Training a U-Net Model for Brain Tumour Segmentation Using Federated Learning

Jeremy Stott
School of Computer Science and
Applied Mathematics, Wits University
Johannesburg, South Africa
2368841@students.wits.ac.za

Hairong Bau
School of Computer Science and
Applied Mathematics, Wits University
Johannesburg, South Africa
hairong.bau@wits.ac.za

Terence L. van Zyl
Institute for Intelligent Systems
University of Johannesburg
Johannesburg, South Africa
tvanzyl@uj.ac.za

***Abstract*—In the critical context of brain tumour diagnostics, where timely and precise segmentation can greatly influence treatment outcomes, this study examines the viability of implementing the U-Net model within a Federated Learning (FL) framework to address privacy concerns associated with centralised medical data. Using the Brain and Tumour Segmentation (BraTS) dataset, we adapt the U-Net architecture, reducing the number of feature maps to suit computational limits of FL without significant loss in segmentation performance. Through a comparative analysis involving varying numbers of FL clients, our research evaluates the model's accuracy against centrally trained equivalents. The findings suggest that, while centralised models outperform in metric evaluations, FL models with fewer clients show promising segmentation performance. However, FL models face diminishing returns as client numbers increase without proportional data diversity.**

***Index Terms*—Federated learning, U-Net, Medical image segmentation, Brain tumour segmentation**

I. Introduction

Medical Image Segmentation (MIS) has become an indispensable tool in the fight against cancer, especially for brain tumours, which are one of the leading causes of global mortality. These tumours pose complex challenges due to their aggressive nature and the non-uniform shapes, making accurate segmentation crucial for guiding treatment and significantly impacting survival outcomes. The U-Net model has notably advanced MIS by enhancing the precision of tumour detection and delineation [38]. Alongside this progress, the necessity for patient data privacy has highlighted the value of Federated Learning (FL). FL facilitates the collaborative development of robust models while keeping patient data localised, thus addressing key privacy concerns in healthcare.

Reflecting the severity of brain tumours and their potential to impair essential cognitive and motor functions, there is a pressing need to improve segmentation. This study probes whether a compact U-Net trained with FL can equal the performance of centrally trained models, hypothesising that a smaller U-Net can maintain satisfactory segmentation accuracy while being more resource-efficient. The hypothesis is tested by analysing models trained on the BraTS dataset, with the goal of directly influencing treatment decisions and enhancing the quality of life for patients.

In the landscape of MIS, several successful methods have been documented, each demonstrating effectiveness on specific datasets. For instance, convolutional neural networks (CNNs) have achieved remarkable segmentation accuracy on datasets like the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [7], which has been crucial in lung cancer detection. Similarly, methods that incorporate deep learning algorithms have shown significant promise in the segmentation of lesions from the International Skin Imaging Collaboration (ISIC) dataset [6], enhancing the diagnostic process for skin cancer. Such methods have not only validated the use of deep learning in MIS, but also set a precedent for the kind of accuracy and efficiency that this study aims to achieve for brain tumour segmentation using a compact U-Net model within an FL framework.

The specific research question that this study addresses is: Can a U-Net model, optimised within an FL framework with fewer feature maps, perform with a segmentation accuracy comparable to that of a centrally trained U-Net model in the context of brain tumour identification? The aim is to bridge the gap between advanced segmentation accuracy and computational efficiency, which is crucial given the aggressive nature of brain tumours and the urgency in treating them. The objectives include: (1) training and evaluating a streamlined U-Net on centralised data as a baseline model; (2) evaluating the performance of a streamlined U-Net model within an FL framework; (3) comparing the model's performance to that of a standard U-Net model trained on centralised data; and (4) analysing the impact of client numbers in an FL setting on model accuracy and training dynamics.

Our findings reveal that, while centralised models retain a lead in performance metrics, FL models exhibit considerable promise, particularly in smaller, more controlled configurations. Yet, an increase in the number of FL clients without a corresponding expansion in data volume can adversely affect model performance. This is a critical insight, given the variable nature of brain tumours and the diverse manifestations of cancer. It suggests that the design of FL systems in healthcare must carefully consider the balance between the number of learning nodes and the volume and variety of data each node processes. The source code is available on a GitHub repository[1].

In summary, this study provides an examination of the viability of FL in MIS. By advancing our understanding of the effectiveness and limitations of FL, we contribute to the goal of improving computer-aided cancer diagnosis and treatment, ultimately impacting patient care and survival rates in the face of one of the most challenging medical conditions.

The remainder of this paper is structured to guide the reader from a broad context to specific details. Section II lays the foundation of the study by presenting a background on MIS and FL, alongside a review of pertinent literature within the domain of FL in healthcare. Section III delineates the proposed methodology, including the dataset, model architectures, FL framework, and evaluation metrics. It also discusses the methods of comparison and carefully considers the limitations and risks associated with the study. Section IV presents the experimental setup and results. Finally, Section V concludes this paper and summarises key insights and implications of the findings.

II. Related Work

This section aims to provide a background of the fundamental concepts pertaining to this paper and to discuss previous work in the MIS and FL domains.

A. Background

1) Image Segmentation: The goal of MIS is to highlight areas of interest in medical images [28]. The need to reduce diagnosis time and expert-related costs has fostered the development of such models. Image segmentation has become an essential component of computer-aided diagnosis due to the models' accuracy and efficiency. MIS has been applied in a wide array of use cases such as liver-tumour segmentation [21], optic disc segmentation and cardiac image segmentation [37]. Brain-tumour segmentation has also been a popular application of deep image segmentation models [23]. The images used by these segmentation models include Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and X-Ray. Typical MIS methods include template matching, edge detection, and more recently: machine learning [37]. Machine learning approaches have generally followed a structure similar to the U-Net [27]. Although significant progress has been made in image segmentation, the application of U-Net models to medical images remains a challenging process due to the complexity of feature representation [37]. The challenges in processing medical images arise from their monochromatic nature and susceptibility to blur, noise, and low contrast [37].

2) Federated Learning: Machine learning models have been traditionally trained on centralised data [32]. However, there are many scenarios in which such training methods are not suitable. [32] highlight issues of creating centralised datasets such as privacy, legal and data ownership issues. Since some scenarios require data that cannot be centralised, new methods of training have been developed that do not have this strict restriction, the most successful of which is FL [31]. There are other decentralised training methods such as cyclic institutional incremental learning (CIIL) and institutional incremental learning (IIL) [31]. Training models using FL often results in poorer performance, especially when trained on non-independent and identically distributed (Non-IID) data [41]. The FL scheme consists of multiple clients where each of them trains local models on their own data [22]. The local models then send their model parameters (or weights) to a server which performs model aggregation, after which the new model is distributed to the local clients [22]. In this way, a model can be trained on decentralised data without compromising the privacy of the data.

3) Evaluation Metrics: Evaluation metrics form an important part of machine learning as models need to be evaluated in an objective manner. Evaluation metrics for centralised (or traditional) ML and FL settings are discussed in the following. Traditional evaluation metrics focus on measuring the quality of the predictions or segmentations produced by a model. Commonly used evaluation metrics include recall, precision, accuracy, and the F1-score [11]. In the context of MIS, metrics such as the Dice Similarity Coefficient (Dice) and the Jaccard index are commonly used to measure the agreement between the segmentation predicted by the model and the ground truth annotations [4]. In addition to traditional evaluation metrics, FL introduces unique measurements. FL-specific evaluation metrics include communication efficiency, privacy preservation, and model convergence across all clients. Communication rounds refer to the number of times the clients and server exchange information during training. Communication overhead refers to the time and resources required for transmitting data between clients and the server [40]. Privacy preservation evaluates the degree to which learnt data representations protect the privacy of the training data. Another common evaluation metric is the training loss measured over communication rounds as used by [22].

B. Traditional Approaches to Brain Tumour Segmentation

In the quest to delineate brain tumours from magnetic resonance imaging, researchers have explored a variety of non-deep learning techniques with varying degrees of success. Traditional methods primarily take advantage of the intrinsic intensity values of image pixels, often divided into thresholding or region-based strategies.

