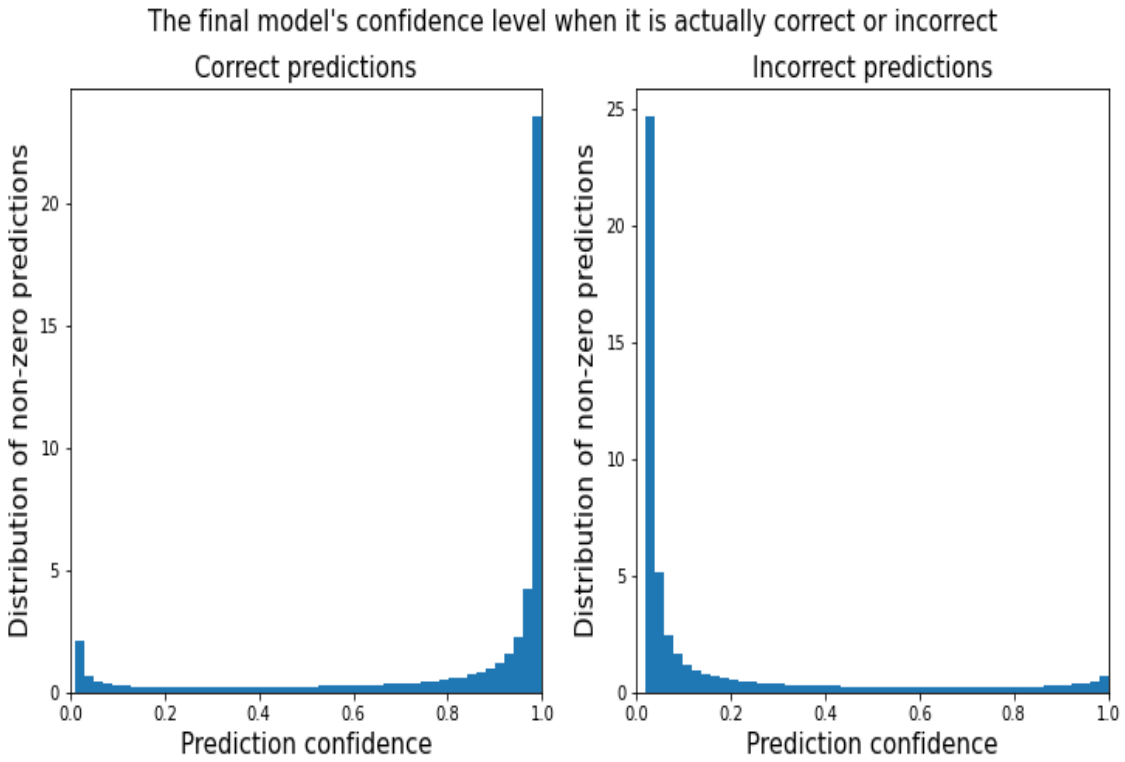


Figure 5.12: As suggested by the promising results of post processing, see Figure 4.1, the model exhibits a strong correlation between its prediction confidence and its ability to be correct. The left hand graph shows that true positives are usually very confident, conversely the right hand graph shows that the false positives are made with very low confidence. However, as seen in Figure 5.13, this does not mean that the confidence score should be taken to literally.

Once it is established that the correlation exist, it start to be interesting to analyse this property quantitatively. The results from such an experiment is presented in Figure 5.13. Based on this examination it is clear that model tends to be over-confident in its prediction, in the sense that if it claims to be x % sure that it has found a tumour the reality is that the likelihood of it actually being a tumour is significantly less than x. This is the equivalent of having a doctor that is overly pessimistic in its diagnosis. However, by having this empirical calibration curve it is now possible to correct for this model bias thus providing increased interpretability of its predictions.

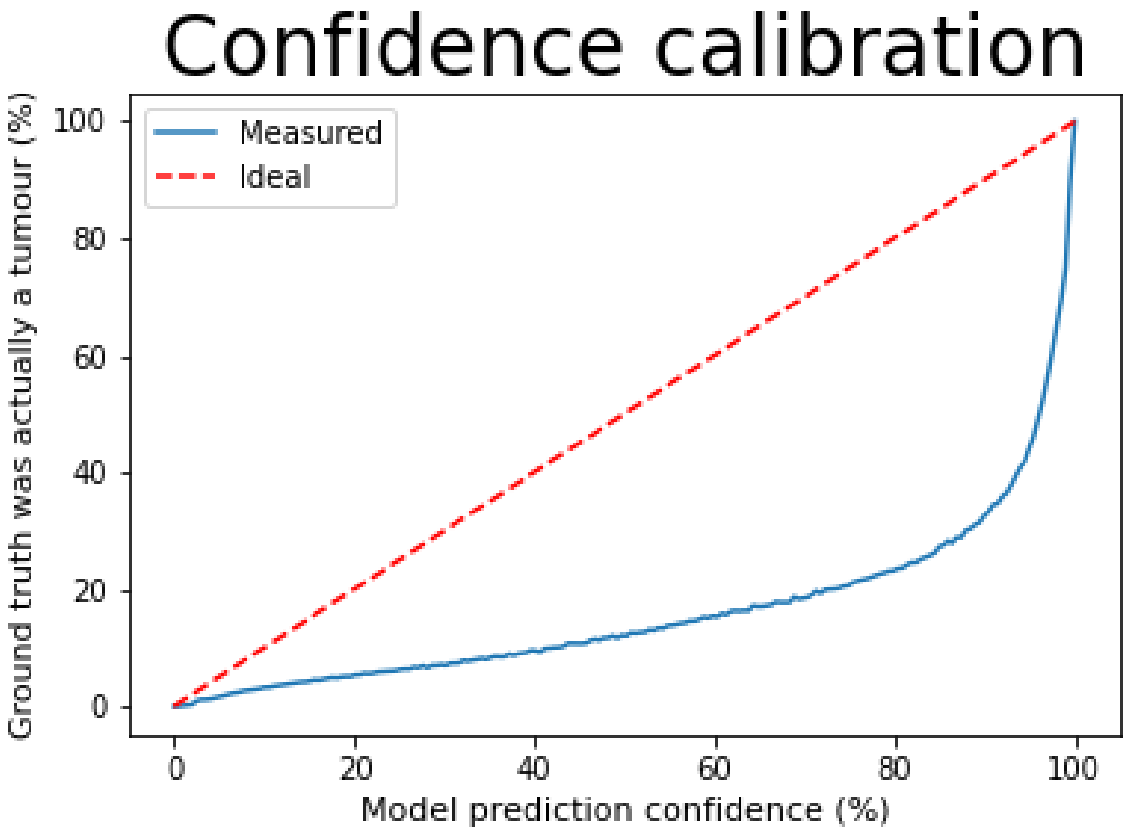
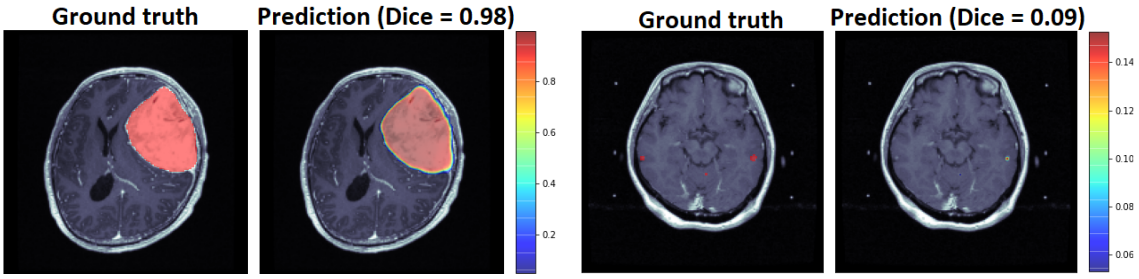


Figure 5.13: From the discussion in Figure 5.12 it is established that there exists a correlation between the confidence score and the ability to predict correctly. This property can be quantified by analysing how often the ground truth for a pixel was actually a tumour, as a function of how confidence the model was of the same. If the model would be perfectly calibrated a confidence of x % would result in that x% of those pixels are actually tumourous. However, as seen in the graph the model exhibits a great overconfidence in its predictions since the empirical curve lies below the ideal at all confidence levels. The model is only perfectly calibrated and thus trustworthy whenever it is either 100 % sure that it is a tumour, or that it is not a tumour. Importantly this calibration curve means that by setting the postprocessing threshold at 80% we are in fact considering all pixels that have at least 25% chance of being a tumour.

