Multi-institutional Travelling Model for Tumor Segmentation in MRI Datasets

Raissa Souza[1,2,3(✉)], Anup Tuladhar[1,2,3], Pauline Mouches[1,2,3], Matthias Wilms[1,2], Lakshay Tyagi[4], and Nils D. Forkert[1,2,5,6]

[1] Department of Radiology, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 4N1, Canada
raissa.souzadeandrad@ucalgary.ca
[2] Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 4N1, Canada
[3] Biomedical Engineering Graduate Program, University of Calgary, Calgary, AB T2N 4N1, Canada
[4] Department of Chemical Engineering, Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh, India
[5] Department of Clinical Neurosciences, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 4N1, Canada
[6] Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB T2N 4N1, Canada

Abstract. Administrative, ethical, and legal reasons are often preventing the central collection of data and subsequent development of machine learning models for computer-aided diagnosis tools using medical images. The main idea of distributed learning is to train machine learning models locally at each site rather than using centrally collected data, thereby avoiding sharing data between health care centers and model developers. Thus, distributed learning is an alternative that solves many legal and ethical issues and overcomes the need to directly share data. Most previous studies simulated data distribution or used datasets that are acquired in a controlled way, potentially misrepresenting real clinical cases. The 2021 Federated Tumor Segmentation (FeTS) challenge provides clinically acquired multi-institutional magnetic resonance imaging (MRI) scans from patients with brain cancer and aims to compare federated learning models. In this work, we propose a travelling model that visits each collaborator site up to five times with three distinct travelling orders (ascending, descending, and random) between collaborators as a solution to distributed learning. Our results demonstrate that performing more training cycles is effective independent of the order that the models are transferred among the collaborators. Moreover, we show that our model does not suffer from catastrophic forgetting and successfully achieves a similar performance (average Dice score 0.676) compared to standard machine learning implementations (Dice score 0.667) trained using the data from all collaborators hosted at a central location.

Keywords: Distributed learning · Travelling model · FeTS · BraTS

A. Crimi and S. Bakas (Eds.): BrainLes 2021, LNCS 12963, pp. 420–432, 2022.
https://doi.org/10.1007/978-3-031-09002-8_37

1 Introduction

In recent years, novel machine learning methods have achieved human-like performance for many problems in the computer vision and medical image analysis domains [1]. However, a major reason that prevents a broad application of these novel artificial intelligence methods for computer-aided diagnosis using medical images is the limited access to datasets due to strict legal, technical, and ethical regulations that aim to protect patient data [2–4]. For instance, regulations such as the United States Health Insurance Portability and Accountability Act (HIPAA) [5] and European General Data Protection Regulation (GDPR) [6] determine how personally identifiable data should be stored, exchanged, and manipulated, limiting healthcare facilities to share information with each other.

To overcome data access problems, a recent approach called distributed learning has emerged, which aims to locally train machine learning models at each data contributor site, thereby avoiding the need to share data at a central location. This approach is especially relevant for applications where a single collaborator does not have enough data for training accurate and generalizable models, which is especially the case for rare diseases for which only few patients are seen at a single health care center [7].

Today, distributed learning is mostly implemented using federated learning of artificial neural networks (ANN). In federated learning, independent machine learning models are trained at each collaborator site in parallel and then combined into a global model. Contrary, in this work, we develop and evaluate a so-called travelling model for distributed learning, also known as single weight transfer and cyclical weight transfer. It consists of training a single model sequentially, at one collaborator site per time, travelling from one collaborator to the next one during training [8–12]. It is hypothesized that the final model trained using a traveling approach is able to leverage knowledge from the diversity of all samples with a similar performance than corresponding models trained using a centralized database.

The 2021 Federated Tumor Segmentation (FeTS) challenge aims at comparing state-of-the-art distributed learning methods that effectively aggregate weights and leverage data from multiple collaborators, given a pre-defined segmentation algorithm. Most previous studies investigating distributed learning described in literature simulated data distribution or used datasets that are acquired in a controlled way, likely misrepresenting real clinical cases [7,9]. The FeTS challenge intends to address these issues by providing clinically acquired multi-institutional magnetic resonance imaging (MRI) scans from previous Brain Tumor Segmentation (BraTS) challenges [13–15], and data from independent collaborators that are part of FeTS initiative [16,18] to provide a common ground to compare different distributed learning approaches.

