nature portfolio

Peer Review File

An International Study Presenting a Federated Learning AI Platform for Pediatric Brain Tumors

Editorial Note: Pages 2, 4, 8 and 14 in this Peer Review File have been amended to follow editorial policy for the ECR Co-Review Scheme.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author): Expert in federated learning, AI, and medical image analysis

- The authors do not review the extensive FL literature and ignore recent progress, even the ones recently published in NatComms. How was the federated mechanism selected? Why not some other FL framework?

- The comparison is basically about CDS and FL. Why? The classification accuracy of FL is around 88%. Would any other be able to beat that? Let say, a foundation model with or without fine-tuning. Would separate fine-tuning of a segmentation model with something like SAM bring the same results (and no communication overhead)? The classification can be also trained in conjunction with separate ML methods.

- I could find any explanation why the holdout accuracy for TK and AU is so much lower than CDS (Table 4)

- What was training/testing behaviour at individual sites? Figure 5 provides insight for classes not for sites.

Reviewer #1 (Remarks on code availability):

At a general glance, the code is of reasonable quality, is commented at a satisfactory level, and will most likely help to reproduce the results.

Reviewer #2 (Remarks to the Author): Clinical expert in brain cancer imaging and radiology, and MRI;

In this study, the authors introduce FL-PedBrain, an extensive federated learning (FL) platform that orchestrated collaborative efforts across global sites for the classification and segmentation of pediatric posterior fossa brain tumors. The study utilizes a large dataset of posterior fossa (PF) tumor patients, consisting of 596 medulloblastomas, 210 ependymomas, 335 pilocytic astrocytomas, and 327 diffuse intrinsic pontine gliomas. The study aims to employ federated learning to allow integration of multiple datasets from different sources (e.g., hospital systems) to create "big data" sets. The goal of these big datasets would be to facilitate AI training and generalizability, without the need to collect and analyze datasets at a central location. The main role of FL is therefore to overcome data sharing limitations due to privacy concerns, by avoiding the sharing of raw data. The categorization of pediatric brain tumors using MRI images is used as a case example. Pediatric brain tumors are relatively rare at individual sites, which motivates the combining of datasets from multiple institutions. This pipeline also has the advantages of circumventing patient privacy of data elements (HIPPA compliance) which can create logistical barriers to analyzing data across hospital systems. Employing a 3D-UNet architecture, the federated learning process utilized federated averaging. A warm-up training stage was proposed, initiating federated learning between the two most data-abundant sites for initial iterations before extending to other sites. The findings indicate that FL can achieve accuracy levels comparable to those achieved by centralized data sharing (CDS) training. Of note, the goal of FL is not to improve the predictive performance of image-based models, as the use of CDS predictive accuracy serves as an upper bound in performance as it has access to all the raw data.

Strengths:

1. Large relatively balanced dataset of diverse pathologies of the 4 most common pediatric PF tumors.

2. Relatively high predictive accuracies achieved for PF tumor diagnosis, for both the centralized data sharing (CDS) and the FL methods.

3. Demonstration of an application of FL to overcome data scarcity and privacy limitations

Critiques:

1. The clinical significance of the findings remains unclear. In many clinical settings, data sharing is allowed with proper anonymization and informed consent. In this context, it is unclear how the logistical costs of FL compare with those of CDS at each institution and across sites deploying FL (e.g., communication overhead, model synchronization, computational resource costs, etc). Also, what are the requirements for technical expertise for implementing FL at each site, and are there steps in place for quality assurance (QA) of imaging parameters/quality, absence of image artifacts, integrity of pathologic diagnosis, visual QA of segmentations, etc? An argument could be made that approaches streamlining the CDS method (e.g., pipelines to automate image anonymization and transfer) could avoid some of the logistical requirements of FL, while providing opportunities for central oversight and QA.

2. Even if data sharing is not an option, with the existing experiments and results, it is not clearly demonstrated that FL is needed. For instance, deep learning for brain tumor segmentation and classification in MRI is a well-studied domain. It may be possible for a site to use more sophisticated

algorithms with only on-site data and outperform the FL approach presented in the work when making predictions on local data. Given that each site is primarily concerned with the performance of the model on its specific data, rather than the model's ability to generalize across different sites, such a result could disincentivize the participation in FL. A formal comparison of on-site vs federated models for each site to demonstrate the need of FL would be helpful to demonstrate the importance of FL in this setting. The evaluation of the siloed model on data from other sites may have less relevance. Given the heterogeneity of data across sites, it would be expected that a model trained on one site would not be generalizable to other sites.