5.7 Qualitative illustrations of model behaviour

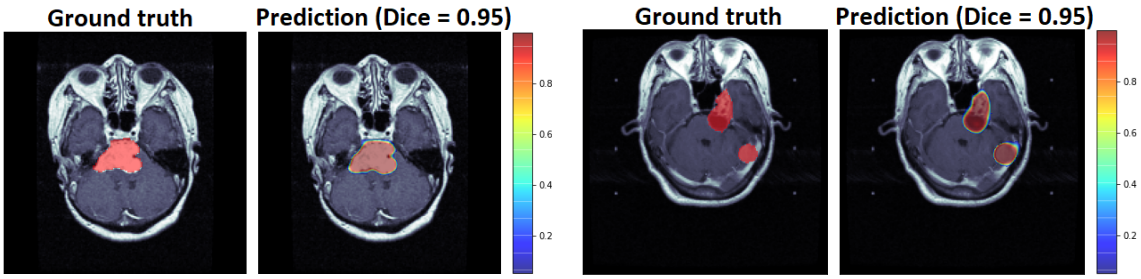
This last section is purely visual. The idea is to show a number of representative cases that illustrates how the model behaves, in order to give some rough intuition of its capabilities and drawbacks. For increased readability the more part of the discussion and comments are provided directly in the captions.



(a) The model can find large tumours. **(b)** The model can find small tumours.

Figure 5.14: The model has the ability to accurately identify both large and small tumours. However the recall is not always perfect - especially for tiny lesions, illustrated in the right hand image where only two out of three tumours have been found.

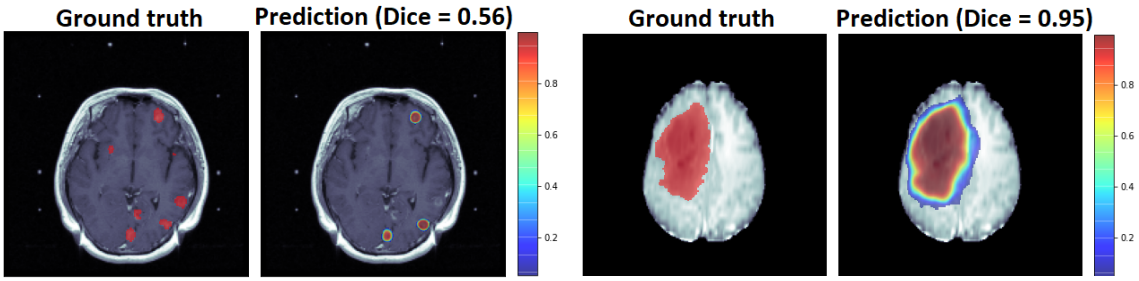
A discussed in section 3.2 the data set contains tumours with a wide array of different characteristics. This implies that a well behaving model must be able to be diverse enough to cope with all these differences. It turns out that the model is versatile enough to accurately identify tumours that are both very small and very large, as illustrated in Figure 5.14b and Figure 5.14a. Furthermore, the model can correctly delineate even complex shapes, see Figure 5.15a, which proves that the model can capture even subtle details.



(a) The model can accurately outline even complex shapes. **(b)** The model can accurately identify multiple tumours in the same patient.

Figure 5.15: The model has the ability to correctly identify several tumours in the same patient. Furthermore, it can correctly provide the outline for even complex, detailed tumour boundaries.

Another necessary trait for a clinically relevant model is that is must be able to identify several lesions in a single patient, especially since Metastases tend to show up in great numbers. The model does indeed exhibit this behaviour, see Figure 5.15b. However, in the specific case of multiple Metastases it proves to be more challenging. This is because this usually coincides with the fact that the lesions are very small, which as already discussed is a know fallacy for the model. However, the model tends to find at least some of them, see Figure 5.16a. This might be considered good enough since it at least makes the doctor vigilant of that the patient is not healthy.



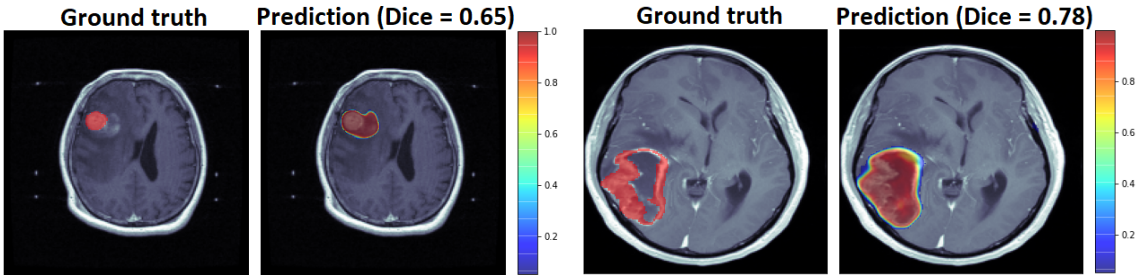
(a) The model identifies some, but not all of the tumours.

(b) The model has issues to accurately outline the peritumoral edema.

Figure 5.16: Although the model recall in general is good, it sometimes fails to correctly identify all lesions in a patient. Furthermore, the model exhibits traits that it has difficulties accurately identifying the exact peritumoral edema boundaries for Gliomas. This is not surprising, since the edema is usually identified by medical professionals using FLAIR or T2-weighted images and not the T1-weighted post contrast used in this study. See Figure 3.3 for a discussion on the different MRI sequences. Besides, the Glioma boundaries are known to be diffuse and difficult to objectively determine in general.

A very important finding concerns the previously discussed annotation ambiguity. Just as expected, the model exhibits some difficulties at accurately delineate the peritumoral edema, see Figure 5.16b. This is not at all surprising since even medical professionals can not always agree on where the edema ends, especially not without also comparing to the FLAIR or T2-weighted images. A more positive notion is that the model seems to have learnt what the majority decision is concerning which annotation definition to use. As illustrated in Figure 5.17b the model has correctly learnt that it should highlight the entire extent of the tumour, not only the enhancing regions. Furthermore, this specific discrepancy might explain why the model performs significantly worse on the TCGA Vallieres data set, discussed and illustrated in Figure 5.7b. This proves that the model is robust to at least some degrees of annotation ambiguity.

Besides the fact that annotation is ambiguous, there are also a significant number of cases where the general annotation quality is just low. However, it turns out that since a majority of samples are adequately annotated, the model has actually learnt to correct these low quality cases. Two illustration of this is seen in Figure 5.17 and Figure 5.18. This proves that the model is robust to at least some degrees of low quality data.



(a) The model corrects inadequate ground truth annotation by correctly identifying the entire tumour.

(b) The model correctly highlights the entire extent of the tumour, not only the enhancing region.

Figure 5.17: Interestingly, the model shows evidence that it is able to correct inadequate ground truth annotations. As discussed in Figure 3.4 different institutions and doctors use different definitions when providing the annotation, which leads to an ambiguity. These illustrated examples show that the model is robust and able to generalise across these discrepancies to some degree. It is also clear that the model tends to annotate using the 'whole tumour' definition, which is expected since the majority of ground truths in the used data follow this convention.