Thresholding methods simplify segmentation by categorising regions according to pixel intensity values. A seminal method within this category is Otsu's thresholding, which computes an optimal global threshold to separate the tumour from the background [26]. The Otsu's algorithm has been further refined with morphological operations to improve tumour detection [14], and has also been adapted for tumour grading by integrating intensity metrics with logical formulas for extraction [19]. Despite these advances, the sensitivity of thresholding to noise remains a significant challenge. To address these limitations, researchers have turned to statistical

[1] https://github.com/jeremyscodes/Training-a-U-Net-Model-for-Brain-Tumour-Segmentation-Using-Federated-Learning

optimisation methods. One such method, Particle Swarm Optimisation (PSO), has been specifically harnessed to enhance the separation between tumour regions and healthy brain tissue, thereby increasing the robustness of the segmentation process. This approach aims to maximise the variance between classes, leading to a more precise identification of tumours [30]. This was combined with feature extraction techniques and various machine learning classifiers to improve grade classification.

Region-based approaches focus on extracting interconnected regions by following a set of conditions related to intensity. Such methods often begin with a seed point, a pre-selected pixel, or voxel, and iteratively incorporate neighbouring points that exhibit similar intensity values, allowing for the segmentation of coherent structures. Notable efforts in this area include the Localised Active Contour Method with Background Intensity Compensation (LACM-BIC), which capitalises on the fusion of MRI modalities T1 and T2, together with the clustering algorithm of k-means, to define tumour boundaries [15]. Further, the symmetry analysis in 3D-MR images by Kermi, Andjouh, and Zidane [18] and the modified level set method of Virupakshappa and Amarapur [35] reflect the continuous pursuit of refinement of segmentation accuracy and robustness against the inherent challenges of MRI data, such as noise and non-uniformity of intensity. These challenges can obscure tumour margins and complicate the accurate extraction of pathological regions, underscoring the importance of such innovative approaches in improving segmentation precision.

C. U-Net Architectures in Medical Imaging

Transitioning from traditional segmentation techniques to modern advances, the field of medical image analysis has seen significant progress with the development of Fully Convolutional Networks (FCNs). These networks improved on earlier architectures by using convolutional layers throughout, thereby preserving spatial information and improving segmentation accuracy.

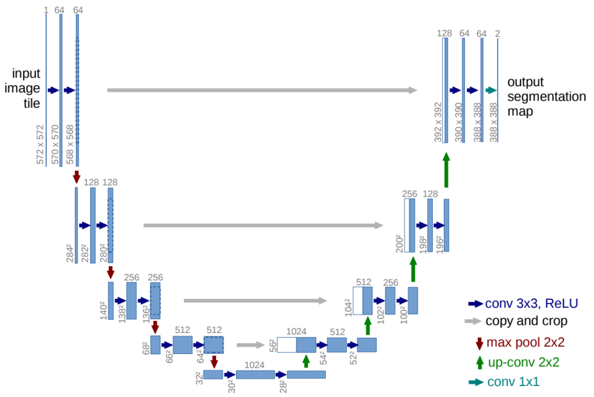

Fig. 1: U-Net architecture, where the boxes represent feature maps, and the numbers denote the number of channels and the spatial size [24].

Building on the principles of FCNs, the U-Net architecture (Fig. 1) emerged as a specifically designed framework for biomedical image segmentation [27]. It has proven versatile across various imaging modalities, including CT [13], MRI scans [39], Ultrasound [1], X-ray, OCT, and PET scans.

The symmetric structure of U-Net, characterised by its encoder-decoder design with skip connections, enables the integration of contextual information and detailed spatial localisation. This design facilitates the production of highly accurate segmentation maps, which are vital for medical diagnosis and treatment planning.

While U-Net has revolutionised medical imaging, it has faced challenges with 3D datasets and complex structures, prompting further innovations. The advent of 3D U-Net [8], for instance, uses 3D convolutional operations to work efficiently with volumetric data. The Residual U-Net [36] integrates ResNet features to combat the vanishing gradients problem, supporting the development of deeper network architectures. Attention U-Net [25] introduces mechanisms to focus on regions of interest, improving generalization by filtering out irrelevant features. These advancements have maintained U-Net's core strengths—particularly its effectiveness with limited data—while extending its capabilities to meet the demands of increasingly complex medical imaging tasks.

U-Net's introduction has been a pivotal development in the evolution of automated medical image analysis, enabling more precise and reliable identification of pathological features, thereby enhancing diagnostics and patient care. Its conceptual legacy continues to inspire the development of new architectures that further the field's progress.

D. Federated Learning Applications Across Medical Frontiers

FL is a suitable solution in health care as it addresses several issues encountered when deploying machine learning in a healthcare context. Both [34] and [12] adopt the use of FL as a solution to the challenges of working with private data that cannot be centralised. Healthcare data is generally difficult to centralise due to the private nature of data, ownership disputes, and legal issues [32]. On a practical level, centralizing large amounts of data is costly from a communication perspective [34]. As such, having each hospital train its own model on its own data, and having a global model average the weights of all the local models, allows hospitals to reap the benefits of training on a larger dataset while avoiding the associated issues. The study by Huang et al. (2019) focused on using a FL system to predict patient outcomes in intensive care units using drug features. The system used a denoising autoencoder and k-means clustering to create privacy-preserving representations of patient data. The study found that the federated learning approach was not always superior to centralised learning.

On the other hand, Thwal et al. (2021) used FL to train a medical diagnosis system using data from multiple clients. Their system used an encoder and decoder network, each containing two deep recurrent neural networks. Unlike the study by Huang et al., no clustering was performed. The study found that the global federated learning model was able to achieve comparable performance to centralised models.

Both studies demonstrated the potential of FL in healthcare settings, particularly in terms of reducing data communication costs and preserving data privacy.

The pioneering application of FL in the context of MIS was demonstrated by Sheller, Reina, Edwards, Martin and Bakas who trained a U-Net Convolutional Neural Network (CNN) on the BraTS dataset for brain tumour segmentation using the FL training technique [32]. The study also compares the performance of FL with institutional incremental learning (IIL) and cyclic institutional incremental learning (CIIL) and concludes on the superior performance of FL. The implementation of the U-Net model takes a single channel image as input and outputs a binary mask. The binary mask classifies each pixel as normal brain tissue or tumour tissue. The study implemented an averaging algorithm that weighted each client's parameter updates by the size of the local dataset often called FedAvg originally presented by McMahan[22]. The hyperparameters used in the study are as follows: epochs per round: 1, clients per round: 100%, which are reused in our experiments. The centralised U-Net that acted as a baseline in the study [32] took 3 epochs to train to a validation DC of 0.862. This study's significance lies in its ability to achieve nearly equivalent performance to a centralised approach, with FL reaching 98.7% of the centralised validation Dice. Such findings underscore the potential of FL to enable collaborative learning across different healthcare institutions, paving the way for widespread adoption of machine learning models that respect the privacy and autonomy of patient data.

III. Research Methodology

This section aims to present the relevant datasets and a detailed discussion of the methodology to be carried out. The relevant model architecture and training technique are also described in detail.

A. Research Design

This research endeavors to empirically substantiate or refute the hypothesis concerning the efficacy of the U-Net architecture within a FL (FL) framework. The primary research question investigates how the segmentation performance of FL-trained U-Nets, with a spectrum of client numbers, compares to those of centrally trained U-Nets, particularly in light of the computational constraints and data distribution intricacies inherent in FL setups. This inquiry further extends to examining whether a leaner U-Net model, with fewer feature maps, can maintain segmentation accuracy while offering computational benefits in a federated context.

B. Data

Classical machine learning performance is greatly influenced by the grade and size of the available dataset on which the model is trained [17]. Models perform better when trained on a larger dataset [10]. However, in the domain of healthcare, large datasets are rare due to the private nature of the data. Much effort has been put into collecting data sets for the purpose of training machine learning models for various medical purposes. Some of these include the Cancer Imaging Archive [9], A Lung Image Dataset with Pathological Information for Lung Cancer Screening [29] and Brain and Tumour Segmentation Dataset (BraTS), the latter of which is used in this research paper. Although large datasets are available for training machine learning models, federated learning offers unique advantages such as the ability for continuous learning, and furthermore FL is useful when centralising a dataset is not possible. Federated Learning systems can be designed to continually update the global model as new data become available [34].

1) The Brain and Tumour Segmentation Dataset: The BraTS (Brain and Tumour Segmentation) dataset comprises MRI scans of brain tumour patients acquired using multiple imaging modalities, obtained from various clinical institutions, as well as associated ground truth annotations for tumour subregions [23, 3, 2]. The subregions of each scan are labelled with the following labels: i) invaded tissue, ii) solid and cystic tumour core, and iii) enhancing tumour. The dataset is part of an initiative to evaluate and advance methods for brain tumour segmentation. Many papers have used this data set to train brain tumour segmentation models in centralised training such as [17] and in decentralised training [32].