For this work, we implemented a travelling machine learning framework to solve the proposed challenge. We compared the proposed distributed learning model to a standard central learning implementation and investigated the influence of the order by which the model travels among the collaborators and the number of visits on its performance.

2 Material and Methods

2.1 Dataset

FeTS provided 341 pre-processed brain imaging datasets from patients with brain tumours including four MRI modalities: T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2-weighted Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes, resulting in a total of 1364 scans for the 341 patients. The scans were acquired across multiple sites with different protocols and scanners.

For each patient, a corresponding brain tumor segmentation is available with separate labels for the enhancing tumor (ET) tissue, core tumor (CT) tissue, and whole tumor (WT) tissue. The manual segmentation was performed by up to four raters using the same annotation protocol, described in [12], that was later approved by expert neuro-radiologists.

Two csv files describing the distribution of the datasets among 17 (partitioning one) and 22 (partitioning two) collaborators were also provided by the FeTS challenge. It is important to note that the data is not equally distributed, with some collaborators providing data for up to 50 patients while other centers provide data for less than ten patients.

Furthermore, the FeTS challenge released datasets from 111 patients containing the same MRI modalities but without ground truths labels for a validation of the distributed models on unseen data.

2.2 Method Training Aggregator

The pre-defined algorithm provided for the challenge is a U-Net segmentation architecture with residual connections. The U-Net is an encoder-decoder structure, which consists of convolutional layers and downsampling layers in the encoder and upsampling layers in the decoder. In order to improve context and feature re-usability, the defined architecture includes skip connections that concatenate feature maps paired across the encoder and the decoder layer. It also includes residual connections that take advantage of information from previous layers to boost its performance [16].

The general idea of the proposed travelling model is to train a single U-Net segmentation model at one collaborator at time. Briefly, we define the order that we want to visit the collaborators and then initiate the model training at the first collaborator of the sequence. When the training is completed at a collaborator, the model with the learned parameters is sent to the next collaborator. This process is repeated until the model reaches the end of the sequence. The whole traveling process can also be repeated for multiple cycles, which may improve the model performance and avoid catastrophic forgetting [17], which means that the model forgets patterns learned at previous collaborators. The final model is always defined as the last model generated after the last visit to all collaborators.

For model development and evaluation using the publicly available training data, the datasets were shuffled and for each collaborator 80% of the datasets

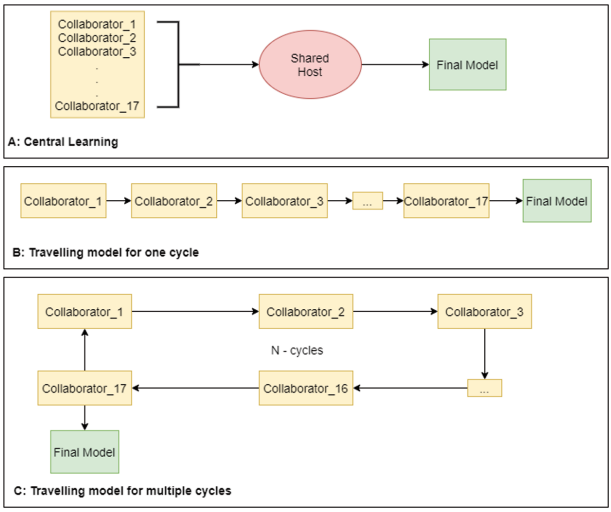

Fig. 1. Model training aggregators investigated (A) central learning, (B) single weight transfer and (C) cyclical weight transfer.

were used for training the model and 20% for internal validation of the trained model. In this work, we tested different aggregation methods and compared their performance. First, we trained the segmentation model provided by the FeTS challenge with a standard central learning approach, pooling the data to a central database (Fig. 1A). Second, we trained a travelling model for one cycle. With this method, the model visits each collaborator only one time as illustrated in Fig. 1B. Finally, as shown in Fig. 1C, we performed multiple cycle visits. This setup only differs from the previous approach by sending the model to each collaborator more than once. We investigated the model performance for 2, 3, 4, and 5 cycles in this work.