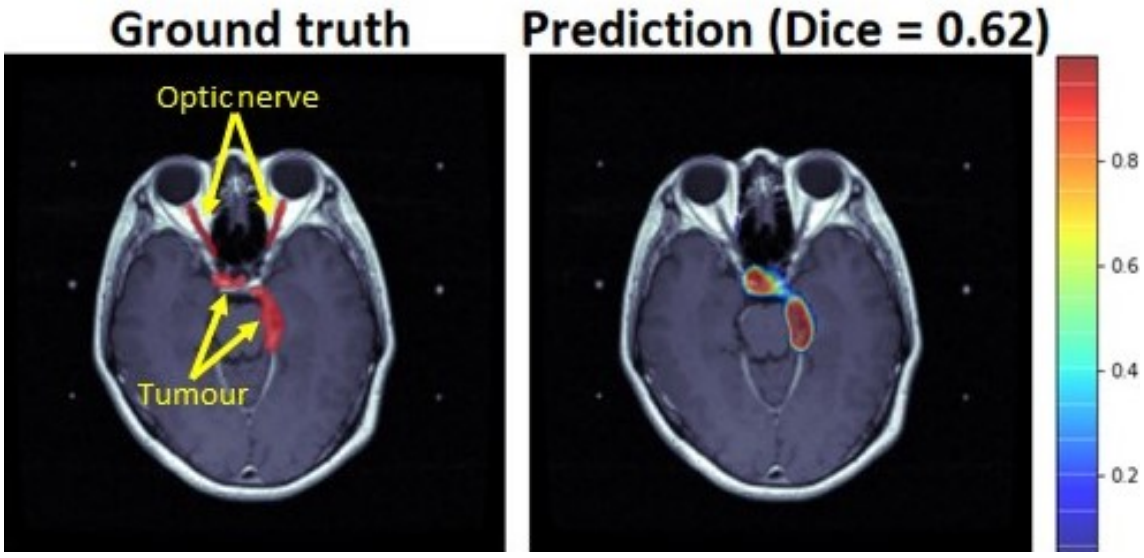
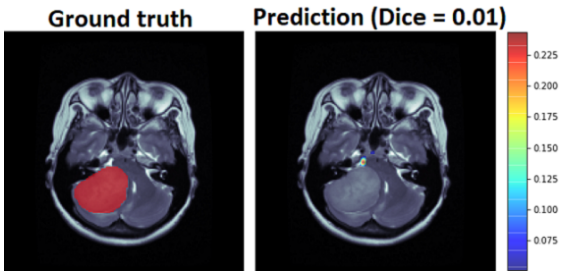


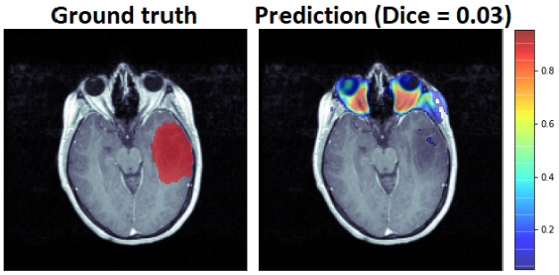
Figure 5.18: In this case the ground truth incorrectly highlights the optic nerves, which is an example of how low quality annotation exist in the data set. Due to that the optic and cochlear (not shown) nerves are sensitive to radiation it is common procedure to highlight these regions before Gamma Knife treatment. This implies that the error is not actually an error in the medical sense, but it becomes problematic from a Data Science perspective since the model is unaware of this. However, as seen from the prediction the model has luckily learnt to disregard and correct for this flaw.

Finally, there are of course also cases where the model does not do very well. In fact, there are examples where it fails dramatically. One such case is illustrated

in Figure 5.19a where the model has completely missed a very large, and visually obvious tumour. Another mistake that occurs from time to time is that the model incorrectly predicts that parts of the eyes are tumours, see Figure 5.19b. However, these spectacular failures are thankfully very rare.



(a) The model fails to identify a very large tumour.



(b) The model accidentally predicts that the eyes are tumours.

Figure 5.19: These two examples show that the model at times performs very poorly. It has dramatically failed to identify two large, and visually simple, tumours. The right hand case illustrates another occurring flaw in that the model incorrectly highlights part of the eyes as tumours due to the fact that they show up bright in the image. However, both of these behaviours are rare. Besides, the error in Figure 5.19a might be explained by the fact that the image is actually a T2-weighted MRI scan. This is yet another example of how occasional flaws exist in the data set due to errors in the data collection process. Since the model is primarily trained to work for T1-weighted post contrast images it is understandable that the performance on other pulse sequences is inferior.

6

Conclusion

In this chapter the concluding remarks derived from the study are presented. Their possible real world implications and applications are discussed to tie back to the purpose of the paper. Lastly, suggestions for future research areas are laid out based on the findings of this study.

6.1 Conclusion

Similarly to previous studies it has been found that it is possible to create a centralised Deep Learning model with human level performance in terms of brain tumour segmentation. The final model achieves a median dice score of 0.87 on the blind test data. A more in depth analysis of the model performance reveal that the major drawback is on small volume tumours, as expected. Additionally it has some difficulties with accurately determining the peritumoral edema boundary. However, this is likely due to the fact that the model is only allowed to view the T1-weighted post contrast sequence when in reality doctors usually use FLAIR or T2-weighted images for this task, since the edema shows up more clearly in such sequences. Lastly, the model shows overconfidence in its predictions which leads to a relatively high false positive rate of 23.1%. These false predictions tend to highlight de facto abnormal tissue, albeit not tumourous. However, the distinction is very subtle and the errors are in line with what a layman would likely do as well. The upside of this model behaviour is that the recall rate is higher, meaning that it is more likely to diagnose healthy patients with tumours than the other way around - which is the lesser evil.

Moreover, others studies have mainly been focused on the BRaTS data set which is significantly more homogeneous than the data set used in this paper. This study thus provides a novel result in that it is indeed possible to create such a model that can learn to generalise across different tumour sub types, institutions, annotation protocols and MRI scanners. Importantly, this implies that a large scale collaboration between hospitals is a feasible possibility. Another contribution, which is in line with the findings from related areas of research, is that more data is better. It was shown that by increasing the data set by a factor of five, the performance rose by 47%.

Furthermore the experiments show that Federated Learning is a feasible method for

creating a cross-hospital model. This moves such a collaboration from a theoretical, idealistic vision to something that could actually be implemented in real life since it circumvents the strict privacy regulations that hinder any traditional methods. As highlighted in the theoretical review, there are still significant practical issues that must be solved before a full scale medical federated system could be rolled out, but this proof of concept has shown that it is at least doable in terms of model convergence. Although the federated model in this study performs 11.4% worse than the ideal, centralised model it is a significant 30% better than the current best option, namely training a hospital unique stand-alone model. The conclusion in terms of Federated Learning is thus that it is still not a perfected method by any means, but it is the de facto best option for cross hospital learning. Lastly, the study proposes a novel technique termed *adaptive momentum* which is shown empirically to improve the previous state-of-the-art.

Finally, a few words on privacy. The approach in this study has been mostly theoretical in this regard, with the exception of implementing differential privacy. It was shown that a federated model is not degraded in any way if moderate levels ($\epsilon \geq 20$) of differential privacy are imposed. However, increasing the privacy into the single digits render the model useless. This might be mitigated slightly if more hospitals than just five are involved, but it proves that differential privacy in itself is not the answer. Instead, the theoretical review suggests that it should be used in conjunction with channel encryption, authentication, homomorphic encryption and Shamir's secret sharing. These are all additional layers of protection that can be used without degrading model performance. Due to the sensitive nature of the data involved, any real world application of federated brain tumour diagnostics should use a mix of all these techniques in order to maximise both protection and performance. This together with the empirical analysis of differential privacy leads to the final conclusion that complete protection of the federated system can be achieved with at most 11% performance reduction, but likely less.

6.2 Future studies

There are several intriguing possibilities for future research that are likely to improve performance even further. Perhaps the most obvious one is to investigate other model architectures and to perform an extensive hyperparameter search. Ideally this grid search should be optimised for the federated setting directly, rather than doing it on the centralised setting as in this study. Other low hanging fruits are to investigate the use of even larger image inputs, N4-ITK bias field correction for preprocessing and CRF for postprocessing. According to previous studies this is likely to increase performance by a few percent. Since ensembling was shown to work well it would be interesting to investigate if even better performance could be gained by retraining several identical versions of the best model, but with different initialisations, and ensemble these. Lastly, more experimental suggestions are to examine using the VAE loss proposed by [4, 5] or to incorporate a lesion prior as in [109]. A potential extension of the proposed lesion prior could be to use different priors at each local client, thus mitigating some of the non-IID difficulties in FL.

Although there is limited empirical evidence to support that these techniques provide any significant performance, they make intuitive sense.