Z-Score Normalisation: To ensure that our model is not unduly influenced by the scale of input features and to aid in faster convergence during training, we apply Z-score normalisation. This process adjusts the data such that it has a mean of zero and a standard deviation of one, ensuring consistency across all slices.

Label Modification: To provide a more holistic understanding of the tumourous regions within the MRI data, we modify the label annotations. Instead of segmenting different subregions of the tumour, we amalgamate them to represent the entire tumour. This simplifies the segmentation task and focusses on the presence and location of the tumour as a whole.

C. Data Augmentation

Data augmentation plays a pivotal role in improving the model's generalisation capabilities by introducing variability in the training data.

Random Flipping and Rotations: Each MRI slice and its corresponding mask are subjected to random horizontal and vertical flipping with a 50% probability. Additionally, with the same probability, the slices may undergo a 90-degree rotation. These augmentations introduce variability in the spatial orientations of the data, enabling the model to recognise tumours in various positions and orientations.

Random Cropping: To further enhance the diversity of the training data, the slices are randomly cropped with a 50% probability. This cropping process ensures that the model is trained on different regions of the MRI slice, making it robust against variations in tumour position within the brain.

By incorporating these data-augmentation techniques, we aim to train a model that is more robust and less prone to overfitting, ensuring improved performance on unseen data.

D. Data Split and Distribution

The data is partitioned into 60-20-20 splits for training, validation, and testing, respectively. This partitioning remains consistent for both centralised training and the FL experiments. In the FL experiments, the training dataset is distributed randomly among the clients, resulting in each client obtaining roughly 131, 66, and 33 3D brain MRI scans for the two, four, and eight client setups, respectively. It should be noted that the total amount of data trained on is consistent across each FL experiment, and as a result clients participating in larger FL setups have less data. Care is taken not to split a single 3D volume's slices between clients. In this way, within a client, we have non-Independent and Identically Distributed (IID) data since they are dependent and come from the same underlying distribution (the same patient). Across clients, the data is roughly IID since each client receives data from a diverse set of patients, ensuring a broad representation of the overall data distribution. This random assignment of entire 3D volumes to clients minimises the risk of any client specialising in a narrow subset of the data, thus promoting a generalised model during the federated aggregation process.

E. Methods

This section discusses the steps that are taken to obtain empirical evidence to accept or reject the hypothesis and to answer research questions.

F. Training U-Net with centralised data

The U-Net has achieved high performance on many MIS tasks [37], and is therefore the segmentation model of choice for this research. The U-Net architecture is characterised by a symmetric U-shaped design [27]. This U shape comprises two sections, an encoder path and a decoder path. The contracting path in the encoder uses a series of convolutional layers followed by max-pooling layers. This allows it to capture context and gradually reduce spatial dimensionality. The expanding path in the decoder uses a series of up-convolutions and concatenations with feature maps from the contracting path (slip connections), which provide localisation cues. The model outputs a mask by which each pixel is classified in a binary manner. The architecture can be understood better with the help of Fig. 1 presented by [27].

Initially, the U-Net is trained in a centralised manner using the BraTS dataset. This centralised training serves as the baseline for our study, providing a point of reference against which the performance of various FL settings can be compared. In refining our methodology, the configuration of the U-Net model required careful consideration, particularly when contrasting our approach to Sheller's study, which used a U-Net with 32 feature maps [33]. In our centralised setup, we evaluated two U-Net variants, one with 16 feature maps and the other with 32, to determine their segmentation efficiency. Remarkably, both configurations demonstrated near-identical performance, each approaching an 80% Dice. Given the computational limitations commonly encountered in federated learning (FL) scenarios, we chose the model with 16 feature maps for its reduced complexity, cutting the parameter count from Sheller's 7.76 million to a more efficient 1.94 million. This decision was not only aimed at reducing communication overhead, an important consideration in a healthcare domain, but was also underpinned by the hypothesis that a smaller model could maintain similar performance levels in an FL setup—a hypothesis that our subsequent FL experiments sought to validate.

G. Federated Learning Framework

Transitioning from centralised training, we then adopt a Federated Learning approach for the U-Net. Our FL framework is constructed using Flower [5]. In this setup, local models independently train on their respective partitions of the dataset for one epoch. Subsequently, these models transmit their learnt weights to a global model, where they are aggregated to a global model as seen in Fig. 2. The aggregated weights are then sent back to each client. This process describes a round of federated learning. This research uses the Federated Averaging algorithm [22] which takes the form:

$$w_{t+1} = \sum_{k=1}^{K} \left(\frac{n_k}{n}\right) w_{t+1}^{k} \tag{1}$$

Where:

- w_{t+1} is the global model weights at time $t+1$.
- K is the number of clients.
- n_k is the number of data samples on client k.
- n is the total number of data samples across all clients.
- w_{t+1}^{k} is the local model weights for client k at time $t+1$.

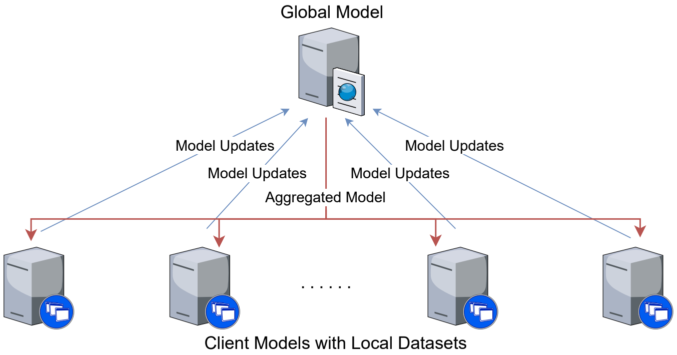

Fig. 2: The FL framework with U-Net local models

H. Evaluation

Evaluation will take place along two axes: the performance of the global model according to various metrics including the Dice [20], and the convergence rate of the FL system.
The Dice metric evaluates the similarity between a predicted segmentation mask and the labelled mask. Dice reflects the precision as well as the completeness of the segmentation results. A higher Dice score signifies greater accuracy where a score of 100% indicates a perfect agreement between the

predicted mask and the ground truth mask. It is calculated as follows:

$$\text{Dice} = \frac{2|P \cap M|}{|P| + |M|} \tag{2}$$

where P and M represent the prediction of the model and the mask of ground truth, respectively.

The Soft Dice metric evaluates the similarity between a predicted segmentation mask and the labelled mask, similar to the Dice metric. However, instead of using hard binary values, Soft Dice works with the probabilistic outputs, allowing for a smoother and more differentiable measure. This makes it particularly useful for training deep learning models using gradient-based optimization. It's important to note that the Soft Dice metric is often lower than the hard Dice score. This is because while the hard Dice metric operates on binary decisions (either a pixel is part of the object or not), Soft Dice incorporates the model's uncertainties and thus penalizes predictions that are not close to 0 or 1 when they should be. A higher Soft Dice score signifies greater accuracy, with a score of 100% indicating a perfect agreement between the predicted probability map and the ground truth mask. It is calculated as follows:

$$\text{Soft Dice} = \frac{2\sum_i P_i M_i}{\sum_i P_i + \sum_i M_i} \tag{3}$$

where P represents the probabilistic predictions of the model and M represents the ground truth mask. The sums run over all pixels i in the images.

In order to evaluate the convergence rates, we calculate and store the metrics of the aggregated global model after each training round with respect to an independent validation dataset.

I. Loss function

The Dice Loss is the loss function minimised in this research, which is commonly paired with the Dice metric [16]:

$$DL(P, M) = 1 - \frac{2|P \cap M| + 1}{|P| + |M| + 1} \tag{4}$$

Here, the numerator and denominator are increased by 1 so that the function is not undefined in edge cases. The Dice loss function is selected due to its proficiency in handling segmentation challenges presented by imbalanced datasets, a frequent occurrence in medical imaging. Traditional loss functions can prioritise the dominant class, often neglecting smaller regions of interest. In contrast, Dice measures the overlap between predicted segmentation and ground truth, treating both classes with equal importance. This focus on balanced segmentation makes Dice a preferred choice for many medical imaging applications.