3. The discussion could be improved by expanding on the real-world utility/feasibility and implementation of the proposed method. While the study mentions some of the benefits of more sophisticated FL techniques, a clear rationale for their exclusion is not provided. The work could also benefit from expanded discussion about algorithm design in terms of computation and communication across sites. The training directly uses large volumes of 3D data, necessitating high-memory GPUs at each site, yet the potential limitations stemming from this design choice remain unexplored. High communication bandwidth and GPU resource requirements should be discussed as limitations to real-world application.

4. There should be expansion of discussion on model design and training procedures, to support an understanding of the reproducibility of the work. For instance, the description of preprocessing is unclear, and questions remain regarding the selection process for the 64 axial slices and the normalization procedure, particularly in the context of different sites. The authors should include information about the hyperparameter tuning strategies. There is ambiguity regarding the need/role of the validation set, as it appears to be treated as an additional test set. The work recognizes the heterogeneity and imbalance of data across different sites, but there is lack of mention of other strategies, such as data augmentation or weighting, that could boost the performance of the models.

5. The paper could benefit from a figure showing the organization of data, including individual sites, splitting of training/validation; hold out sites, etc.

Reviewer #2 (Remarks on code availability):

appears adequate

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports as part of the Nature Communications initiative to facilitate training in peer review and appropriate recognition for co-reviewers.

REVIEWER COMMENTS AND AUTHOR RESPONSES IN BULLET POINTS

Reviewer #1 (Remarks to the Author): Expert in federated learning, AI, and medical image analysis

- The authors do not review the extensive FL literature and ignore recent progress, even the ones recently published in NatComms. How was the federated mechanism selected? Why not some other FL framework?

- Thank you for the review. We apologize if any literature important to this topic was omitted. We had previously cited FL works including the NatComms paper [17]. On this revision, we now include the following literature [27-29] that discusses advanced strategies, as well as a recent work on gastric tumors with 4 FL participating sites published in NatComms [26].

References above:

- [17] Pati, Sarthak, et al. "Federated learning enables big data for rare cancer boundary detection." Nature communications 13.1 (2022): 7346.
- [26] Feng, B., Shi, J., Huang, L. et al. Robustly federated learning model for identifying high-risk patients with postoperative gastric cancer recurrence. Nature Communications. 15, 742 (2024).
- [27] Ye, Mang, et al. "Heterogeneous federated learning: State-of-the-art and research challenges." ACM Computing Surveys 56.3 (2023): 1-44.
- [28] Shao, J., Wu, F. & Zhang, J. Selective knowledge sharing for privacy-preserving federated distillation without a good teacher. Nat Commun 15, 349 (2024).
- [29] Rahimi, M.M., Bhatti, H.I., Park, Y., Kousar, H., Kim, D.Y. and Moon, J., 2024. EvoFed: Leveraging Evolutionary Strategies for Communication-Efficient Federated Learning. Advances in Neural Information Processing Systems, 36.

- We chose the federated mechanism based on simplicity, popularity, and ease of use. These came down to Federated Averaging (FedAvg) and Federated Proximal learning (FedProx). In our custom FL framework, we also included custom training strategies such as Federated Warm-up. More precisely, in Federated Warm-up, FL is first performed on the first two of the largest sites (ST and SE) prior to turning on FL on all 16 participating sites.
- We have tried other advanced FL strategies (synthetic data augmentation, and weight transfer-based strategies) that are aimed at heterogeneous, non-IID data but found that FedAvg with a FedProx hyperparameter of $\mu=0$ works the best for segmentation across all sites.
- We now add a new figure (**Figure 1b**) that further clarifies on our FL training procedure and the specific optimization function(s) used and the hyperparameters involved, as well as additional discussion in lines **523-530.**
- For reproducibility, our code and models can be found on GitHub.

- The comparison is basically about CDS and FL. Why? The classification accuracy of FL is around 88%. Would any other be able to beat that? Let say, a foundation model with or without fine-tuning. Would separate fine-tuning of a segmentation model with something like SAM bring the same results (and no communication overhead)? The classification can be also trained in conjunction with separate ML methods.