The qualitative analysis of the model performance revealed that the major issues are to identify small tumours, accurately outlining the peritumoral edema and that the model has a tendency to predict parts of the eyes as tumours. The edema issue would likely be mitigated if other MRI sequences were included, especially T2-weighted or FLAIR. In general it is intuitive that the performance would improve if other MRI views or sequences are incorporated, since that means that the model have access to the same information that doctors would analyse. The issue with falsely predicting the eyes and missing tiny tumours might be solved by adding a weighted loss function that punishes these mistakes harder. As discussed in the theoretical review there are indeed several previous publications that advocate for the use of such weightings. Another promising method designed for improving the recall on small tumours is to implement an attention mechanisms as in [93].

So far all suggestions have been directed at improving the model performance, without changing the setting. However, another relevant point to discuss is how to make the solution more practically feasible. As discussed a major bottleneck is to collect samples, while model performance is largely driven by the raw data size. This suggests that more intelligent use of data is the best way forward. The theoretical review makes it clear that there is no obvious way for how to leverage Transfer Learning in this application, but it might be worth looking into in more depth for this reason. Otherwise, using Model Genesis [106] is another promising proxy for TL. Furthermore, several papers show promising results using Curriculum and Active Learning, as well as leveraging weakly annotated samples. All of these three methods would reduce the resource requirement for gathering data significantly which motivates a thorough examination of them.

Besides improving the general model architecture or the applications practical feasibility, there are several intriguing possibilities that related directly to the federated learning. Firstly, the fact that momentum was shown to work well suggests that implementing a novel FedAvg analogue of the Adam optimiser would improve convergence. Secondly, the perhaps most interesting solution to the non-IID issue is the Agnostic FL model proposed by [74]. Due to the limited work on this topic it is not yet practically feasible, but the method makes intuitive sense which suggests that it is worth pursuing further. Lastly, in order to gain more understanding of how the non-IID and heterogeneity impacts performs in FL it would be interesting to replicate the same procedure as in this study but with other local distributions. To get a baseline the experiment should first be replicated using the local IID data sets, since that would reveal to what extent the non-IID condition actually matters. Furthermore the inter-hospital heterogeneity could be examined by splitting the data according to the used annotation definition, the MRI scanner make, original data source etc. Furthermore it is likely instructive to examine how the FL performance is effected if the local data sets are not the same size, which was the case in this study.

A practical issue concerning the federated setting is that hospitals are autonomous, meaning that the data scientist can not assume anything concerning the data quality or availability. The latter motivates an examination of using vertical Federated learning [39, 36], as opposed to the horizontal setting assumed in this study. This would allow the model to train on whatever MRI sequences or views that happen to be available at each location, rather than rejecting all patient cases that are do not have for example T1-weighted post contrast images. Such a vertical setting would thus mean that a larger proportion of available data can be used for training. Finally, the data quality issue implies that future work should examine how the server can selectively chose only well behaving local models, thus rejecting poor updates that are caused by sub par data quality. Some preliminary studies on this are presented in the theoretical review, but it is a relatively new area of research that will require more extensive scrutiny.

A relatively weak part of this study is the extent to which differential privacy was analysed empirically. It is likely that higher levels of privacy than $\epsilon = 20$ can be achieved if this is examined more in depth. Future work should implement the Moments Accountant discussed in the theoretical review to provide a tighter bound. Furthermore it is likely that the level of DP can be increased if more clients are used, since the average is taken over more samples. Another possible improvement could be to add the Gaussian mechanism at each local gradient update instead of to the entire model update, thus providing samples level DP rather than the more restrictive client level case in this study.

Finally, it is necessary to briefly mention the commonly quoted argument against using DL in medical applications, namely that it is a black box approach. As discussed by [19] it is in all data scientists' interest to investigate how these methods can be motivated to sceptical patients, doctors and approving agencies such as the FDA. Currently it is very difficult to get approval for clinical tests due to this reason [19]. Consequently, this must be given more thought if we are to have any chance of moving from research to real world use cases. This is probably the most difficult of all issues left to solve. One step in the right direction might be if we could ensure that models are properly calibrated, as opposed to the result in Figure 5.13. By incorporating an explicit loss that punishes bad calibration it might be possible to make the model more trustworthy, in the sense that the model predictions could be taken more literally. Although this does not solve the black box issue it at least ensures that the model output is more interpretable.

Bibliography

[1] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," 2018, arXiv:1812.00535v3[cs.LG].

[2] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: Information leakage from collaborative deep learning," 2017, arXiv:1702.07464v3[cs.CR].

[3] Schmainda KM, Prah M, "Data from brain-tumor-progression," *The Cancer Imaging Archive*, 2018.

[4] A. Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," 2018, arXiv:1810.11654v3[cs.CV].

[5] A. Myronenko and A. Hatamizadeh, "Robust semantic segmentation of brain tumor regions from 3d mris," 2020, arXiv:2001.02040[eess.IV].

[6] MICCAI BRaTS, "Validation phase leaderboard 2019," 2020. [Online]. Available: https://www.cbica.upenn.edu/BraTS19/lboardValidation.html Accessed: 2020-02-27.

[7] American Cancer Society, *Global Cancer. Facts & Figures*, 4th ed., Atlanta, USA, 2018.

[8] World Health Organization, "Global cancer rates to rise by 50% by 2020," 2003. [Online]. Available: https://scielosp.org/pdf/bwho/2003.v81n5/385-386/en Accessed: 2020-02-26.

[9] World Health Organisation, *World Cancer Report: Cancer Research for Cancer Prevention*. Lyon, France: International Agency for Research on Cancer, 2020.

[10] R. T. Merrell, "Brain tumors," *Disease-a-Month*, vol. 58(12), pp. 678–689, 2012, doi:10.1016/j.disamonth.2012.08.009.

[11] Biomapas, "Brain cancer clinical trials," 2020. [Online]. Available: https://biomapas.eu/indications/brain-cancer-clinical-trials/ Accessed: 2020-02-26.

[12] Brain Tumour Alliance Australia, "Brain tumour facts," 2011. [Online].

Available: https://web.archive.org/web/20140125234503/http://www.btaa.org.au/BrainTumourFactSheet2011.pdf Accessed: 2020-02-26.

[13] National Brain Tumour Society, "Quick brain tumor facts," 2020. [Online]. Available: https://braintumor.org/brain-tumor-information/brain-tumor-facts/ Accessed: 2020-02-26.

[14] National Cancer Institute (a), "Adult central nervous system tumors treatment (pdq®)–health professional version," 2020. [Online]. Available: https://www.cancer.gov/types/brain/hp/adult-brain-treatment-pdq#section/all Accessed: 2020-02-26.

[15] National Cancer Institute (b), "Cancer stat facts: Brain and other nervous system cancer," 2020. [Online]. Available: https://seer.cancer.gov/statfacts/html/brain.html Accessed: 2020-02-26.

[16] Arti Tiwari, Shilpa Srivastava, Millie Pant, Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019, Pattern Recognition Letters, Volume 131, 2020, Pages 244-260, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2019.11.020. (http://www.sciencedirect.com/science/article/pii/S016786551930340X).

[17] S. Margiewicz, C. Cordova, A. S. Chi, and R. Jain, "State of the art treatment and surveillance imaging of glioblastomas," *Seminars in Roentgenology*, vol. 53(1), pp. 23–36, 2018, doi:10.1053/j.ro.2017.11.003.

[18] M. Iv, B. C. Yoon, J. J. Heit, N. Fischbein, and M. Wintermark, "Current clinical state of advanced magnetic resonance imaging for brain tumor diagnosis and follow up," *Seminars in Roentgenology*, vol. 53(1), pp. 45–61, 2018, doi:10.1053/j.ro.2017.11.005.

[19] Pesapane, F., Suter, M. B., Codari, M., Patella, F., Volonté, C., & Sardanelli, F. (2020). Regulatory issues for artificial intelligence in radiology. In Precision Medicine for Investigators, Practitioners and Providers (pp. 533-543). Academic Press.

[20] T. Ben-Nun and T. Hoefler, "Demystifying parallel and distributed deep learning: An in-depth concurrency analysis," 2018, arXiv:1802.09941v2[cs.LG].