J. Comparisons

The performance metrics of all models will be compared with each other to determine the relative effectiveness and to highlight the differences between models trained on centralised data versus those trained with Federated Learning. The convergence rates of the FL system with local U-Nets with various numbers of participating clients will be compared to each other to show the effect of the number of clients on the overall training dynamics and system efficiency. We will also analyse the number of rounds required for each system to attain different percentiles of accuracies based on the centrally trained baseline U-Net. Systems that achieve higher accuracy in fewer rounds will be said to have better convergence rates comparatively. We will also be interested in observing the stability of the systems over the course of their training. This stability can be assessed by examining the variance in performance metrics across training rounds.

IV. Results and Analysis

A. Training

The experimental setup is designed to evaluate the performance of the U-Net model under both centralised and FL conditions. This section delineates the training protocols, model evaluation criteria, and specific configurations used during the experiments.

Data Partitioning: For FL scenarios, the BraTS dataset is divided among the participating clients to simulate a distributed data environment. The partitioning is performed to reflect a realistic scenario where different clients, representing different clinics or hospitals, have access to distinct subsets of data.

Model Configuration: The U-Net model architecture is configured with 16 feature maps at each layer to maintain computational efficiency and feasibility within the FL framework. This configuration aligns with the findings from preliminary centralised training, which indicated that a smaller model does not significantly compromise segmentation accuracy compared to a larger variant.

Optimisation and Evaluation Metrics: A comprehensive set of metrics, including loss, Dice, soft Dice, intersection over union (IOU), accuracy, precision, and specificity, are employed to assess model performance rigorously. These metrics provide a multifaceted view of the model's segmentation capabilities, ensuring a thorough evaluation.

Model Training and Validation: Training is conducted with careful consideration of overfitting and underfitting phenomena. A validation set is used to monitor the model's generalisability and to facilitate early stopping, thereby preventing overfitting to the training data.

Following this setup, the specific training and evaluation details are as follows:

For the training process, Adam Optimiser is used with an initial learning rate of 0.0001, and dropout regularisation is incorporated.

Epochs and Rounds: The centralised U-Net is trained for 30 epochs. The FL clients are trained for one epoch per FL round. All FL simulations are run for 200 rounds.

Early Stopping: Early stopping is implemented based on the Dice validation loss, with patience of 10 epochs for the centralised U-Net and 60 FL rounds for the FL simulations. The model weights retained and subsequently tested are those corresponding to the epoch (or round) with the lowest validation loss. In order to correctly obtain the best model in the FL simulations, early stopping had to be implemented based on a moving average with a ten-point window. This approach was necessary due to the pronounced fluctuation in validation loss from one epoch to the next. The centralised U-Net training ran for 5.3 hours. The FL experiments with two, four, and eight clients took 14.80 hours, 14.52 hours, and 10.37 hours, respectively. The experiments were run on a pair of Quadro RTX 8000 GPUs. The experiments with two and eight clients were run with four concurrent clients, and the experiment with four clients was run with two concurrent clients, as dictated by resource availability.

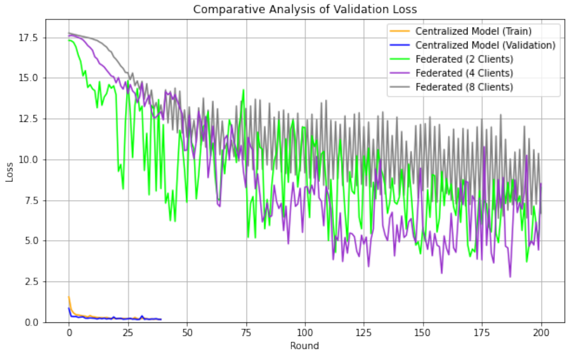

Fig. 3: Validation Loss with Moving Average for both centralised and FL models

From Fig. 3 we see that the FL models have much more variance than the central model. Furthermore, we see that an increase in the number of clients results in a decrease to the convergence stability.

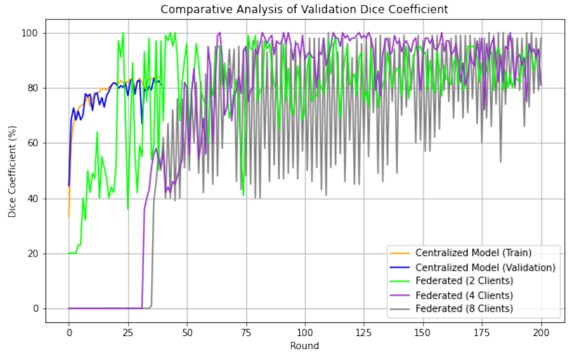

Fig. 4: Dice Coefficient for Centralised and Federated Models

It is interesting to note that in Fig. 4 the aggregated models with four and eight clients take several FL rounds before being able to improve their DC scores, which is consistent with previous research [33]. Comparing the DC of the centralised model and the DC of the FL models, we see that the U-Net trained on centralised data has already achieved state-of-the-art performance before the aggregated model has shown any signs of learning.

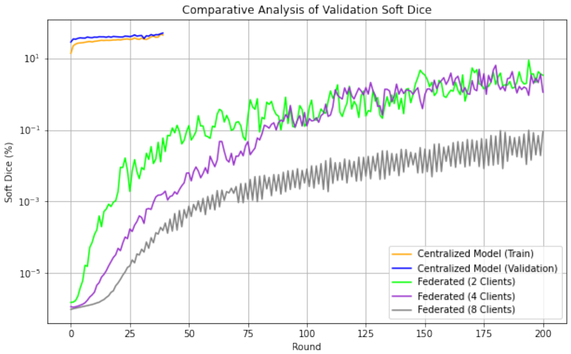

Fig. 5: Soft Dice for Centralised and Federated Models

As the number of clients increases, we also notice a decrease in the ability of the aggregated model to improve the dice and soft dice scores. In order to determine the reason for this we look at the individual client's loss during training. When looking at Fig. 7, it is clear that client zero is unable to learn as well as client one, round after round. Given that all clients use identical local model configurations, the observed variance in learning performance likely stems from the nature of the data subsets received by each client, which may not be comprehensive enough to facilitate adequate learning. Although the dataset on each client is assumed to be roughly IID, random partitioning can inadvertently lead to sample sets that do not capture the comprehensive statistical properties of the global dataset, thus impeding a client's ability to learn generalisable features effectively. This insight suggests a potential revision of the initial IID assumption. The disparity in performance across clients persists as the number of participants in the experiments increases, as detailed in the Appendix. A possible solution to the issue of struggling clients is to use an adaptive aggregation strategy where the contribution of each client to the global model is weighted based on a performance metric evaluated on a validation set, which could be held centrally. This metric might include factors such as the client's loss on the validation set or the improvement of the global model's performance when including the client's update. Clients that consistently perform poorly would have their updates down-weighted, thereby minimising their impact on the global model. Conversely, clients that perform well would have a greater influence. When analyzing the 2-client setup, it is observed that each client individually shows an improvement in their soft dice scores, reaching approximately 30% and 20%, respectively, as shown in 6. However, when these individual client models are aggregated, the soft dice score of the resulting global model on the validation set is considerably lower throughout the training process. This discrepancy suggests that the specific features and the predictive confidence captured by each local model

TABLE I: Quantitative results showing the performance of all models on held out test data

	Loss	Dice	Soft dice	IOU	Accuracy	Precision	Specificity
Centralised	**0.1982**	**0.8034**	**0.3918**	**0.7593**	**0.9870**	**0.8732**	0.9948
2 Clients	3.8979	0.6915	0.2680	0.6485	0.9790	0.7250	0.9870
4 Clients	3.1142	0.7637	0.2560	0.7211	0.9812	0.8653	**0.9950**
8 Clients	4.5907	0.7312	0.2260	0.6872	0.9788	0.8307	0.9938

do not combine effectively during aggregation, resulting in a loss of generalisation when the global model is evaluated on the validation set. This could imply that, while each client is learning useful patterns within their own data subsets, these patterns may not be complementary or may even be conflicting, leading to a decrease in the aggregated model's performance on unseen data.

