Federated Learning in Prediction of Dementia Stage: An Experimental Study

Boyun Eom, Muhammad Zubair, Dong-Hwan Park, Hyunhak Kim,Young-Ho Suh, Sunhwan Lim, Chanwon Park
Autonomous IoT Research Section,
Electronics and Telecommunications Research Institute
Daejeon, South Korea
Email : {eby, zubair5608, dhpark, hh.kim, yhsuh, shlim, cwp}@etri.re.kr

Abstract*—Federated Learning(FL) has emerged as the optimal approach for training machine learning models when dealing with data containing sensitive information, making data sharing impractical. Particularly in contexts where privacy is a primary concern, such as medical applications, Federated Learning demonstrates its efficacy as a solution. Motivated by this, we have conducted comprehensive experiments using a medical image dataset. One of the key objectives of these experiments is to evaluate the influence of Non-IID data which is frequently encountered in Federated Learning, especially within the medial field. We present our exploration of Federated Learning in classification of OASIS medical images, along with the results obtained from various experiments.

Keywords— machine learning, DNN, federated learning, CNN, Non-IIDness, flower

I. INTRODUCTION

While Federated Learning (FL) is a relatively recent development in the field of deep learning, it has garnered increasing interest as a potential solution for training machine learning (ML) models without disclosing the underlying data. As the domains for applying FL continue to widen, there have been increasing efforts to implement this technology in medical fields, particularly where privacy preservation is required by regulations [1,2]. FL often needs to deal with not independent and identically distributed(Non-IID) data, which arises from statistically imbalanced local datasets and the heterogeneity of participating devices. It is widely acknowledged that FL may not deliver good performance on all clients due to this Non-IIDness [2,3]. In this paper, we present our findings from study of FL framework in predicting Alzheimer Disease(AD) stage. First, we compare FL to traditional ML approach which trains models on single server using aggregated data, to predict Alzheimer's disease stage on MRI(Magnetic Resonance Imaging) images. Then, we delve into how FL works with Non-IIDness.

The remaining sections of this paper are structured as follows; In section 2, we provide brief overviews of related work. Section 3 explains our methodology, and Section 4 contains the experimental results and discussions. Finally, we conclude in section 5.

II. RELATED WORK

A. Federated Learning and Non-IIDness

In general, DNN(Deep Neural Network) models comprise millions of parameters, referred as weights, and these transform input values into each layers, as illustrated in Fig.1. During training processes, these DNN parameters are tuned to the purpose of performing various tasks such as detection, classification, segmentation, and more.

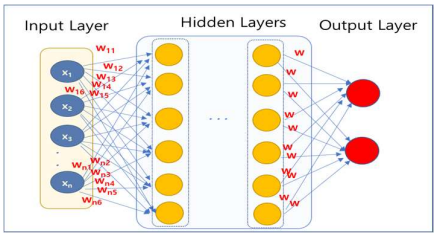

Fig. 1. DNN Network

In the context of Federated Learning (FL), the training procedure entails the model being exposed to datasets located on individual clients' systems. Rather than sharing the data itself, the process involves sending the model weights to a central server over several rounds of federation. Subsequently, the server aggregates the weights obtained from all clients, performs updates, and returns refined model parameters to the respective clients. The clients then employ these updated parameters to retrain the model using their local datasets. This cycle of processes is repeated iteratively until the completion of the FL. The primitive objective of FL can be written as (1);

$$\min_{(w_1,w_2,\ldots,w_n)} \frac{1}{n}\sum_{i=1}^{n}\frac{|D_i|}{K} l_i\,(W_i; D_i) \qquad (1)$$

here, n is the number of FL clients and K is the total sample count over all clients. l_i is the local loss function, W_i denotes the model parameter, and D_i is the local dataset on client i. There have been many studies on aggregation algorithms on a server which make great impact on the overall performance of FL[4,5]. FedAVG, FedProx and Scaffold are the most well-known aggregators. The objective of FedAvg and its variants, is to get an optimal global model parameter, w, across all clients so that the parameters W_i for all clients are same as w [4]. It is known that those parameters will be tuned more accurately and result in better performance with more federation rounds and more extensive training in FL [4]. Exhausted experiments on Non-IID using above mentioned algorithms has conducted in [6].