[21] K. Bhardwaj, N. Suda, and R. Marculescu, "Edgeai: A vision for deep learning in iot era," 2019, arXiv:1910.10356v1[cs.LG].

[22] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," 2016, arXiv:1510.00149v5[cs.CV].

[23] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," 2018, arXiv:1712.07557v2[cs.CR].

[24] D. Xinxin, M. H. Ang, S. Karaman, and D. Rus, "A general pipeline for 3d detection of vehicles," *IEEE International Conference on Robotics and Automation*, pp. 3194–3200, 2018, doi: 10.1109/ICRA.2018.8461232.

[25] C. Peng, T. Xiao, Z. Li, Y. Jiang, X. Zhang, K. Jia, G. Yu, and J. Sun, "Megdet: A large mini-batch object detector," 2018, arXiv:1711.07240v4[cs.CV].

[26] J. Liang, N. Homayounfar, W.-C. Ma, Y. Xiong, R. Hu, and R. Urtasun, "Polytransform: Deep polygon transformer for instance segmentation," 2019, arXiv:1912.02801[cs.CV].

[27] P. Burlina, K. D. Pacheco, N. Joshi, and D. E. F. andNeil M. Bressler, "Comparing humans and deep learning performance for grading amd: A study in using universal deep features and transfer learning for automated amd analysis," *Computers in Biology and Medicine*, vol. 82, pp. 80–86, 2017, doi:10.1016/j.compbiomed.2017.01.018.

[28] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," 2020, arXiv:1911.04252v2 [cs.LG].

[29] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, arXiv:1704.06857[cs.CV].

[30] W. Jin, M. Fatehi, K. Abhishek, M. Mallya, B. Toyota, and G. Hamarneh, "Applying artificial intelligence to glioma imaging: Advances and challenges," 2019, arXiv:1911.12886[eess.IV].

[31] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?" 2015, arXiv:1511.06348[CS.LG].

[32] P. Vepakomma, O. Gupta, A. Dubey, and R. Raskar, "Reducing leakage in distributed deep learning for sensitive health data," 2019.

[33] K. Chang, N. Balachandar, C. K. Lam, D. Yi, J. M. Brown, A. Beers, B. R. Rosen, D. L. Rubin, and J. Kalpathy-Cramer, "Institutionally distributed deep learning networks," 2017, arXiv:1709.05929[cs.CV].

[34] L. Zhao, Q. Wang, Q. Zou, Y. Zhang, and Y. Chen, "Privacy-preserving collaborative deep learning with unreliable participants," 2019, arXiv:1812.10113v3[cs.CR].

[35] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10(2), 2019, arXiv:1902.04885v1[cs.AI].

[36] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," 2019, arXiv:1907.09693v3[cs.LG].

[37] K. Hsieh, "Machine learning systems for highly-distributed and rapidly-growing data," 2019, arXiv:1910.08663v1[cs.LG].

[38] C. Fung, J. Koerner, S. Grant, and I. Beschastnikh, "Dancing in the

dark: Private multi-party machine learning in an untrusted setting," 2019, arXiv:1811.09712v2[cs.CR].

[39] K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, and Q. Yang, "Secureboost: A lossless federated learning framework," 2019, arXiv:1901.08755v1[cs.LG].

[40] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou, "A hybrid approach to privacy-preserving federated learning," 2019, arXiv:1812.03224v2[cs.LG].

[41] Juba, Brendan & Le, Hai. (2019). Precision-Recall versus Accuracy and the Role of Large Data Sets. Proceedings of the AAAI Conference on Artificial Intelligence. 33. 4039-4048. 10.1609/aaai.v33i01.33014039.

[42] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan, "Do we need more training data?" *International Journal of Computer Vision*, vol. 119, no. 1, p. 76–92, Mar 2015. [Online]. Available: http://dx.doi.org/10.1007/s11263-015-0812-2

[43] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," 2017, arXiv:1707.02968[CS.CV].

[44] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," 2017, arXiv:1712.00409v1[cs.LG].

[45] X. Sun, A. Bommert, F. Pfisterer, J. Rahnenführer, M. Lang, and B. Bischl, "High dimensional restrictive federated model selection with multi-objective bayesian optimization over shifted distributions," 2019, arXiv:1902.08999v2[cs.LG].

[46] Stanford Vision Lab, Stanford University, Princeton University, "Imagenet," 2016. [Online]. Available: http://www.image-net.org/ Accessed: 2020-04-06.

[47] Forbes, "Rethinking privacy for the ai era," 2019. [Online]. Available: https://www.forbes.com/sites/insights-intelai/2019/03/27/rethinking-privacy-for-the-ai-era/#3829d4507f0a Accessed: 2020-02-27.

[48] D. Gabel and T. Hickman, "The rapid evolution of data protection laws," 2019. [Online]. Available: https://iclg.com/practice-areas/data-protection-laws-and-regulations/1-the-rapid-evolution-of-data-protection-laws Accessed: 2020-02-27.

[49] European Union, "REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL," *Official Journal of the European Union*, 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 Accessed: 2020-02-27.

[50] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," 2020, arXiv:1909.11875v2[cs.NI].

[51] C. Xie, S. Koyejo, and I. Gupta, "Asynchronous federated optimization," 2019, arXiv:1903.03934v4[cs.DC].

[52] E. Bakopoulou, B. Tillman, and A. Markopoulou, "A federated learning approach for mobile packet classification," 2019, arXiv:1907.13113v1[cs.LG].

[53] CALIFORNIA DEPARTMENT OF JUSTICE, "California Consumer Privacy Act (CCPA), FACT SHEET," 2018. [Online]. Available: https://oag.ca.gov/system/files/attachments/press_releases/CCPA%20Fact%20Sheet%20%2800000002%29.pdf Accessed: 2020-02-27.

[54] P. Vepakomma, T. Swedish, R. Raskar, O. Gupta, and A. Dubey, "No peek: A survey of private distributed deep learning," 2018, arXiv:1812.03288v1[cs.LG].

[55] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Úlfar Erlingsson, "Scalable private learning with pate," 2018, arXiv:1802.08908v1[stat.ML].

[56] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," pp. 1322–1333, 2015.

[57] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," 2019, arXiv:1908.07873[cs.LG].

[58] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," 2019, arXiv:1906.08935v2[cs.LG].

[59] N. Guha, A. Talwalkar, and V. Smith, "One-shot federated learning," 2019, arXiv:1902.11175v2[cs.LG].

[60] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," 2019, arXiv:1812.06127v4[cs.LG].

[61] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," 2018, arXiv:1810.04304[cs.LG].

[62] W. Li, F. Milletarì, D. Xu, N. Rieke, J. Hancox, W. Zhu, M. Baust, Y. Cheng, S. Ourselin, M. J. Cardoso, and A. Feng, "Privacy-preserving federated brain tumour segmentation," 2019, arXiv:1910.00962v1[cs.CV].

[63] S. Remedios, S. Roy, J. Blaber, C. Bermudez, V. Nath, M. B. Patel, J. A. Butman, B. A. Landman, and D. L. Pham, "Distributed deep learning for robust multi-site segmentation of ct imaging after traumatic brain injury," *Proceedings SPIE, Medical Imaging: Image Processing*, vol. 10949, 2019, doi:10.1117/12.2511997.

[64] S. Silva, B. Gutman, E. Romero, P. M. Thompson, A. Altmann, and M. Lorenzi, "Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data," 2018, arXiv:1810.08553[stat.ML].

[65] D. Liu, T. A. Miller, and K. D. Mandl, "Confederated machine learning on horizontally and vertically separated medical data for large-scale health system intelligence," 2019, arXiv:1910.02109[cs.LG].

[66] J. Jeon, J. Kim, J. Kim, K. Kim, A. Mohaisen, and J.-K. Kim, "Privacy-preserving deep learning computation for geo-distributed medical big-data platforms," 2020, arXiv:2001.02932[cs.LG].

[67] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," 2018, arXiv:1812.00564[cs.LG].

[68] L. Huang and D. Liu, "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records," 2019, arXiv:1903.09296[cs.LG].

[69] Y. Chen, J. Wang, C. Yu, W. Gao, and X. Qin, "Fedhealth: A federated transfer learning framework for wearable healthcare," 2019, arXiv:1907.09173[cs.LG].