- Yes, this is a great point. The comparison is mainly between FL and CDS because CDS is how we typically train AI models and therefore have served as the benchmark. In practice, by pooling data together, CDS achieves the highest *average* accuracy. This is because with larger datasets, larger and more complex models can be developed that can reason well across the data distribution shared across the sites. In our work, we show that FL can enable us to still build the same large models with near-CDS performance in very heterogeneous cohorts and without pooling data.
- Other advanced AI methods such as using a foundation model can work and is a topic for future research. However, this still requires a model to first be trained on a large corpus of 3D imaging data.
- On the related topic of pretraining, we have found an uplift by leveraging a pretrained "Foundation" model that was pretrained on normal pediatric MRI brain and age data (n=1667) with ages ranging from 0.1 month to 20 years.
- The state-of-the-art segmentation models like SAM have been trained on >11M images including descriptions. They are considered good few-shot learners that can be finetuned on a handful of segmentation data. However, SAM does not support 3D or even video inputs ("Currently the model only supports images or individual frames from videos."). However, there are on-going efforts to create foundation models in radiology that can be proficient at zero-shot segmentation tasks. This requires data and (centralized) data sharing or Federated approaches. Here, our FL work aims to better understand the extent to which we can create large models *without* data share.

- I could [not] find any explanation why the holdout accuracy for TK and AU is so much lower than CDS (Table 4)

- We found that FL is slightly worse than CDS in classification accuracy. Class imbalance within the TK and AU sites likely affected the classification performance. Nonetheless, the segmentation performance is almost equivalent between CDS and FL for these two sites. To clarify, we have now added the following statement in the limitation section (lines **562-566)**.
  Class imbalance within TK and AU sites likely contribute to a slightly lower classification performance. For segmentation task, where the scores are calculated per pixel across the entire 256 x 256 x 64 head volume, we observe similar performance between FL and CDS.

- What was training/testing behaviour at individual sites? Figure 5 provides insight for classes not for sites.

- Re: training/testing behavior, we had saved all the per-site checkpoints for FL *and* performed cross-site validation across the 18 sites for every FL round [from Round 1 to Round 300 (5400 experiments)]. However, this data behavior was too complex to display in a summary format (for the purpose of the manuscript), and thus we show behavior for classes (previously **Figure 5**, now **Figure 5a, b**) *and* final site-specific performance (**Figure 2c**). If the reviewer wishes, we are glad to add the training convergence per FL round per site (16 sites) on the GitHub page.
- Upon request by Reviewer 2, we have performed a new ablation study (**Figure 5c**) showing the effect of adding sites into the FL network on the FL performance. We have added the following in the revised manuscript (Methods section, study design in lines **458-466**):

  Better performance with more active FL sites
  In our study, we assess the impact of site activity on FL performance by conducting an ablation experiment. This experiment measures the FL system's performance relative to the quantity of active training sites, as depicted in Figure 5(c). We rerun the full FL experiment by integrating more sites into the training process (x-axis), prioritizing those with larger datasets. The performance evaluation is based on the F1 score—specifically, the classification accuracy of the label that performs the poorest. Our findings indicate a positive correlation between the number of active sites and the F1 score: as more sites participate in the FL network, the F1 score improves, eventually equaling the peak score achieved when all available sites are active.

Reviewer #1 (Remarks on code availability):

At a general glance, the code is of reasonable quality, is commented at a satisfactory level, and will most likely help to reproduce the results.

- Thank you! We appreciate the feedback.

Reviewer #2 (Remarks to the Author): Clinical expert in brain cancer imaging and radiology, and MRI;