While FL has its outstanding strengths, such as preserving privacy and making efficient use of hardware, the main drawbacks of current FL frameworks lie in excessive computation and communication overhead [3]. Non-IIDness caused from the heterogeneity of data and device has been considered as a challenge which degrades the performance of FL [3,5]. Previous work well summarizes the Non-IID cases in terms of the distribution perspective [5].

B. Deep Learning for AD Stage Recognization

Alzheimer's Disease (AD) is the most common form of dementia in the elderly, leading to a high degree of neuronal

death in various areas of the brain. Early diagnosis of AD is known to be significantly important for delaying the progression of the disease [7]. Clinical Dementia Rating(CDR) scale has been used as an indicator of the status of AD [8]. A CDR score of 0 corresponds to no dementia, while scores of 0.5, 1, 2, and 3 represent 'very mild', 'mild', 'moderate', and 'severe dementia', respectively. The medial temporal lobe and the hippocampus are typically the brain's initial regions affected by AD, and the degeneration can be detected through MRI. The continuous development of neuroimaging techniques, combined with the rapid progress of AI technology, has paved the way for extensive research in automated classification of AD, enabling early detection using deep learning algorithms that leverage multimodal neuroimaging data [9]. In recent years, there has been a notable surge in applying deep learning models to medical images, including CT and MRI scans and it has been indicated that using two or more multimodal neuroimaging data types has produced higher accuracies than a single neuroimaging type [10]. However, due to its high cost of PET, many studies have concentrated on MRI data as resource for training models. Convolutional Neural Network (CNN) stands out as the most extensively used deep learning technique in computer vision, demonstrating superior performance compared to other methods. The comparison of accuracy with several pre-defined models as well as CNN in predicting AD stages has been provided in [7, 11].

III. Methodology

We have validated the performance of FL for classification of AD (Alzheimer's Disease) images. In this section, we explain our method for predicting AD stage using FL.

A. FL Framework

In our FL setting, we utilized Flower, an open-source FL framework[12]. One server and three participant clients implemented using Keras were executed and FedAVG was used for the aggregation of weights on FL server.

B. Data Acquisition

Our experiments for predicting the AD stage were conducted on a dataset derived from OASIS (Open Access Series of Imaging Studies), which comprises approximately 80,000 brain MRI images [13]. The main reason why we chose OASIS brain image set instead of easily used benchmark such as MNIST or CIFAR, is we believed that the domain of medical is the most appropriate to adopt FL technology. These images are labeled with four CDR scales, 0, 0.5, 1 and 2. Fig. 2 shows those four samples of each AD stages of MRI images we obtained.

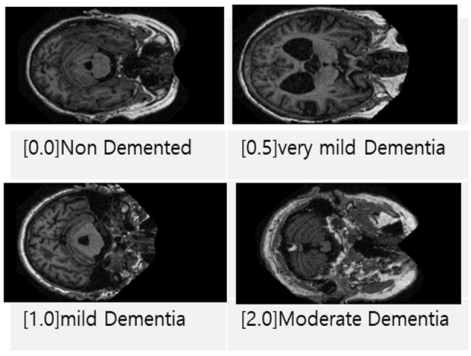

Fig. 2. OASIS Dataset

C. Local Training

The underlying framework of our employed model takes cues from VGG16. The basic network of the model we used is inspired by VGG16. VGG16 stands as a deep neural network comprising 16 layers, featuring compact convolution filters that enable it to categorize images across 1,000 classes. In our case, we have adapted this pretrained model to perform classification across four distinct categories. Fig. 3 provides a visual depiction of both the original VGG16 architecture and our tailored modification.

Regarding hyperparameters, we employed a learning rate of 1e-3, coupled with an Adam optimizer and a categorical crossentropy loss function. While local training in FL transpired over 2 and 10 epochs, centralized learning underwent experimentation with variations of 20 and 40 epochs.

When training models within a Non-IID data environment which is our experiment case no 2, we tackled class imbalance by implementing class weights during the training process.