[70] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *International journal of medical informatics*, vol. 112, pp. 59–67, 2018, doi: 10.1016/j.ijmedinf.2018.01.007.

[71] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2016, arXiv:1602.05629v3[cs.LG].

[72] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," 2019, arXiv:1807.00459v3[cs.CR].

[73] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," 2019, arXiv:1810.05512v4[eess.AS].

[74] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," 2019, arXiv:1902.00146v1[cs.LG].

[75] TensorFlow, "Tensorflow federated: Machine learning on decentralized data," 2020. [Online]. Available: https://www.tensorflow.org/federated Accessed: 2020-03-06.

[76] Facebook Research, "Crypten," 2019. [Online]. Available: https://github.com/facebookresearch/CrypTen Accessed: 2020-03-06.

[77] H. R. Yuhong Wen, Wenqi Li and P. Dogra, "Federated learning powered by nvidia clara," 2019. [Online]. Available: https://devblogs.nvidia.com/federated-learning-clara/ Accessed: 2020-03-06.

[78] OpenMind, "Answer questions using data you cannot see," 2020. [Online]. Available: https://www.openmined.org/ Accessed: 2020-03-06.

[79] Privacy AI, "A platform for secure, privacy-preserving machine learning," 2020. [Online]. Available: https://privacy.ai/ Accessed: 2020-03-06.

[80] Arkhn, "Federated learning," 2020. [Online]. Available: https://arkhn.org/en/federated/ Accessed: 2020-03-06.

[81] Scaleout Systems, "Scaleout platform," 2020. [Online]. Available: https://scaleoutsystems.com/scaleout-platform Accessed: 2020-03-06.

[82] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, and J. Passerat-Palmbach, "A generic framework for privacy preserving deep learning," 2018, arXiv:1811.04017[cs.LG].

[83] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," 2018, arXiv:1805.04049v3[cs.CR].

[84] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, H. B. McMahan, T. V. Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," 2019, arXiv:1902.01046v2[cs.LG].

[85] T. Yang, G. Andrew, H. Eichner, H. Sun, W. Li, N. Kong, D. Ramage, and F. Beaufays, "Applied federated learning: Improving google keyboard query suggestions," 2018, arXiv:1812.02903v1[cs.LG].

[86] F. Callegati, W. Cerroni, and M. Ramilli, "Man-in-the-middle attack to the https protocol," *IEEE Security & Privacy*, vol. 7(1), pp. 78–81, 2009.

[87] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, M. Prastawa, E. Alberts, J. Lipkova, J. Freymann, J. Kirby, M. Bilello, H. Fathallah-Shaykh, R. Wiest, J. Kirschke, B. Wiestler, R. Colen, A. Kotrotsou, P. Lamontagne, D. Marcus, M. Milchenko, A. Nazeri, M.-A. Weber, A. Mahajan, U. Baid, E. Gerstner, D. Kwon, G. Acharya, M. Agarwal, M. Alam, A. Albiol, A. Albiol, F. J. Albiol, V. Alex, N. Allinson, P. H. A. Amorim, A. Amrutkar, G. Anand, S. Andermatt, T. Arbel, P. Arbelaez, A. Avery, M. Azmat, P. B., W. Bai, S. Banerjee, B. Barth, T. Batchelder, K. Batmanghelich, E. Battistella, A. Beers, M. Belyaev, M. Bendszus, E. Benson, J. Bernal, H. N. Bharath, G. Biros, S. Bisdas, J. Brown, M. Cabezas, S. Cao, J. M. Cardoso, E. N. Carver, A. Casamitjana, L. S. Castillo, M. Catà, P. Cattin, A. Cerigues, V. S. Chagas, S. Chandra, Y.-J. Chang, S. Chang, K. Chang, J. Chazalon, S. Chen, W. Chen, J. W. Chen, Z. Chen, K. Cheng, A. R. Choudhury, R. Chylla, A. Clérigues, S. Colleman, R. G. R. Colmeiro, M. Combalia, A. Costa, X. Cui, Z. Dai, L. Dai, L. A. Daza, E. Deutsch, C. Ding, C. Dong, S. Dong, W. Dudzik, Z. Eaton-Rosen, G. Egan, G. Escudero, T. Estienne, R. Everson, J. Fabrizio, Y. Fan, L. Fang, X. Feng, E. Ferrante, L. Fidon, M. Fischer, A. P. French, N. Fridman, H. Fu, D. Fuentes, Y. Gao, E. Gates, D. Gering, A. Gholami, W. Gierke, B. Glocker, M. Gong, S. González-Villá, T. Grosges, Y. Guan, S. Guo, S. Gupta, W.-S. Han, I. S. Han, K. Harmuth, H. He, A. Hernández-Sabaté, E. Herrmann, N. Himthani, W. Hsu, C. Hsu, X. Hu, X. Hu, Y. Hu,

Y. Hu, R. Hua, T.-Y. Huang, W. Huang, S. V. Huffel, Q. Huo, V. HV, K. M. Iftekharuddin, F. Isensee, M. Islam, A. S. Jackson, S. R. Jambawalikar, A. Jesson, W. Jian, P. Jin, V. J. M. Jose, A. Jungo, B. Kainz, K. Kamnitsas, P.-Y. Kao, A. Karnawat, T. Kellermeier, A. Kermi, K. Keutzer, M. T. Khadir, M. Khened, P. Kickingereder, G. Kim, N. King, H. Knapp, U. Knecht, L. Kohli, D. Kong, X. Kong, S. Koppers, A. Kori, G. Krishnamurthi, E. Krivov, P. Kumar, K. Kushibar, D. Lachinov, T. Lambrou, J. Lee, C. Lee, Y. Lee, M. Lee, S. Lefkovits, L. Lefkovits, J. Levitt, T. Li, H. Li, W. Li, H. Li, X. Li, Y. Li, H. Li, Z. Li, X. Li, Z. Li, X. Li, W. Li, Z.-S. Lin, F. Lin, P. Lio, C. Liu, B. Liu, X. Liu, M. Liu, J. Liu, L. Liu, X. Llado, M. M. Lopez, P. R. Lorenzo, Z. Lu, L. Luo, Z. Luo, J. Ma, K. Ma, T. Mackie, A. Madabushi, I. Mahmoudi, K. H. Maier-Hein, P. Maji, C. Mammen, A. Mang, B. S. Manjunath, M. Marcinkiewicz, S. McDonagh, S. McKenna, R. McKinley, M. Mehl, S. Mehta, R. Mehta, R. Meier, C. Meinel, D. Merhof, C. Meyer, R. Miller, S. Mitra, A. Moiyadi, D. Molina-Garcia, M. A. B. Monteiro, G. Mrukwa, A. Myronenko, J. Nalepa, T. Ngo, D. Nie, H. Ning, C. Niu, N. K. Nuechterlein, E. Oermann, A. Oliveira, D. D. C. Oliveira, A. Oliver, A. F. I. Osman, Y.-N. Ou, S. Ourselin, N. Paragios, M. S. Park, B. Paschke, J. G. Pauloski, K. Pawar, N. Pawlowski, L. Pei, S. Peng, S. M. Pereira, J. Perez-Beteta, V. M. Perez-Garcia, S. Pezold, B. Pham, A. Phophalia, G. Piella, G. N. Pillai, M. Piraud, M. Pisov, A. Popli, M. P. Pound, R. Pourreza, P. Prasanna, V. Prkovska, T. P. Pridmore, S. Puch, Élodie Puybareau, B. Qian, X. Qiao, M. Rajchl, S. Rane, M. Rebsamen, H. Ren, X. Ren, K. Revanuru, M. Rezaei, O. Rippel, L. C. Rivera, C. Robert, B. Rosen, D. Rueckert, M. Safwan, M. Salem, J. Salvi, I. Sanchez, I. Sánchez, H. M. Santos, E. Sartor, D. Schellingerhout, K. Scheufele, M. R. Scott, A. A. Scussel, S. Sedlar, J. P. Serrano-Rubio, N. J. Shah, N. Shah, M. Shaikh, B. U. Shankar, Z. Shboul, H. Shen, D. Shen, L. Shen, H. Shen, V. Shenoy, F. Shi, H. E. Shin, H. Shu, D. Sima, M. Sinclair, O. Smedby, J. M. Snyder, M. Soltaninejad, G. Song, M. Soni, J. Stawiaski, S. Subramanian, L. Sun, R. Sun, J. Sun, K. Sun, Y. Sun, G. Sun, S. Sun, Y. R. Suter, L. Szilagyi, S. Talbar, D. Tao, D. Tao, Z. Teng, S. Thakur, M. H. Thakur, S. Tharakan, P. Tiwari, G. Tochon, T. Tran, Y. M. Tsai, K.-L. Tseng, T. A. Tuan, V. Turlapov, N. Tustison, M. Vakalopoulou, S. Valverde, R. Vanguri, E. Vasiliev, J. Ventura, L. Vera, T. Vercauteren, C. A. Verrastro, L. Vidyaratne, V. Vilaplana, A. Vivekanandan, G. Wang, Q. Wang, C. J. Wang, W. Wang, D. Wang, R. Wang, Y. Wang, C. Wang, G. Wang, N. Wen, X. Wen, L. Weninger, W. Wick, S. Wu, Q. Wu, Y. Wu, Y. Xia, Y. Xu, X. Xu, P. Xu, T.-L. Yang, X. Yang, H.-Y. Yang, J. Yang, H. Yang, G. Yang, H. Yao, X. Ye, C. Yin, B. Young-Moxon, J. Yu, X. Yue, S. Zhang, A. Zhang, K. Zhang, X. Zhang, L. Zhang, X. Zhang, Y. Zhang, L. Zhang, J. Zhang, X. Zhang, T. Zhang, S. Zhao, Y. Zhao, X. Zhao, L. Zhao, Y. Zheng, L. Zhong, C. Zhou, X. Zhou, F. Zhou, H. Zhu, J. Zhu, Y. Zhuge, W. Zong, J. Kalpathy-Cramer, K. Farahani, C. Davatzikos, K. van Leemput, and B. Menze, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," 2019, arXiv:1811.02629v3[cs.CV].