In this study, the authors introduce FL-PedBrain, an extensive federated learning (FL) platform that orchestrated collaborative efforts across global sites for the classification and segmentation of pediatric posterior fossa brain tumors. The study utilizes a large dataset of posterior fossa (PF) tumor patients, consisting of 596 medulloblastomas, 210 ependymomas, 335 pilocytic astrocytomas, and 327 diffuse intrinsic pontine gliomas. The study aims to employ federated learning to allow integration of multiple datasets from different sources (e.g., hospital systems) to create “big data” sets. The goal of these big datasets would be to facilitate AI training and generalizability, without the need to collect and analyze datasets at a central location. The main role of FL is therefore to overcome data sharing limitations due to privacy concerns, by avoiding the sharing of raw data. The categorization of pediatric brain tumors using MRI images is used

as a case example. Pediatric brain tumors are relatively rare at individual sites, which motivates the combining of datasets from multiple institutions. This pipeline also has the advantages of circumventing patient privacy of data elements (HIPPA compliance) which can create logistical barriers to analyzing data across hospital systems. Employing a 3D-UNet architecture, the federated learning process utilized federated averaging. A warm-up training stage was proposed, initiating federated learning between the two most data-abundant sites for initial iterations before extending to other sites. The findings indicate that FL can achieve accuracy levels comparable to those achieved by centralized data sharing (CDS) training. Of note, the goal of FL is not to improve the predictive performance of image-based models, as the use of CDS predictive accuracy serves as an upper bound in performance as it has access to all the raw data.

Strengths:
1. Large relatively balanced dataset of diverse pathologies of the 4 most common pediatric PF tumors.
2. Relatively high predictive accuracies achieved for PF tumor diagnosis, for both the centralized data sharing (CDS) and the FL methods.
3. Demonstration of an application of FL to overcome data scarcity and privacy limitations

Critiques:
1. The clinical significance of the findings remains unclear. In many clinical settings, data sharing is allowed with proper anonymization and informed consent. In this context, it is unclear how the logistical costs of FL compare with those of CDS at each institution and across sites deploying FL (e.g., communication overhead, model synchronization, computational resource costs, etc.). Also, what are the requirements for technical expertise for implementing FL at each site, and are there steps in place for quality assurance (QA) of imaging parameters/quality, absence of image artifacts, integrity of pathologic diagnosis, visual QA of segmentations, etc? An argument could be made that approaches streamlining the CDS method (e.g., pipelines to automate image anonymization and transfer) could avoid some of the logistical requirements of FL, while providing opportunities for central oversight and QA.

- We appreciate the feedback. CDS and FL each offer unique advantages. As the reviewer correctly notes, CDS is certainly feasible and remains attractive where data sharing across the hospitals is executable, legally acceptable, and attainable in a timely fashion. However, where data share is prohibitively time-intensive or not possible due to privacy, security, and legal challenges, we believe FL can be useful. For example, some of our data use agreements (DUAs) took 2-3 years to execute. As our study shows, FL can achieve CDS-like performance for both classification and segmentation of pediatric brain tumors even in highly heterogeneous settings (a large range of imaging sequence parameters, missing classes, dataset imbalances). Also, setting up the FL network among partnering sites allows us to rapidly develop models from *newer, updated* datasets that could be a hassle to continuously update in real-time with CDS.

- We have now added a new section dedicated to address Q1 (as well as Q3, and Q4), as shown below, on this revised manuscript (Methods section lines **458-492**).

In our study, we assess the impact of site activity on FL performance by conducting an ablation experiment. This experiment measures the FL system's performance relative to the quantity of active training sites, as depicted in Figure 5(c). We rerun the full FL experiment by integrating more sites into the training process (x-axis), prioritizing those with larger datasets. The performance evaluation is based on the F1 score—specifically, the classification accuracy of the label that performs the poorest. Our findings indicate a positive correlation between the number of active sites and the F1 score: as more sites participate in the FL network, the F1 score improves, eventually equaling the peak score achieved when all available sites are active.

Future Challenges and Practical Implementation

FL-PedBrain introduces logistical challenges, communication overhead, model synchronization, and computational demands.

Communication and Logistical Challenges. In FL, every participating hospital must regularly exchange model updates — specifically, the model weights after each FL training round. For our classification-segmentation model, this equates to transmitting approximately 125 MB of model weights per round. With 19 hospitals involved, this culminates in a data transfer of around 38 GB per hospital for each training session with 300 rounds. Training the largest dataset for 1 epoch consumes approximately 3-4 minutes on a V100 GPU, and the time to then transfer all 16 models from each hospital to the central parameter server (coordinating hospital) in Figure 1 (a) is roughly 2 minutes at a 1 MB/s internet. This equates to about 6 minutes per round (1 epoch per round) and 1200 minutes to ship one trained model. And although CDS only requires a one-time collection of 200-1000 GB of DICOMs, FL offers benefits by removing the need for data use agreements and the need for deidentification, which can take a long time to establish and verify. Finally, FL provides advantages such as continuous quality control and oversight from each of the sites' technical model builders.