We also investigated how the order by which the model is transferred between collaborators affects its ability to learn and generalize. Therefore, we investigated three distinct transfer schemes: ascending and descending orders according to the number of samples available at each collaborator, and a random order.

We used one epoch per cycle for all experiments, which means that we used all available data at least once at each collaborator. An optimizer rate of 0.05 was used during training.

2.3 Validation

After training our models for the given dataset partitioning, we inferred the labels for the provided unseen validation dataset. Then, generated segmentation labels were uploaded to the challenge system to obtain the evaluation results.

The FeTS evaluation system provides the Dice score and the Hausdorff95 metrics for each image. Briefly, the Dice score measures the overlap of two

segmentations (ground truth and the model's output). It is defined between 0 and 1 where 1 indicates a perfect match. Hausdorff95 measures the 95th percentile of distances between points on one edge set (our segmentation) to points on the other edge set (ground truth).

3 Results

3.1 Experiments with 17 Collaborators

First, we verified the performance of our three experiments using the data consisting of 17 institutions. The data distribution and travelling order for all experiments are illustrated in Fig. 2. It is important to note that the distribution for the ascending and descending orders is smooth while the random order randomly placed the collaborators with more samples towards the end of the sequence.

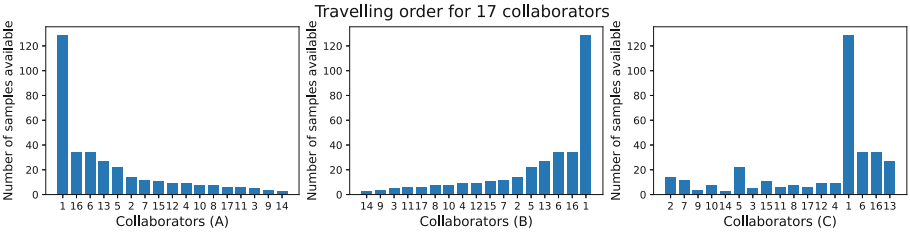

Fig. 2. Data distribution and travelling order for 17 collaborators: (A) descending order, (B) ascending order, and (C) random order.

The internal validation results for the above distributions are shown in Figs. 3 and 4. We evaluated travelling models for 1–5 cycles and compared the results to our baseline model (central learning) represented as a green line in the graphs.

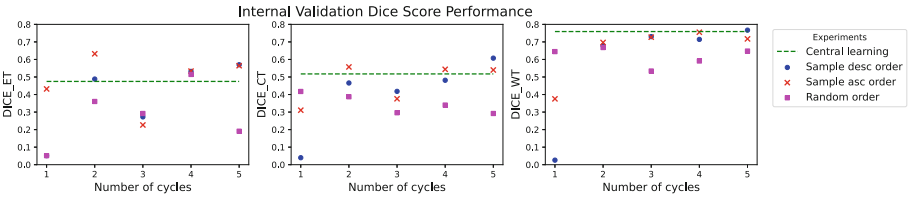

Fig. 3. Dice score performance evaluation results for travelling models with 1–5 cycles for 1 epoch per round travelling in different orders for partitioning one (17 collaborators).

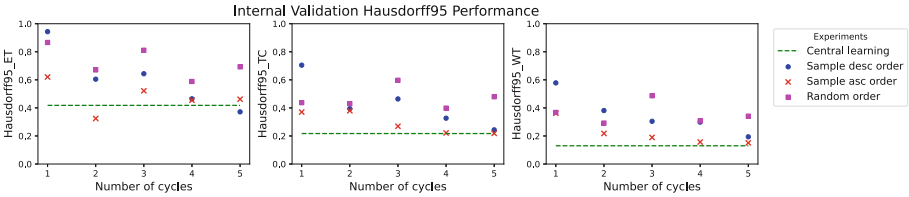

Fig. 4. Hausdorff95 performance evaluation results (values are divided by one hundred) for travelling models with 1–5 cycles for 1 epoch per round travelling in different orders for partitioning one (17 collaborators).