[88] Mamta Mittal, Lalit Mohan Goyal, Sumit Kaur, Iqbaldeep Kaur, Amit Verma, D. Jude Hemanth, Deep learning based enhanced tumor segmentation approach for MR brain images, Applied Soft Computing, Volume 78, 2019, Pages 346-354, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2019.02.036. (http://www.sciencedirect.com/science/article/pii/S1568494619301000).

[89] T.-W. Ho, H. Qi, F. Lai, F.-R. Xiao, and J.-M. Wu, "Brain tumor segmentation using u-net and edge contour enhancement," *Proceedings of the 2019 3rd International Conference on Digital Signal Processing*, p. 75–79, 2019, doi: https://doi.org/10.1145/3316551.3316554.

[90] Pablo Ribalta Lorenzo, Jakub Nalepa, Barbara Bobek-Billewicz, Pawel Wawrzyniak, Grzegorz Mrukwa, Michal Kawulok, Pawel Ulrych, Michael P. Hayball, Segmenting brain tumors from FLAIR MRI using fully convolutional neural networks, Computer Methods and Programs in Biomedicine, Volume 176, 2019, Pages 135-148, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2019.05.006. (http://www.sciencedirect.com/science/article/pii/S0169260718315955).

[91] Sun Li, Zhang Songtao, Chen Hang, Luo Lin, "Brain tumor segmentation and survival prediction using multimodal mri scans with deep learning," 2019, doi: 10.3389/fnins.2019.00810.

[92] M. H. Vu, T. Nyholm, and T. Löfstedt, "Tunet: End-to-end hierarchical brain tumor segmentation using cascaded networks," 2019, arXiv:1910.05338v3[eess.IV].

[93] Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache "Deep learning with mixed supervision for brain tumor segmentation," Journal of Medical Imaging 6(3), 034002 (10 August 2019). https://doi.org/10.1117/1.JMI.6.3.034002.

[94] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, arXiv:1505.04597[cs.CV].

[95] I. Brahim, D. Fourer, V. Vigneron and H. Maaref, "Deep Learning Methods for MRI Brain Tumor Segmentation: a comparative study," 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey, 2019, pp. 1-6.

[96] Shengcong Chen, Changxing Ding, Minfeng Liu, Dual-force convolutional neural networks for accurate brain tumor segmentation, Pattern Recognition, Volume 88, 2019, Pages 90-100, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2018.11.009.

[97] Liu H., Shen X., Shang F., Ge F., Wang F. (2019) CU-Net: Cascaded U-Net with Loss Weighted Sampling for Brain Tumor Segmentation. In: Zhu D. et al. (eds) Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy. MBIA 2019, MFCA 2019. Lecture Notes in Computer Science, vol 11846. Springer, Cham.

[98] Linmin Pei, Lasitha Vidyaratne, Md Monibor Rahman, Khan M. Iftekharuddin, "Deep learning with context encoding for semantic brain tumor segmentation and patient survival prediction," Proc. SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis, 113140H (16 March 2020); https://doi.org/10.1117/12.2550693.

[99] J. Zhang, Z. Jiang, J. Dong, Y. Hou and B. Liu, "Attention Gate ResU-Net for automatic MRI brain tumor segmentation," in IEEE Access, 2020, doi: 10.1109/ACCESS.2020.2983075.

[100] Kumar S., Negi A., Singh J.N. (2019) Semantic Segmentation Using Deep Learning for Brain Tumor MRI via Fully Convolution Neural Networks. In: Satapathy S., Joshi A. (eds) Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies, vol 106. Springer, Singapore.

[101] S. Bakas, "2018 international miccai brats challenge," *Pre-Conference Proceedings of the 7th MICCAI BraTS Challenge*, 2018. [Online]. Available: https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2018_proceedings_shortPapers.pdf Accessed: 2020-04-06.

[102] Yang, T., Song, J., Li, L., Tang, Q. 'Improving Brain Tumor Segmentation on MRI Based on the Deep U-net and Residual Units'. 2020 : pp 95 – 110.

[103] Sharma D., Shanis Z., Reddy C.K., Gerber S., Enquobahrie A. (2019) Active Learning Technique for Multimodal Brain Tumor Segmentation Using Limited Labeled Images. In: Wang Q. et al. (eds) Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data. DART 2019, MIL3ID 2019. Lecture Notes in Computer Science, vol 11795. Springer, Cham.

[104] Abd-Ellah M.K., Khalaf A.A.M., Awad A.I., Hamed H.F.A. (2019) TPUAR-Net: Two Parallel U-Net with Asymmetric Residual-Based Deep Convolutional Neural Network for Brain Tumor Segmentation. In: Karray F., Campilho A., Yu A. (eds) Image Analysis and Recognition. ICIAR 2019. Lecture Notes in Computer Science, vol 11663. Springer, Cham.

[105] S. Nema, A. Dudhane, S. Murala, and S. Naidu, "Rescuenet: An unpaired gan for brain tumor segmentation," 2020, https://doi.org/10.1016/j.bspc.2019.101641.

[106] Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models genesis: Generic autodidactic models for 3d medical image analysis," 2019, arXiv:1908.06912[eess.IV].

[107] Chandan Ganesh Bangalore Yogananda, Sahil S. Nalawade, Gowtham K. Murugesan, Ben Wagner, Marco C. Pinho, Baowei Fei, Ananth J. Madhuranthakam, Joseph A. Maldjian, "Fully automated brain tumor segmentation and survival prediction of gliomas using deep learning and mri," 2019, doi: https://doi.org/10.1101/760157.

[108] Hu, X., Luo, W., Hu, J. et al. Brain SegNet: 3D local refinement network for brain lesion segmentation. BMC Med Imaging 20, 17 (2020). https://doi.org/10.1186/s12880-020-0409-2.

[109] P.-Y. Kao, J. W. Chen, and B. S. Manjunath, "Improving 3d u-net for brain tumor segmentation by utilizing lesion prior," 2020, arXiv:1907.00281v3[cs.CV].