Need for on-site Technical Expertise. Additionally, having both clinical and AI experts per site would greatly enhance and streamline the FL workflow, enabling them to 1) inspect the training and evaluation data for any obvious imaging artifacts or integrity of diagnosis and 2) monitor the training process as the model evolves. We intend our FL framework not to be used just for static datasets like in the CDS case but rather a bedrock for active learning on growing datasets. Therefore, human integration into the FL pipeline is a very promising future direction.

2. Even if data sharing is not an option, with the existing experiments and results, it is not clearly demonstrated that FL is needed. For instance, deep learning for brain tumor segmentation and classification in MRI is a well-studied domain. It may be possible for a site to use more

sophisticated algorithms with only on-site data and outperform the FL approach presented in the work when making predictions on local data. Given that each site is primarily concerned with the performance of the model on its specific data, rather than the model's ability to generalize across different sites, such a result could disincentivize the participation in FL. A formal comparison of on-site vs federated models for each site to demonstrate the need of FL would be helpful to demonstrate the importance of FL in this setting. The evaluation of the siloed model on data from other sites may have less relevance. Given the heterogeneity of data across sites, it would be expected that a model trained on one site would not be generalizable to other sites.

- Deep learning for adult brain tumors (mainly gliomas) is indeed a well-studied domain (e.g., Sheller et al., Pati et al. [16,17]). However, pediatric brain tumor studies remain sparse, if any, and of our magnitude and scale – that includes multiple international sites. Also, pediatric tumors are distinct as they commonly arise in the posterior fossa (brainstem and cerebellum), unlike the typical adult gliomas (cerebral), and harbor biologically divergent tumor pathologies (e.g., embryonal origins pilocytic astrocytomas, and other glial cell origins) with their own unique albeit, heterogeneous MRI features. Even pediatric supratentorial gliomas are considered biologically distinct from adult counterparts.

- It certainly is a great idea to consider pre-existing models. We had initially tried finetuning a state-of-the-art adult segmentation model trained on adult BRATS to our pediatric dataset. However, we found that the results were not as promising as when using a very clean FL strategy from scratch with pediatric normal controls.

- We also felt that an FL study and model development dedicated to pediatric tumors is of interest to both the clinical and research community, in light of their divergent tumor compositions, surgical implications (curative versus risk-mitigating approaches), and many new and ongoing immunotherapy trials that could benefit from more precise pediatric tumor evaluation strategies (e.g., CAR-T cell therapies for DIPG; [Majzner et al. Nature 2022; Vitanza et al. Cancer Disc. 2023]). Unlike some of the prior adult tumor 3D segmentation models (e.g., BRATS dataset) or prior adult glioma FL works (Sheller et al., Pati et al. [16,17]), our 3D Ped-FL model does not require skull stripping or brain registration into a common atlas, which we think would facilitate clinician user experience and translation.

In response to: A formal comparison of on-site vs federated models for each site to demonstrate the need of FL would be helpful to demonstrate the importance of FL in this setting.

- Thank you for bringing this up; the reviewer is right. We built a siloed model built for each site and tested against that site's validation set. It performs ~5 to 30% worse in classification and 40-50% worse in Dice score on the segmentation task. As shown in **Figure 2c**, even the siloed model trained solely on the site with the largest cohort

performs significantly worse in segmentation Dice Score (~40% lower) on the validation set of the same site when compared to FL or CDS.

- Despite being the most common solid cancer of childhood, pediatric brain tumors, are relatively sparse in each hospital. The reviewer is correct to note that some sites with a large data of its own may be de-incentivized to participate in FL, as long as they can build their own site-specific models that outperforms FL. But even a model built proprietarily with the largest site (ST), the siloed model fails to work well even on its own cohort (as shown by **Figure 2c**). In our study, we show that federated models offer CDS-like levels of performance without the need for data share.