Analyzing the Dice scores for the whole tumor segmentation and one training cycle in Fig. 3, it becomes obvious that the random order led to the best results (0.647) followed by the ascending (0.375) and descending (0.025) order, but they performed worse compared to the baseline model (0.759). This might indicate that the network is mainly learning from the last collaborators of the sequence, forgetting the data seen at the first collaborators. However, investigating the models with more cycles, it becomes evident that this scenario changes. For two cycles, the order does not seem important in any experiment resulting in indistinguishable scores (0.667, 0.678, 0.696). For three, four, and five cycles, the ascending and descending orders led to similar results, which were also comparable to the baseline model. On the other hand, the random order performance improved as the number of cycles increased.

The Hausdorff95 metric demonstrated that the ascending order led to results that are more similar to the baseline model for all segmentations and cycles. It also indicates that as the number of cycles increases, the travelling order does not matter and all experiments tend to converge to similar results.

The results of the validation of the developed models using the unseen data is shown in Figs. 5 and 6. The metrics clearly show that the number of cycles is the major determinant for the generalizability of the network. It also reinforces the observation that with more cycles, the order of data contributors is less important, catastrophic forgetting does not happen, and all experiments (average Dice score for five cycles 0.672) achieve results comparable to the baseline score (0.667).

Thus, the results show that our distributed learning model does not forget patterns learned from previous collaborators. Travelling models with more than one cycle demonstrated to be more effective, and all orders converge to central learning results with an increasing number of cycles.

3.2 Experiments with 22 Collaborators

Finally, we evaluated the performance of the proposed distributed learning models based on the training data from 22 collaborators as illustrated in Fig. 7. It is important to highlight that this distribution is not as smooth as the previous one.

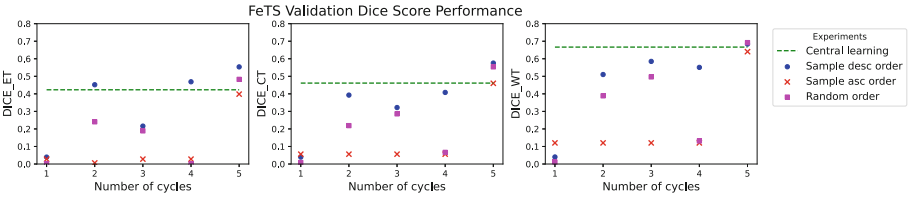

Fig. 5. Validation Dice scores for experiments with 1 epoch per round for partitioning one (17 collaborators).

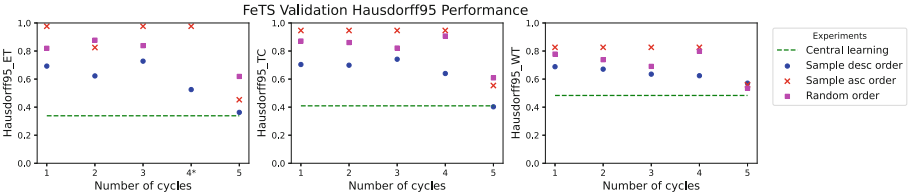

Fig. 6. Validation Hausdorff95 scores, values are divided by one hundred, for experiments with 1 epoch per round for partitioning one (17 collaborators). *Notice that random order with 4 cycles results in an outlier for the enhancing tumour segmentation measuring 3.7.

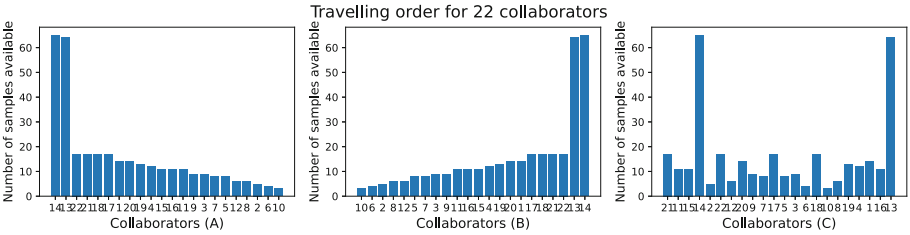

Fig. 7. Data distribution and travelling order for 22 collaborators: (A) descending order, (B) ascending order, and (C) random order.

As can be seen in Fig. 7(C), the database for 22 collaborators is more evenly distributed for the random order compared to the random distribution for the 17 collaborator database. Therefore, the collaborators that provide the majority of data are not concentrated towards the end of the sequence in this case.