[110] Sajid, S., Hussain, S. & Sarwar, A. Brain Tumor Detection and Segmentation in MR Images Using Deep Learning. Arab J Sci Eng 44, 9249–9261 (2019). https://doi.org/10.1007/s13369-019-03967-8.

[111] Jie Chang, Luming Zhang, Naijie Gu, Xiaoci Zhang, Minquan Ye, Rongzhang Yin, Qianqian Meng, A mix-pooling CNN architecture with FCRF for brain tumor segmentation, Journal of Visual Communication and Image Representation, Volume 58, 2019, Pages 316-322, ISSN 1047-3203, https://doi.org/10.1016/j.jvcir.2018.11.047.

[112] Muhammad Sajjad, Salman Khan, Khan Muhammad, Wanqing Wu, Amin Ullah, Sung Wook Baik, Multi-grade brain tumor classification using deep CNN with extensive data augmentation, Journal of Computational Science, Volume 30, 2019, Pages 174-182, ISSN 1877-7503, https://doi.org/10.1016/j.jocs.2018.12.003. (http://www.sciencedirect.com/science/article/pii/S1877750318307385).

[113] M. Soltaninejad, L. Zhang, T. Lambrou, G. Yang, N. Allinson, and X. Ye, "Mri brain tumor segmentation using random forests and fully convolutional networks," 2019, arXiv:1909.06337v1[cs.CV].

[114] Fatih Özyurt, Eser Sert, Engin Avci, Esin Dogantekin, Brain tumor detection based on Convolutional Neural Network with neutrosophic expert maximum fuzzy sure entropy, Measurement, Volume 147, 2019, 106830, ISSN 0263-2241, https://doi.org/10.1016/j.measurement.2019.07.058. (http://www.sciencedirect.com/science/article/pii/S0263224119306876).

[115] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," 2015, arXiv:1512.04150v1[cs.CV].

[116] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," 2019, arXiv:1610.02391v4[cs.CV].

[117] Muhammad Irfan Sharif, Jian Ping Li, Muhammad Attique Khan, Muhammad Asim Saleem, Active deep neural network features selection for segmentation and recognition of brain tumors using MRI images, Pattern Recognition Letters, Volume 129, 2020, Pages 181-189, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2019.11.019.

[118] COCO - Common Objects in Context, "Coco 2019 object detection task," 2019. [Online]. Available: http://cocodataset.org/#detection-2019 Accessed: 2020-04-06.

[119] Stanford University, Center for Artificial Intelligence in Medicine & Imaging, "Medical imagenet," 2017. [Online]. Available: https://aimi.stanford.edu/research/medical-imagenet Accessed: 2020-04-06.

[120] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," pp. 1725–1732, 2014.

[121] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," 2019, arXiv:1907.06987[cs.CV].

[122] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, P. Bilic, P. F. Christ, R. K. G. Do, M. Gollub, J. Golia-Pernicka, S. H. Heckers, W. R. Jarnagin, M. K. McHugo, S. Napel, E. Vorontsov, L. Maier-Hein, and M. J. Cardoso, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019, arXiv:1902.09063v1[cs.CV].

[123] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning (extended version)," 2019, arXiv:1812.11494v3[cs.IT].

[124] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data," 2018, arXiv:1811.11479v1[cs.LG].

[125] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," 2019, arXiv:1812.11750v3[cs.LG].

[126] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," 2019, arXiv:1811.12470v4[cs.LG].

[127] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," 2019, arXiv:1812.00984v2[stat.ML].

[128] "Robust and communication-efficient federated learning from non-iid data," 2019, arXiv:1903.02891[cs.LG].

[129] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," 2018, arXiv:1712.01887v2[cs.CV].

[130] Fortune Business Insights, "Magnetic resonance imaging mri systems market," 2019. [Online]. Available: https://www.fortunebusinessinsights.com/industry-reports/magnetic-resonance-imaging-mri-systems-market-100087 Accessed: 2020-03-05.

[131] M. Shayan, C. Fung, C. J. M. Yoon, and I. Beschastnikh, "Biscotti: A ledger for private and secure peer-to-peer machine learning," 2019, arXiv:1811.09904v4[cs.LG].

[132] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017, arXiv:1412.6980v9[cs.LG].

[133] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," *Association for Computing Machinery*, 2017.

[134] C. Dwork, "Differential privacy," *ICALP*, vol. 4052, pp. 1–12, 2006.

[135] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," 2006, arXiv:0610105v2[cs.CR].

[136] M. Wjst, "Caught you: threats to confidentiality due to the public release of large-scale genetic data sets," *BMC medical ethics*, vol. 11, 2010.

[137] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA law review*, vol. 57(1701), 2010.

[138] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks," 2018, arXiv:1812.00910v1[stat.ML].

[139] National Institute of Standards and Technology, "Announcing the advanced encryption standard (aes)," *Federal Information Processing Standards Publication*, vol. 197, 2001.

[140] W. Penard and T. van Werkhoven, "On the secure hash algorithm family (chapter 1)," 2008. [Online]. Available: https://web.archive.org/web/20160330153520/http://www.staff.science.uu.nl/~werkh108/docs/study/Y5_07_08/infocry/project/Cryp08.pdf Accessed: 2020-03-16.

[141] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Trans. Information Forensics and Security*, vol. 13(5), p. 1333–1345, 2018.

[142] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," *IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60, 2010, doi: 10.1109/FOCS.2010.12.

[143] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. T. Li, and Zhang, "Deep learning with differential privacy," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16*, 2016, arXiv:1607.00133v2[stat.ML].

[144] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," 2013, arXiv:1311.0776[cs.DS].

[145] C. Dwork, "Differential privacy: A survey of results," *Theory and applications of models of computation*, pp. 1–19, 2008.

[146] "Paillier, P. (1999, May). Public-key cryptosystems based on composite degree

residuosity classes. In International conference on the theory and applications of cryptographic techniques (pp. 223-238). Springer, Berlin, Heidelberg."

[147] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34(10), 1993-2024, 2015, doi:10.1109/TMI.2014.2377694.

[148] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Nature Scientific Data*, vol. 4:170117, 2017, doi:10.1038/sdata.2017.117.

[149] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," 2018, arXiv:1811.02629.

[150] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al., "Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection," *The Cancer Imaging Archive*, 2017, doi:10.7937/K9/TCIA.2017.KLXWJJ1Q.

[151] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al. , "Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection," *The Cancer Imaging Archive*, 2017, doi:10.7937/K9/TCIA.2017.GJQ7R0EF.

[152] Pedano, N., Flanders, A. E., Scarpace, L., Mikkelsen, T., Eschbacher, J. M., Hermes, B., ... Ostrom, Q. (2016). Radiology Data from The Cancer Genome Atlas Low Grade Glioma [TCGA-LGG] collection. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2016.L4LTD3TK.

[153] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, Journal of Digital Imaging, Volume 26, Number 6, December, 2013, pp 1045-1057.

[154] Su, C., Vallières, M., & Bai, H. (2017). ROI Masks Defining Low-Grade Glioma Tumor Regions In the TCGA-LGG Image Collection [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2017.BD7SGWCA.

[155] Hao Zhou, Martin Vallières, Harrison X. Bai, Chang Su, Haiyun Tang, Derek Oldridge, Zishu Zhang, Bo Xiao, Weihua Liao, Yongguang Tao, Jianhua Zhou, Paul Zhang, Li Yang; MRI features predict survival and molecular markers in diffuse lower-grade gliomas. Neuro Oncol 2017 now256. DOI: 10.1093/neuonc/now256.

[156] J. Cheng, "Brain tumor dataset," 2017. [Online]. Available: https://figshare.com/articles/brain_tumor_dataset/1512427

[157] S. Jadhav, "3d mri brain tumor segmentation using autoencoder regularization," 2020. [Online]. Available: https://github.com/IAmSuyogJadhav/3d-mri-brain-tumor-segmentation-using-autoencoder-regularization Accessed: 2020-05-13.

[158] TOP 500 List, Taiwania 2, 2019. [Online]. Available: https://www.top500.org/system/179590 Accessed: 2020-04-06.

[159] TOP 500 List, November 2019, 2019. [Online]. Available: https://www.top500.org/list/2019/11/?page=1 Accessed: 2020-04-06.