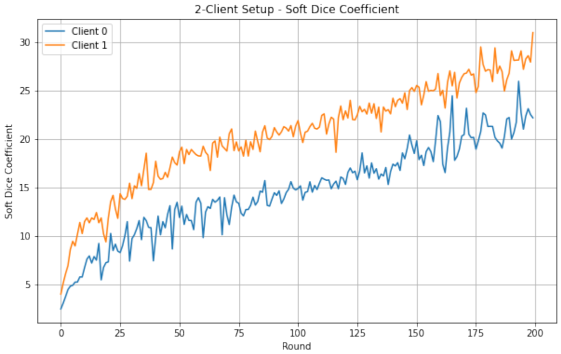

Fig. 6: Loss for Two Client Setup

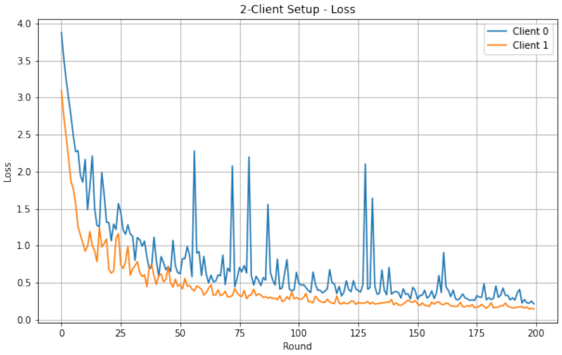

Fig. 7: Soft Dice for two Client Setup

We notice that across all metrics, FL models with more clients have slower convergence.

As we increase the number of clients and hence distribute the data even further, we notice that this affects convergence as each model has less data to train on.

Limitation: In a typical setting, increasing the number of participating clients will typically also increase the amount of data in the system since each client will be based at a clinic that has its own scans. In this regard, the research is limited by the dataset size as we were not able to keep the amount of data consistent per client as the experiments scaled the number of clients. This limitation unfairly penalises setups with more clients and has a profound impact segmentation performance.

B. Testing

1) Quantitative Results: The final models are evaluated on the held-out test dataset using several metrics, namely, Dice score, soft dice, intersection over union (IOU), accuracy, precision, and specificity. The following equations give the definitions of IOU, accuracy, precision, and specificity:

$$IOU = \frac{|P \cap M|}{|P \cup M|} \tag{5}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

Where:

- P is the set of pixels predicted by the model to belong to the segmentation class.
- M is the set of pixels actually belonging to the segmentation class according to the ground truth (manual segmentation).
- TP (True Positives) is the count of pixels correctly identified as belonging to the segmentation class.
- TN (True Negatives) is the count of pixels correctly identified as not belonging to the segmentation class.
- FP (False Positives) is the count of pixels incorrectly identified as belonging to the segmentation class.
- FN (False Negatives) is the count of pixels that belong to the segmentation class but were not identified by the model.

The table I presents a comparative analysis of model performance metrics across different setups: centralised and decentralised with 2, 4, and 8 clients. A notable observation is the superior performance of the centralised model across all metrics, emphasising the often more straightforward optimisation landscape of centralised training, which is consistent with previous research [33]. The loss values increase as the client number increases, hinting at the challenges associated with coordinating and aggregating model updates across a distributed setup. Despite an increase in loss, the Dice score, a measure of model accuracy, remains relatively stable though slightly reduced with an increase in client number. However, the Soft Dice score demonstrates a downward trend as client number increases, potentially indicating challenges with model generalization or data heterogeneity among clients. The IOU (Intersection over Union) scores also exhibit a decline with increasing client number, reflecting a diminishing overlap between the predicted and actual positive classes. On the other hand, accuracy, precision, and specificity metrics remain relatively high across all setups, although there is a subtle decline with increasing client numbers. Collectively, these data hint at the nuanced trade-offs involved in FL architectures. While decentralisation empowers models to learn from distributed data sources, the accompanying challenges in model aggregation, communication overheads, and potential data skewness among clients might compromise the model's performance on certain metrics. Hence, the insights from this table could serve as a valuable guide for optimising FL strategies, balancing the benefits of decentralised data learning against the imperatives of model performance and accuracy. Upon comparison with the findings from [33], it is observed that the Dice score of the centralised model (80%) is within five percent of the centralised model reported in their research. Furthermore, in their study, the four- and eight client configurations reported Dice scores marginally approaching or exceeding the score of their centralised model by 99. 9% and 100. 2%, respectively. In contrast, within our research, the four- and eight-client setups attain 95.1% and 91% of our centralised model's Dice score, respectively. The observed decrease in performance can likely be ascribed to the use of a more compact U-Net model.

2) Qualitative Results: The visual assessment reveals segmentation results across varying model configurations. Figure 8 shows 2D slices for which all models demonstrate a high level of segmentation accuracy, which attests to their robustness in typical imaging scenarios. However, when smaller tumours are considered, such as in Fig. 9, the models' performance reflects a significant variation. Notably, the two-client aggregated model shows a distinct improvement in managing these complex cases, suggesting that the aggregation of insights from a smaller subset of clients may result in a more finely-tuned model within the federated learning framework. This phenomenon indicates a key limitation in the design of the study: increasing the number of clients does not equate to a proportional increase in the diversity or volume of training data, as each client is effectively learning from a smaller and potentially less varied subset of the total dataset.

On the contrary, the penultimate row illustrates that the advantage of a two-client configuration over a centralised model is not consistent, particularly for the detection of less common tumour features. This variation underscores a critical limitation of FL in medical imaging: as the number of clients grows, the quantity of data available to each client diminishes. This may not may not provide a comprehensive enough learning experience to consistently outperform centralised models. It also suggests that the mere expansion of the federated network, without a corresponding increase in data diversity and volume, may not yield the expected improvements in model performance.

This insight is crucial for the deployment of FL models in MIS. It implies that beyond a certain point, simply adding more clients to the network cannot compensate for the constrained data each client receives, which is a fundamental deviation from typical FL scenarios where more clients usually mean access to more data. Future research should thus explore mechanisms to ensure that increasing the number of clients in federated learning does not compromise the richness of data necessary for the nuanced task of tumour segmentation.

V. Conclusion

In summarising the outcomes of this study, we have critically analysed the performance of a streamlined U-Net model within a Federated Learning (FL) framework for segmenting brain tumors from medical images. The research question probed whether a U-Net model with fewer feature maps, optimised for FL, could rival the segmentation accuracy of a centrally trained counterpart. While our streamlined U-Net model proved capable within the FL framework, the results indicate that its performance as an aggregated global model is influenced by the diversity of data across clients, which presents a challenge to achieving the equivalent accuracy of centralised training.

The study successfully achieved its primary objective of training and evaluating a compact U-Net model, establishing a baseline for performance within a centralised framework. Subsequent objectives aimed at evaluating the model's performance in an FL environment and comparing it against a standard U-Net model were also met. The analysis revealed that the performance of FL models, particularly in configurations with a few clients, holds considerable promise, although it is not yet equivalent to centralised models. The decrease in performance with an increasing number of clients highlights the challenges in data distribution and model generalisation across a federated network.

Addressing the research problem, it is evident that the hypothesis, asserting that a smaller U-Net would maintain satisfactory accuracy, is partially supported. The streamlined model demonstrated adequate segmentation capabilities, albeit with some performance drop, suggesting that while the goal of computational efficiency was met, the trade-off in segmentation accuracy needs to be carefully managed.

Future work should focus on investigating dynamic aggregation strategies that weigh individual client contributions more effectively. An expansion of the dataset size and diversity would likely yield deeper insights into the performance and scalability of FL in medical imaging. Finally, implementing and testing these adaptive strategies in real-world settings is crucial for validating the practical applicability of the research findings, with the ultimate aim of enhancing FL systems for healthcare applications and addressing the urgent need for accurate brain tumour segmentation.