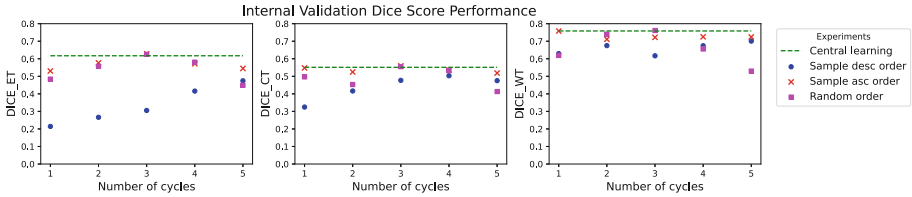

Fig. 8. Dice score performance evaluation results for travelling models with 1–5 cycles for 1 epoch per round travelling in different orders for partitioning two (22 collaborators).

In contrast to the previous results, the model trained with one cycle led to a performance comparable to the baseline model. While the ascending order achieved the same result as central learning (0.758), the descending and random orders achieved similar results (0.629, 0.619), which resulted in Dice scores only 0.14 worse than central learning. This difference might be due to the data distribution that does not have the majority of samples available at a single collaborator (Fig. 8).

Furthermore, travelling models with more than two cycles demonstrated the same trends as seen before. For two cycles, all order schemes resulted in Dice score differences of up to 0.1 (0.675, 0.711, 0.740). In addition to that, the ascending order achieved scores similar to the baseline model for all cycles numbers. Moreover, for five cycles, the descending order achieved a similar score compared to the ascending order, demonstrating once again that increasing the number of cycles improves performance and indicating that the order is not important as no catastrophic forgetting was observed.

Even though the Hausdorff metric performance was not as good for enhancing tumor and core tumor segmentations, with some outliers, it showed acceptable results for the whole tumor segmentation with all experiments leading to similar results compared to the baseline model (Fig. 9).

Validation metrics demonstrated similar trends compared to the training metrics. The ascending order performance was better for all cycles except for two cycles, resulting in scores similar to the descending and random orders (average Dice score 0.521). However, comparing these metrics to the validation metrics for 17 collaborators, the 22 collaborators results were worse and showed more variations when compared to the training results (Figs. 10 and 11).

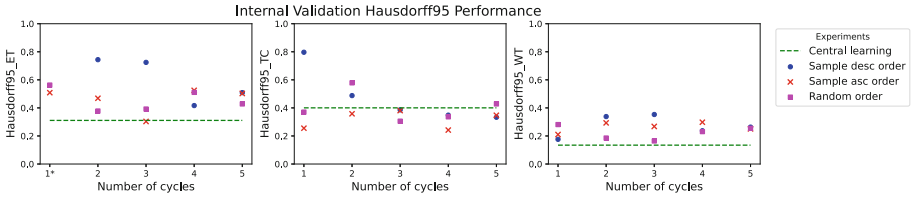

Fig. 9. Hausdorff95 performance evaluation results, values are divided by one hundred, for travelling models with 1–5 cycles for 1 epoch per round travelling in different orders for partitioning two (22 collaborators). *Notice that for tumor core segmentation, the central learning model results in an outlier measuring 3.7 and ascending order for 1 cycle is an outlier for the enhancing tumour segmentation measuring 1.2.

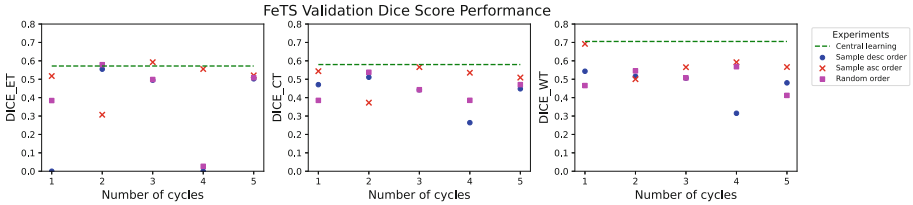

Fig. 10. Dice Score validation of experiments with 1 epoch per round for partitioning two (22 collaborators).