- In response to: Given the heterogeneity of data across sites, it would be expected that a model trained on one site would not be generalizable to other sites. We think that this depends on the source of heterogeneity. The models built with FL and CDS learn to capture >95% of the common knowledge across all sites about posterior fossa tumors, structural brain development across age (using the 1667 pediatric normal brains). The last ~5% we believe are due site-specific difference or out-of-distribution sources such as scanner hardware and imaging software parameters and protocols, and other site-specific factors [e.g., tertiary care pediatric centers may capture tumors at an earlier phase among those who are screened for genetic risks (e.g., P53 mutations, NF1), whereas in some community hospitals, patients may present at a later symptomatic phase when the tumors are larger or more aggressive]. We believe that an FL or CDS-trained model can be further fine-tuned (if necessary) on the site before deployment to account for that last 5% of variation.

3. The discussion could be improved by expanding on the real-world utility/feasibility and implementation of the proposed method. While the study mentions some of the benefits of more sophisticated FL techniques, a clear rationale for their exclusion is not provided. The work could also benefit from expanded discussion about algorithm design in terms of computation and communication across sites. The training directly uses large volumes of 3D data, necessitating high-memory GPUs at each site, yet the potential limitations stemming from this design choice remain unexplored. High communication bandwidth and GPU resource requirements should be discussed as limitations to real-world application.

- Thank you for the feedback. We have included real-world and practical challenges in the section including communication overheads (see above under Q1). We have also included time of FL training (minutes) with communication overhead. Furthermore, we have also included a web interactive demo of an FL learned model that anyone can use. See the GitHub webpage where we show a demo movie as an example of an FL learned model that a non-technical clinician can use. We share a link interactive website in the GitHub.

4. There should be expansion of discussion on model design and training procedures, to support an understanding of the reproducibility of the work. For instance, the description of

preprocessing is unclear, and questions remain regarding the selection process for the 64 axial slices and the normalization procedure, particularly in the context of different sites. The authors should include information about the hyperparameter tuning strategies. There is ambiguity regarding the need/role of the validation set, as it appears to be treated as an additional test set. The work recognizes the heterogeneity and imbalance of data across different sites, but there is lack of mention of other strategies, such as data augmentation or weighting, that could boost the performance of the models.

- All hyperparameters are kept extremely simple. We now add these numbers in the GitHub, as well as a new figure (**Figure 1b**), which contains the math and the exact optimization loss that was used. We have now added in lines **337-346 (next paragraph)**, specific details about the preprocessing procedure, as well as the preprocessing script, which is kept extremely simple and naive to imaging sequence parameters and slice thicknesses. This model can accommodate the extremely large variation of different T2 sequence parameters and slice thicknesses. We have also tried other methods such as data augmentation using synthetic data and other advanced FL methods. However, we have found that these results were mixed and would only complicate and dilute the main novelty of our work. Future research is always welcome with the dataset provided and potential challenges that we are excited to organize in the future.

  Data Preprocessing
  Each site must possess the small but important knowledge to manage consistent data preprocessing, a task that, under CDS, would typically be centralized by a trusted party. To streamline preprocessing, we have minimized any complex preprocessing steps (e.g., brain registration to a common atlas or skull-stripping.). Preprocessing only includes: 1) normalization of each 3D image to a simple 0-255 intensity range and 2) volume extraction of 64 congruent axial slices of 256 x 256. These preprocessing steps are executed via an automated script applied to the DICOM data across all 19 sites. The number of 64 slices was chosen such that it can handle virtually all of the variations of the individual sites' T2 sequence parameters (e.g., TSE, FSE, Propeller, etc.) with a large range of slice thicknesses 1-5 mm based on site scanner technology and protocol. Therefore, our FL system can accommodate a large range of sequence parameters and axial slices...

- Finally, we have also added another section in lines **458-466**, ***Better performance with more active FL sites,*** which describes an ablation study showing the impact of active FL sites on model performance.

5. The paper could benefit from a figure showing the organization of data, including individual sites, splitting of training/validation; hold out sites, etc.

- Thank you! Done, we now include a new figure (**Figure S1**) in the Supplement.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports as part of the Nature Communications initiative to facilitate training in peer review and appropriate recognition for co-reviewers.

- Thank you again for your time and the feedback!