A. Acknowledgments

We extend our gratitude to the Wits Cluster staff for their support and the provision of the high-performance computing

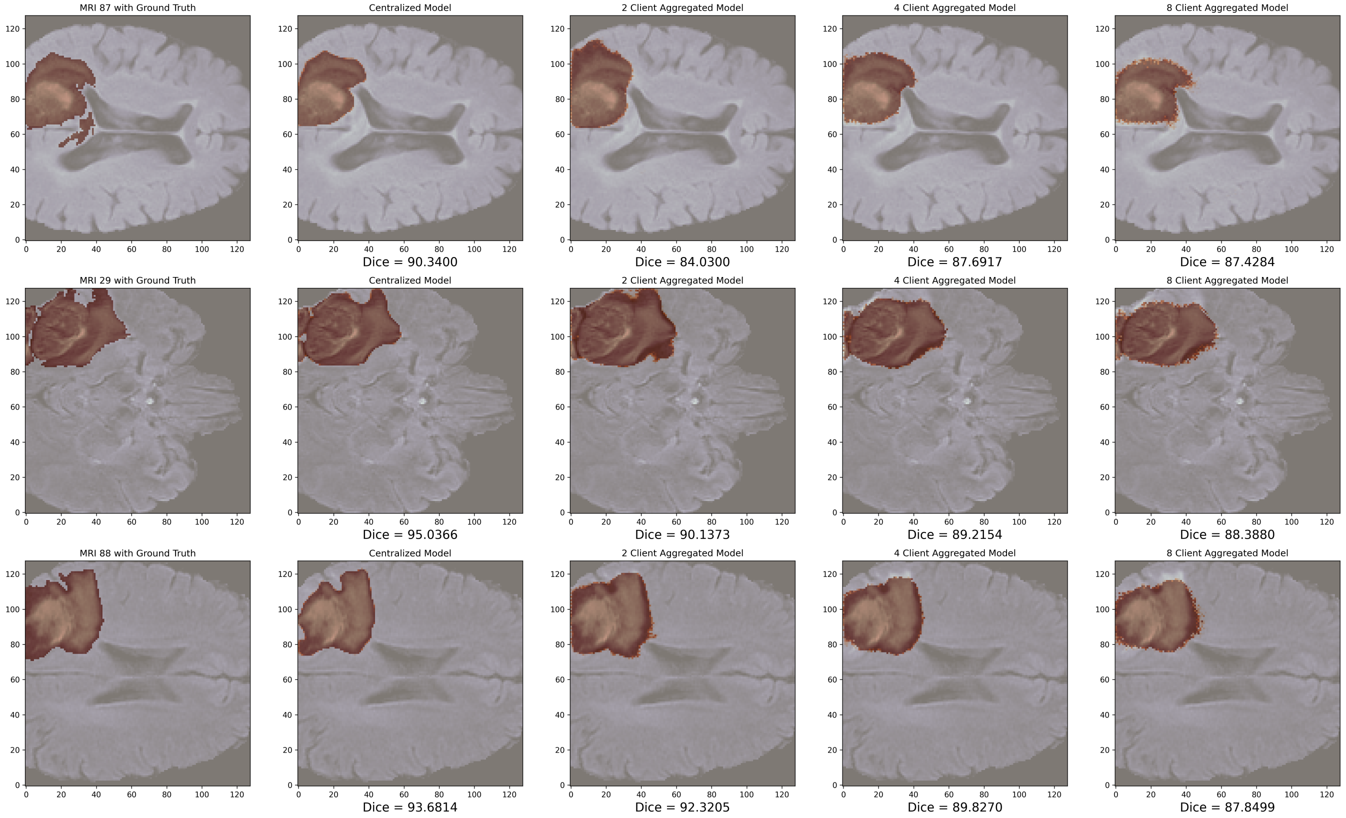

Fig. 8: Qualitative Comparison showing samples where model performance is high.

infrastructure. Additionally, we express our appreciation to the developers of the Flower framework for their robust and versatile tool, which has laid the foundation for our work.

References

[1] Nabila Abraham and Naimul Mefraz Khan. "A novel focal tversky loss function with improved attention u-net for lesion segmentation". In: *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE. 2019, pp. 683–687.

[2] Spyridon Bakas et al. "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific data* 4.1 (2017), pp. 1–13.

[3] Spyridon Bakas et al. "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection". In: *The cancer imaging archive* 286 (2017).

[4] Jeroen Bertels et al. "Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer. 2019, pp. 92–100.

[5] Daniel J Beutel et al. "Flower: A friendly federated learning research framework". In: *arXiv preprint arXiv:2007.14390* (2020).

[6] Alceu Bissoto et al. "Deep-learning ensembles for skin-lesion segmentation, analysis, classification: RECOD titans at ISIC challenge 2018". In: *arXiv preprint arXiv:1808.08480* (2018).

[7] Ying Chen et al. "A lung dense deep convolution neural network for robust lung parenchyma segmentation". In: *IEEE Access* 8 (2020), pp. 93527–93547.

[8] Özgün Çiçek et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer. 2016, pp. 424–432.

[9] Kenneth Clark et al. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository". In: *Journal of digital imaging* 26 (2013), pp. 1045–1057.

[10] Alon Halevy, Peter Norvig, and Fernando Pereira. "The unreasonable effectiveness of data". In: *IEEE intelligent systems* 24.2 (2009), pp. 8–12.

[11] Guy S Handelman et al. "Peering into the black box of artificial intelligence: evaluation metrics of machine

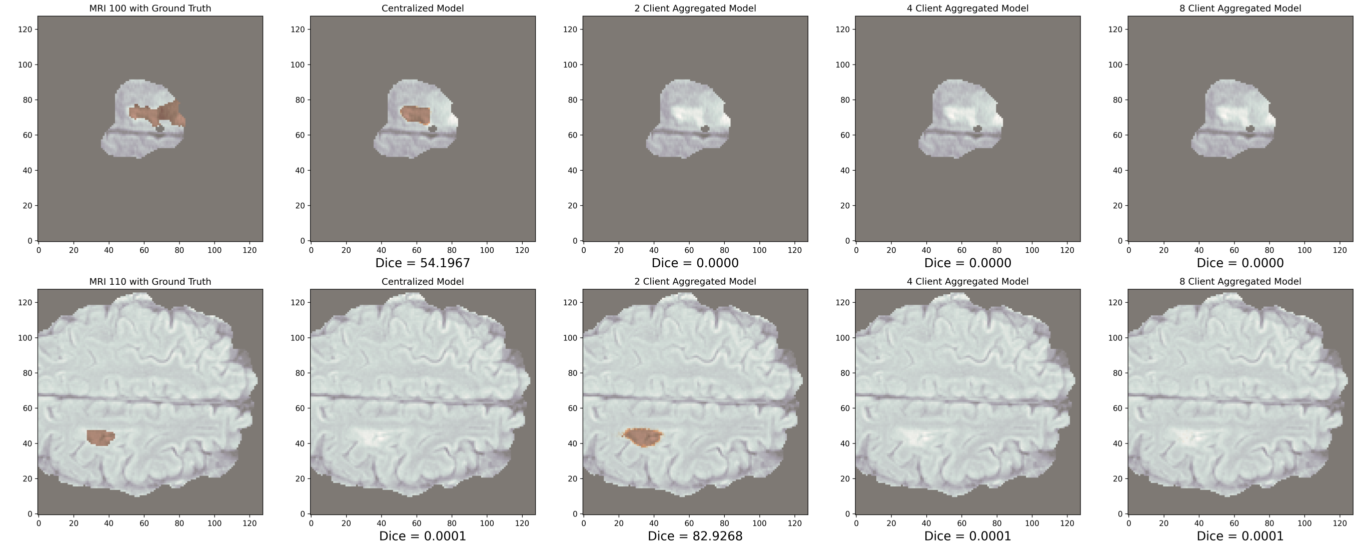
MRI 100 with Ground Truth
Centralized Model
2 Client Aggregated Model
4 Client Aggregated Model
8 Client Aggregated Model
Dice = 54.1967
Dice = 0.0000
Dice = 0.0000
Dice = 0.0000
MRI 110 with Ground Truth
Centralized Model
2 Client Aggregated Model
4 Client Aggregated Model
8 Client Aggregated Model
Dice = 0.0001
Dice = 82.9268
Dice = 0.0001
Dice = 0.0001

Fig. 9: Qualitative Comparison showing samples where model performance is poor.

learning methods". In: *American Journal of Roentgenology* 212.1 (2019), pp. 38–43.

[12] Li Huang et al. "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records". In: *Journal of biomedical informatics* 99 (2019), p. 103291.

[13] Qing Huang et al. "Robust liver vessel extraction using 3D U-Net with variant dice loss function". In: *Computers in biology and medicine* 101 (2018), pp. 153–162.

[14] Umit Ilhan and Ahmet Ilhan. "Brain tumor segmentation based on a new threshold approach". In: *Procedia computer science* 120 (2017), pp. 580–587.

[15] Elisee Ilunga-Mbuyamba et al. "Localized active contour model with background intensity compensation applied on automatic MR brain tumor segmentation". In: *Neurocomputing* 220 (2017), pp. 84–97.