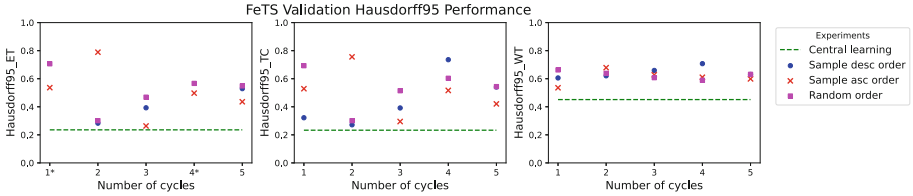

Fig. 11. Hausdorff95 validation results, values are divided by one hundred, with 1 epoch per round for partitioning two (22 collaborators). *Notice that for ascending order for 1 and 4 cycles are outliers for the enhancing tumour segmentation measuring 3.5.

3.3 Leaderboard Validation

For the final validation of our work, we needed to select one of our models to be evaluated by the challenge leaderboard on a hidden test set. In order to verify our assumption that as number of cycles increases the order loses importance, we submitted our random order travelling model to be evaluated for five cycles, using the same learning rate (0.05) used for our validations. A comparison of our evaluation and leaderboards' evaluation can be seen in Figs. 12 and 13.

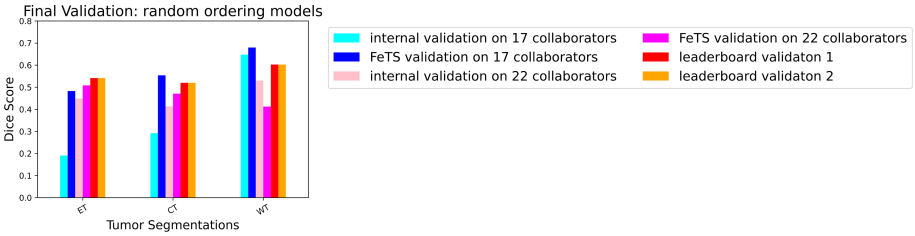

Fig. 12. Dice score comparison for random order models for 5 cycles. Light blue represents our internal validation for 17 collaborators, dark blue represents FeTS validation for 17 collaborators, light pink represents our internal validation for 22 collaborators, dark pink represents FeTS validation for 22 collaborators, and red and orange represent the results provided by FeTS challenge for leaderboard 1 and 2.

Figure 13 compares Dice scores for our internal and external validations for 17 and 22 institutions, and leaderboards validation as provided by the FeTS challenge on a hidden test set. The leaderboard evaluation achieved results (WT: 0.602) comparable to evaluations done for 17 institutions (WT internal: 0.647, WT external: 0.679), which suggests that our assumption that the number of cycles is more important for travelling models than travelling order is indeed correct.

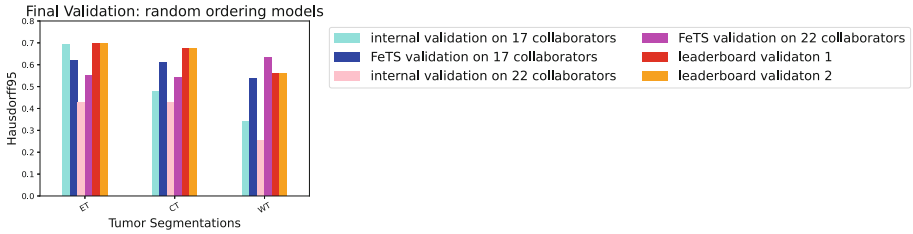

Fig. 13. Hausdorff95 comparison for random ordering models for 5 cycles. Light blue represents our internal validation for 17 collaborators, dark blue represents FeTS validation for 17 collaborators, light pink represents our internal validation for 22 collaborators, dark pink represents FeTS validation for 22 collaborators, and red and orange represent the results provided by FeTS challenge for leaderboard 1 and 2.

Furthermore, Fig. 13 shows Hausdorff95 results for the same models. The results are consistent with the results found for the Dice score results whereas the leaderboard evaluation and our evaluations for 17 institutions led to similar results. This demonstrates that our approach is robust and generalizable enough to be applied in different scenarios.

4 Discussion

In this work, we demonstrated that travelling models are able to leverage knowledge from multiple collaborators and could be used as a new approach avoiding pooling data from different collaborators together. We also compared three distinct travelling orders and demonstrated that the segmentation network did not suffer from catastrophic forgetting, as two cycles or more achieved similar results for the three different travelling orders and their performance converge with increasing number of cycles.