Final author comments:

Thank you very much for everyone's review. We have started the data upload process for the FL data training and testing under https://doi.org/10.25740/bf070wx6289. And will continually upload till all of the participating sites are represented. For this and other details posted in this review, we'd like to guide you to the github page for future updates.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

All my comments have been addressed. Though authors have added some references to be comprehensive, still some recent ones like "proxyFL" are missing (Kalra, et al. "Decentralized federated learning through proxy model sharing." Nature communications 14.1 (2023): 2899.).

Reviewer #2 (Remarks to the Author):

The authors have addressed nearly all of our previous questions/suggestions. This has strengthened the manuscript. There are a couple aspects we ask the authors to address:

1. The authors have justified the need for federated learning by training models on each site individually and showing their poorer performance compared to federated learning. Specifically, the authors “built a siloed model built for each site and tested against that site’s validation set. It performs ~5 to 30% worse in 260 classification and 40-50% worse in Dice score on the segmentation task. As shown in Figure 2c, even the siloed model trained solely on the site with the largest cohort 7 performs significantly worse in segmentation Dice Score (~40% lower) on the validation set of the same site when compared to FL or CDS. “

We believe the manuscript would be strengthened by including these results in the published work, in addition to inclusion of figures (such as in Figure 2c). These can be included in the supplementary data/figures. This would help to reflect the data heterogeneity across sites to the readership, instead of only showing a single siloed model trained on the largest site,

2. The authors added a section titled "Future Challenges and Practical Implementation." In this section, an idealized scenario is presented, concluding that 1200 minutes are needed to train the federated learning model. While the use of real numbers is appreciated, this section could be improved as there are a few mistakes and unrealistic assumptions:

o This section mentions training for 300 rounds, whereas the rest of the text uses 200 rounds.

o The parameter data transfer calculation only considered the time for one-way transfer. In reality, each round requires model parameters to be sent from the hospitals to the central server and back.

o The time required for data transfer only considered non-central server hospitals. The central server would need to transfer 15x the data. Considering this and the previous point, solely the data transfer in each round would take over 1 hour with 1MB/s internet, rather than the stated 2 minutes.

o Synchronization challenges were not expanded upon. In practice, not all hospitals would have the same computation speeds or internet bandwidth. Federated learning would be bottlenecked by the slowest site each round.

o Computational cost challenges were not expanded upon. The work assumed each site would train on a V100 GPU, which may not be commonly available at hospitals. Using 3rd party cloud services with capable GPUs is possible, but could introduce privacy risks. These points may be better suited for the limitations section rather than results.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

REVIEWER COMMENTS AND AUTHOR RESPONSES IN BULLET POINTS

Reviewer #1 (Remarks to the Author): Expert in federated learning, AI, and medical image analysis

REVIEWER COMMENTS Reviewer #1 (Remarks to the Author): All my comments have been addressed. Though authors have added some references to be comprehensive, still some recent ones like "proxyFL" are missing (Kalra, et al. "Decentralized federated learning through proxy model sharing." Nature communications 14.1 (2023): 2899.).

Thank you for your review. We have added this reference in line 497:
Recent methods for heterogeneous FL [30] can potentially alleviate communication and compute overheads.

Reviewer #2 (Remarks to the Author): The authors have addressed nearly all of our previous questions/suggestions. This has strengthened the manuscript. There are a couple aspects we ask the authors to address: 1. The authors have justified the need for federated learning by training models on each site individually and showing their poorer performance compared to federated learning. Specifically, the authors "built a siloed model built for each site and tested against that site's validation set. It performs ~5 to 30% worse in 260 classification and 40-50% worse in Dice score on the segmentation task. As shown in Figure 2c, even the siloed model trained solely on the site with the largest cohort 7 performs significantly worse in segmentation Dice Score (~40% lower) on the validation set of the same site when compared to FL or CDS. "

We believe the manuscript would be strengthened by including these results in the published work, in addition to inclusion of figures (such as in Figure 2c). These can be included in the supplementary data/figures. This would help to reflect the data heterogeneity across sites to the readership, instead of only showing a single siloed model trained on the largest site,

Thank you, we strongly agree, and have added another table in the Supplemental Figure section (Supplemental Table 1). We also added this section in lines 432-436:
Supplementary Table 1 presents the classification and segmentation results for the 16 independently trained models, each using its respective site-specific dataset. The outcomes suggest subpar performance across the board, attributable to the limited size of individual datasets. Notably, models from sites UT and CP showed the highest segmentation Dice Scores, reaching 0.57. However, models from five sites did not converge.