[16] Shruti Jadon. "A survey of loss functions for semantic segmentation". In: *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*. IEEE. 2020, pp. 1–7.

[17] Haozhe Jia et al. "H2NF-Net for Brain Tumor Segmentation Using Multimodal MR Imaging: 2nd Place Solution to BraTS Challenge 2020 Segmentation Task". In: *Lecture Notes in Computer Science* (2021), pp. 58–68. ISSN: 1611-3349. DOI: 10.1007/978-3-030-72087-2_6. URL: http://dx.doi.org/10.1007/978-3-030-72087-2_6.

[18] Adel Kermi, Khaled Andjouh, and Ferhat Zidane. "Fully automated brain tumour segmentation system in 3D-MRI using symmetry analysis of brain and level sets". In: *IET Image Processing* 12.11 (2018), pp. 1964–1971.

[19] Shah Rukh Khan et al. "IoMT-based computational approach for detecting brain tumor". In: *Future Generation Computer Systems* 109 (2020), pp. 360–367.

[20] Jiaxuan Li et al. "Multi-scale GCN-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary OCT images". In: *Biomedical Optics Express* 12.4 (2021), pp. 2204–2220.

[21] Wen Li et al. "Automatic segmentation of liver tumor in CT images with deep convolutional neural networks". In: *Journal of Computer and Communications* 3.11 (2015), p. 146.

[22] Brendan McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[23] Bjoern H Menze et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[24] David Ojika et al. "Addressing the memory bottleneck in AI model training". In: *arXiv preprint arXiv:2003.08732* (2020).

[25] Ozan Oktay et al. "Attention u-net: Learning where to look for the pancreas". In: *arXiv preprint arXiv:1804.03999* (2018).

[26] Nobuyuki Otsu. "A threshold selection method from gray-level histograms". In: *IEEE transactions on systems, man, and cybernetics* 9.1 (1979), pp. 62–66.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[28] Natalia Salpea, Paraskevi Tzouveli, and Dimitrios Kollias. "Medical image segmentation: A review of modern architectures". In: *Computer Vision–ECCV 2022 Work-*

shops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII. Springer. 2023, pp. 691–708.

[29] Yanbo Shao et al. “Lidp: A lung image dataset with pathological information for lung cancer screening”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*. Springer. 2022, pp. 770–779.

[30] Muhammad Sharif et al. “An integrated design of particle swarm optimization (PSO) with fusion of features for detection of brain tumor”. In: *Pattern Recognition Letters* 129 (2020), pp. 150–157.

[31] Micah J Sheller et al. “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data”. In: *Scientific reports* 10.1 (2020), pp. 1–12.

[32] Micah J Sheller et al. “Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation”. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*. Springer. 2019, pp. 92–104.

[33] Micah J Sheller et al. “Multi-institutional Deep Learning Modeling without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation, CoRR abs/1810.04304”. In: *arXiv preprint arXiv:1810.04304* (2018).

[34] Chu Myaet Thwal et al. “Attention on personalized clinical decision support system: Federated learning approach”. In: *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE. 2021, pp. 141–147.

[35] Virupakshappa and Basavaraj Amarapur. “Cognition-based MRI brain tumor segmentation technique using modified level set method”. In: *Cognition, Technology & Work* 21.3 (2019), pp. 357–369.

[36] Guotai Wang et al. “Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation”. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*. Springer. 2019, pp. 61–72.

[37] Risheng Wang et al. “Medical image segmentation using deep learning: A survey”. In: *IET Image Processing* 16.5 (2022), pp. 1243–1267.

[38] Rammah Yousef et al. “U-Net-Based Models towards Optimal MR Brain Image Segmentation”. In: *Diagnostics* 13.9 (2023), p. 1624.

[39] Ziang Zhang et al. “DENSE-INception U-net for medical image segmentation”. In: *Computer methods and programs in biomedicine* 192 (2020), p. 105395.

[40] Zihao Zhao et al. “Towards Efficient Communications in Federated Learning: A Contemporary Survey”. In: *Journal of the Franklin Institute* (2023).

[41] Hangyu Zhu et al. “Federated learning on non-IID data: A survey”. In: *Neurocomputing* 465 (2021), pp. 371–390.

Appendix

Summary: The first seven graphs show the accuracy (Fig. 10), Dice metric (Fig. 11), soft Dice metric (Fig. 12), IOU (Fig. 13), precision (Fig. 14), specificity (Fig. 15) and loss (Fig. 16) of the baseline U-Net with 16 feature maps used in this study. These graphs show the training performance and convergence as the U-Net trains on centralised data. The next graph shows the losses of both the centralised and FL models plotted using a moving average with a window size of 10 training rounds (Fig. 17). This graph allows for an easy comparison between the performance of the centralised and FL models. The remaining graphs show the performance of individual clients within the two-client (Figs. 18,19), four-client (Figs. 20,21, 22) and eight-client (Figs. 23, 24, 25) FL experiments.

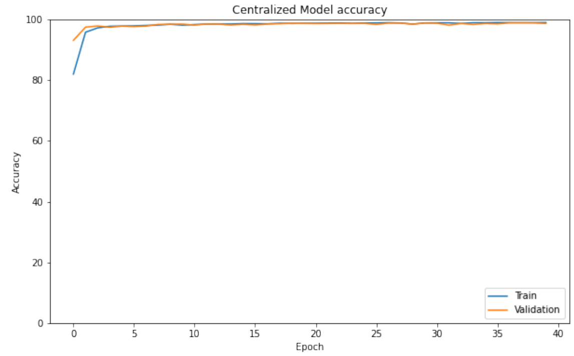

Fig. 10: Graph showing accuracy of model trained on centralised data.