Moreover, we observed that when the data is more evenly distributed among the collaborators and there is not a single collaborator holding the majority of the data, the ascending travelling order achieved results comparable to the baseline model for the different number of cycles, while the descending and random travelling orders achieved improved results when the number of cycles is increasing. On the other hand, when the data distribution is smooth and there is a single institution providing most of the data, all experiments showed that performance improves as number of cycles increases. However, the one cycle models did not perform as well as the baseline model.

Overall, the validation performance for 17 collaborators indicated that fewer collaborators and a smoother data distribution is better when compared to the 22 collaborators results with a more uneven distribution.

Furthermore, we believe that as the number of cycles increases, the network performance will improve and, when reaching a certain number of cycles, all experiments should converge to results similar to the baseline model's accuracy level. However, the use of additional cycles should be investigated in future work to confirm this assumption.

5 Conclusion

We implemented distributed learning methods that aggregate weights effectively within a traveling model, leveraging knowledge from data of multiple collaborators with real world distribution, thereby avoiding the need to share data centrally.

The simplicity of the methodology and ability to learn remotely are important benefits of this method, which might convince health care centers to collaborate and take advantage of it in the future eliminating data sharing bureaucracy.

Acknowledgement. This work was supported by University of Calgary BME Research Scholarship (RS), the Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship (AT) and Discovery Grant (NDF), MITACS Globalink Research Internship (LT), Canada Research Chairs program (NDF), the River Fund at Calgary Foundation (NDF), and Canadian Institutes of Health Research (NDF).

Disclosures. The authors declare that there is no conflict of interest.

References

1. Lo Vercio, L., et al.: Supervised machine learning tools: a tutorial for clinicians. J. Neural Eng. **17**(6), 062001 (2020). https://doi.org/10.1088/1741-2552/abbff2
2. Hinton, G.: Deep learning-a technology with the potential to transform health care. JAMA **320**(11), 1101–1102 (2018). https://doi.org/10.1001/jama.2018.11100
3. Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F.: Secure, privacy-preserving and federated machine learning in medical imaging. Nat. Mach. Intell. **2**(6), 305–311 (2020). https://doi.org/10.1038/s42256-020-0186-1
4. MacEachern, S.J., Forkert, N.D.: Machine learning for precision medicine. Genome **64**(4), 416–425 (2021). https://doi.org/10.1139/gen-2020-0131
5. HIPAA. US Department of Health and Human Services (2020). https://www.hhs.gov/hipaa/index.html
6. GDPR. Intersoft Consulting (2016). https://gdpr-info.eu
7. Tuladhar, A., Gill, S., Ismail, Z., Forkert, N.D.: Building machine learning models without sharing patient data: a simulation-based analysis of distributed learning by ensembling. J. Biomed. Inform. **106**, 103424 (2020). https://doi.org/10.1016/j.jbi.2020.103424
8. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. **10**(2), 1–19 (2019). https://doi.org/10.1145/3298981
9. Chang, K., et al.: Distributed deep learning networks among institutions for medical imaging. J. Am. Med. Inform. Assoc. **25**(8), 945–954 (2018). https://doi.org/10.1093/jamia/ocy017
10. Remedios, S.W., et al.: Distributed deep learning across multisite datasets for generalized CT hemorrhage segmentation. Med. Phys. **47**(1), 89–98 (2020). https://doi.org/10.1002/mp.13880
11. Reina, G.A., Gruzdev, A., Foley, P., Perepelkina, O., Sharma, M., Davidyuk, I., et al.: OpenFL: an open-source framework for Federated Learning. arXiv preprint arXiv:2105.06413 (2021)
12. Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., et al.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Nat. Sci. Rep. **10**, 12598 (2020). https://doi.org/10.1038/s41598-020-69250-1
13. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Nat. Sci. Data **4**, 170117 (2017). https://doi.org/10.1038/SDATA.2017.117
14. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
15. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF

16. Pati, S., Baid, U., Zenk, M., Edwards, B., Sheller, M.J., Reina, G.A., et al.: The Federated Tumor Segmentation (FeTS) Challenge, arXiv preprint arXiv:2105.05874 (2021)
17. Kirkpatrick, J., et al.: Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. **114**, 3521–3526 (2017)
18. The Federated Tumor Segmentation (FeTS) Challenge https://www.fets.ai/. Accessed 22 July 2021