2. The authors added a section titled "Future Challenges and Practical Implementation." In this section, an idealized scenario is presented, concluding that 1200 minutes are needed to train

the federated learning model. While the use of real numbers is appreciated, this section could be improved as there are a few mistakes and unrealistic assumptions:
o This section mentions training for 300 rounds, whereas the rest of the text uses 200 rounds.
o The parameter data transfer calculation only considered the time for one-way transfer. In reality, each round requires model parameters to be sent from the hospitals to the central server and back.

Thank you for the catch. The 2x factor was indeed a typo but we tried to keep everything somewhat on the approximate order. The 300 rounds should be 200 rounds. We have fixed this.

o The time required for data transfer only considered non-central server hospitals. The central server would need to transfer 15x the data. Considering this and the previous point, solely the data transfer in each round would take over 1 hour with 1MB/s internet, rather than the stated 2 minutes.

Yes, the central hospital (parameter server) will be burdened by a 15-way transfer. We had initially assumed that this server's network would accommodate for this and the transfers would be bottlenecked by the other 15 sites' bandwidth. But with calculation, you are correct. The server in practice would not necessarily be a physical point like a hospital but some cloud service (e.g., AWS, Azure) that has end-points in different regions (e.g., Ireland hospital would send parameters to the UK endpoint) and each of these cloud endpoints would transmit to a central endpoint. However, even with cloud end-points, the FL times are still bottlenecked by the slowest internet connection to and from one of the hospitals (1 MB/s).

o Synchronization challenges were not expanded upon. In practice, not all hospitals would have the same computation speeds or internet bandwidth. Federated learning would be bottlenecked by the slowest site each round.

o Computational cost challenges were not expanded upon. The work assumed each site would train on a V100 GPU, which may not be commonly available at hospitals. Using 3rd party cloud services with capable GPUs is possible but could introduce privacy risks. These points may be better suited for the limitations section rather than results.

We appreciate the feedback. With the above comments, we have revised the section in lines 480-496 as shown in purple.

Communication and Logistical Challenges. In FL, every participating hospital must regularly exchange model updates — specifically, the model weights after each FL training round. For our classification-segmentation model, this equates to transmitting approximately 125 MB of model weights per round. This culminates in a data transfer of around 74 GB per hospital for each training session with 200 rounds. Training the largest dataset for 1 epoch consumes approximately 3-4 minutes on a V100 GPU, and the time to then transfer all 16 models from each hospital to the central parameter server (coordinating hospital) in Figure 1 (a) is roughly 7 minutes at 1 MB/s internet upload rate, assuming that the central server's download rate is

much faster than 1 MB/s. This equates to about 10 minutes per round (1 epoch per round) and 2000 minutes to ship one trained model. And although CDS only requires a one-time collection of 200-1000 GB of DICOMs, FL offers benefits by removing the need for data use agreements and the need for deidentification, which can take a long time to establish and verify. Finally, FL provides advantages such as continuous quality control and oversight from each of the sites' technical model builders. The provided figures are rough estimates; actual performance will vary as hospitals differ in computing power, communication standards, and data transfer speeds. Asynchronous Federated Learning (FL) is particularly beneficial in environments where hospitals exhibit diversity not just in data but also in computational and networking resources.

Reviewer #3 (Remarks to the Author): I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

REVIEWERS' COMMENTS

Reviewer #2 (Remarks to the Author):

In the revised manuscript, the authors have added a supplemental table showing that individual siloed training can not rival FL in performance. The authors have also corrected the typographical errors and provided justifications for the new time estimates for model training. As such, the authors have adequately addressed the previous comments.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Thank you everyone for the reviews!

The revised manuscript contains changes made based on the author checklist. Any changes are made in green this time (as opposed to blue and red previous revisions).

REVIEWERS' COMMENTS Reviewer #2 (Remarks to the Author): In the revised manuscript, the authors have added a supplemental table showing that individual siloed training can not rival FL in performance. The authors have also corrected the typographical errors and provided justifications for the new time estimates for model training. As such, the authors have adequately addressed the previous comments.

Thank you for your review.

Reviewer #3 (Remarks to the Author): I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Thank you for your review.