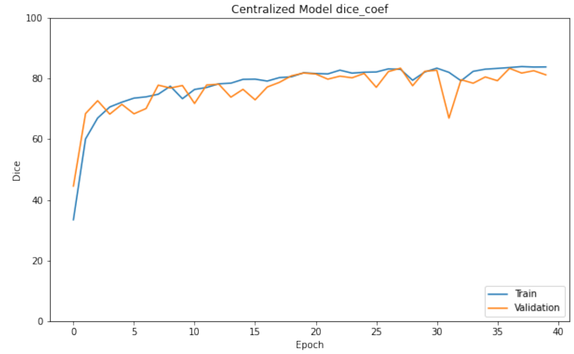

Fig. 11: Graph showing Dice of model trained on centralised data

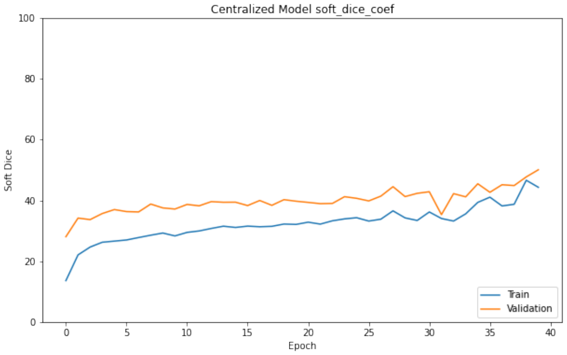

Fig. 12: Graph showing soft Dice of model trained on centralised data

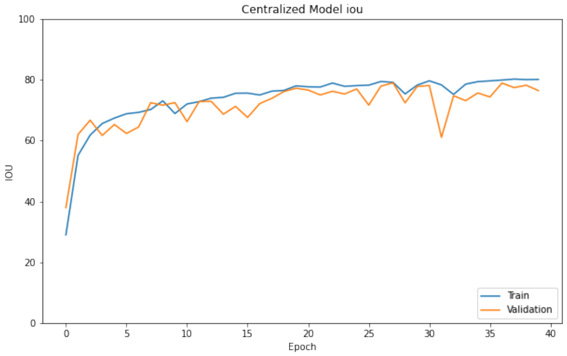

Fig. 13: Graph showing IOU of model trained on centralised data

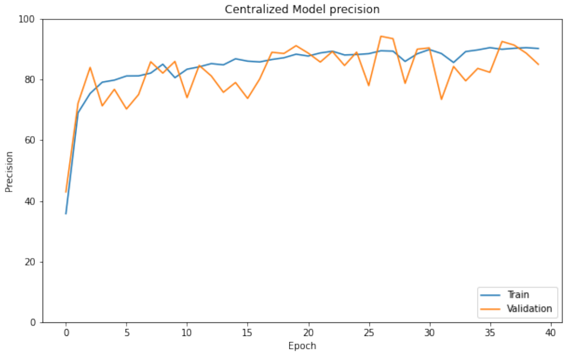

Fig. 14: Graph showing precision of model trained on centralised data

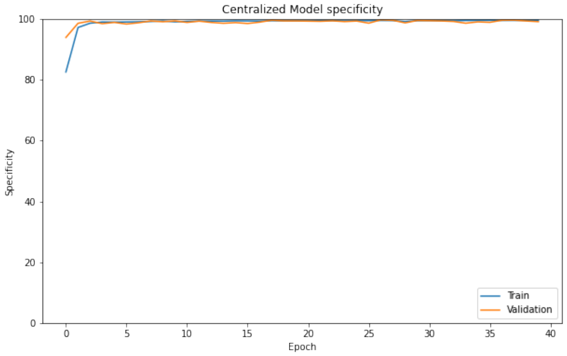

Fig. 15: Graph showing specificity of model trained on centralised data

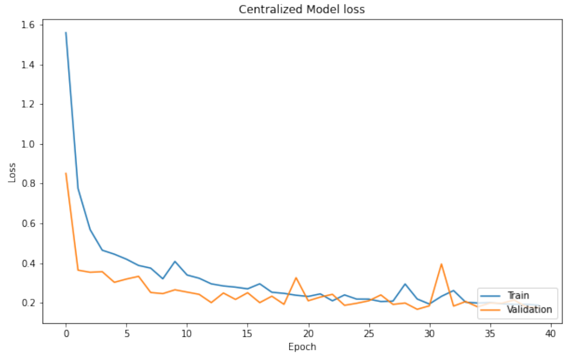

Fig. 16: Graph showing training and validation loss of model trained on centralised data

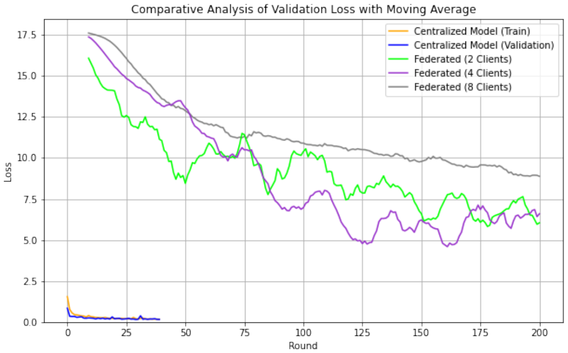

Fig. 17: Graph showing loss of centralised and FL models with moving average

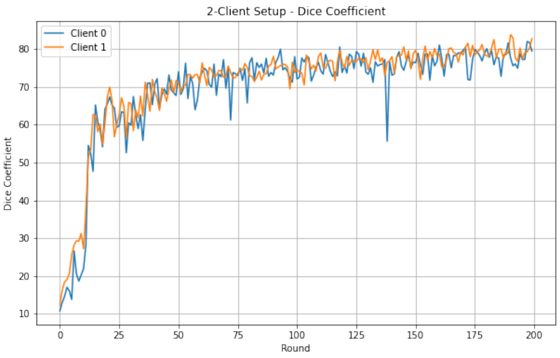

Fig. 18: Graph showing Dice of clients in a two client FL setup

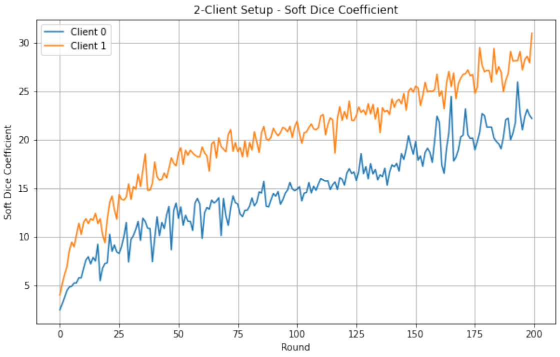

Fig. 19: Graph showing soft Dice of clients in a two client FL setup

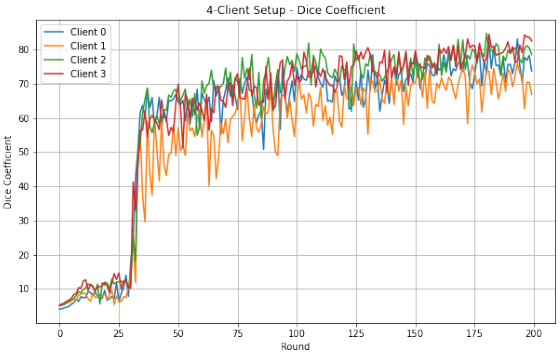

Fig. 20: Graph showing Dice of clients in a four client FL setup

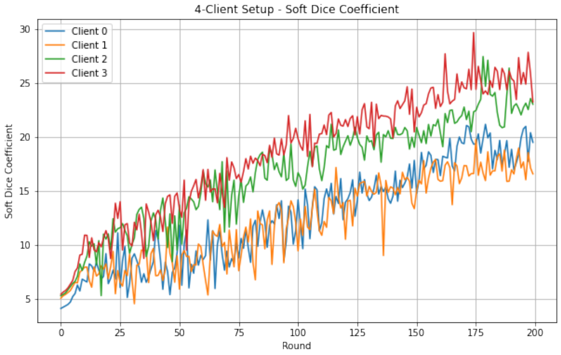

Fig. 21: Graph showing soft Dice of clients in a four client FL setup

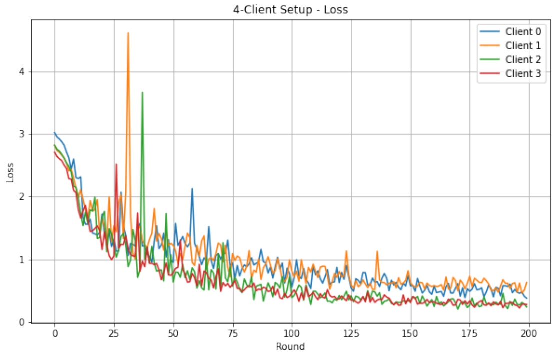

Fig. 22: Graph showing loss of clients in a four client FL setup

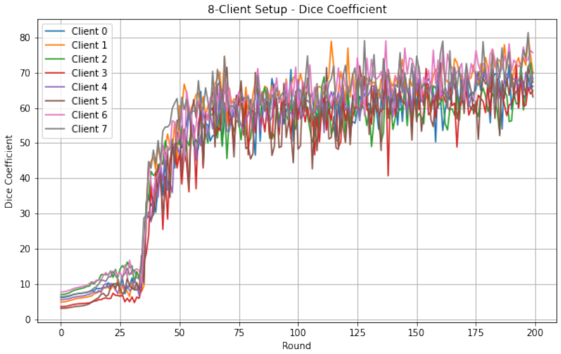

Fig. 23: Graph showing Dice of clients in an eight client FL setup

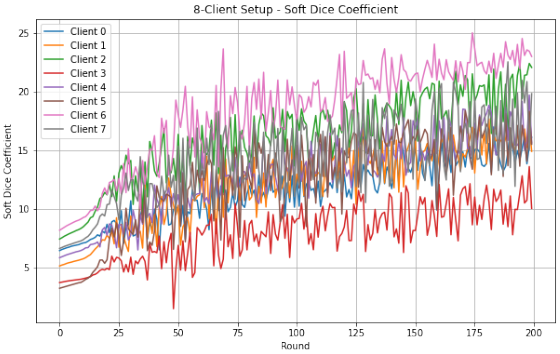

Fig. 24: Graph showing soft Dice of clients in an eight client FL setup

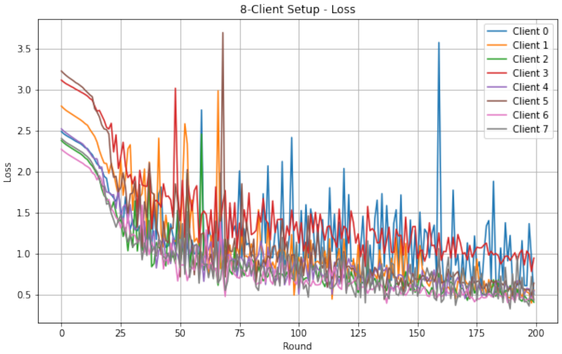

Fig. 25: Graph showing loss of clients in an eight client FL setup