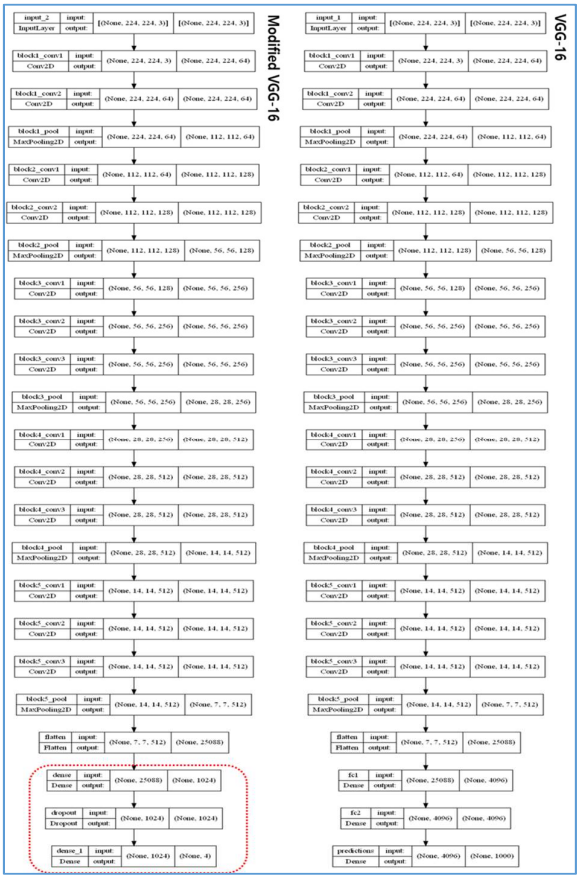

Fig. 3. Pretrained VGG and Our Modification

D. Experimental Design

We designed a couple of experimental settings to investigate FL, particularly with skewed datasets, focusing on two distinct cases: quantity-based label skew and distribution-based label skew. In the context of federated learning scenarios, the term Non-IIDness is commonly associated with substantial differences between data distributions P_i and P_j for different clients i and j. Therefore, we call the first experimental case as just "imbalanced datasets" instead of

Non-IID, since all clients have the same amount of data distributed uniformly across classes. The primary distinction emerges from the presence of distinct quantity skew across these classes.

For both experimental scenarios, we also conducted centralized approaches to facilitate meaningful comparisons and insights.

a) Experiment Case No.1 : Imbalanced Datasets

For the inherent quantity skewness across classes in real-world scenarios, we allocated the acquired OASIS images evenly among three participating clients before initiating FL. Conversely, for the centralized ML approach, all acquired data was consolidated onto a single server. Fig. 4 presents the image distribution for the FL experiment. The hardware specifications for servers and clients were same for both approaches, though the epoch was increased for the centralized ML method as mentioned in the previous section.

b) Experiment Case No.2 : Non-IID Datasets

In the second experiment, we intentionally arranged the images to simulate the skewed feature distribution of Non-IID situation. In this setup, three clients were assigned distinct data distributions spanning the four classes. Client 1 experienced a scarcity of samples for 'MildDementia' and 'ModerateDementia' classes, whereas it possessed an ample dataset of over 35,000 samples for 'NonDementia' class. We speculated that such an occurrence could be plausible if the hospital were situated in a college town with a notably high young population. In contrast, for client 2, the 'VeryMildDementia' class showcased the most substantial dataset compared to the other clients. Meanwhile, client 3 exhibited fewer samples for the 'NonDemented' and 'VeryMildDementia' classes, yet it had the highest number of samples for the 'MildDementia' and 'ModerateDementia' classes among the three clients. The imbalanced data distribution within the FL environment is visually depicted in Fig. 5.

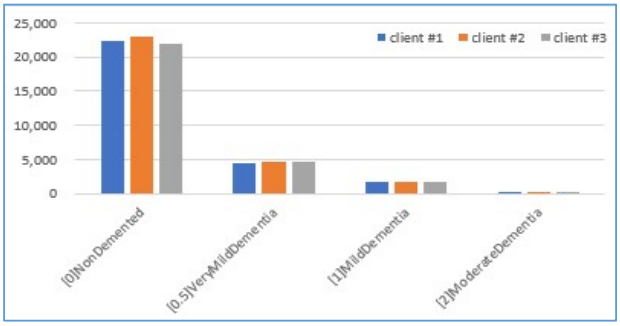

Fig. 4. Imbalanced Data Distribution for FL

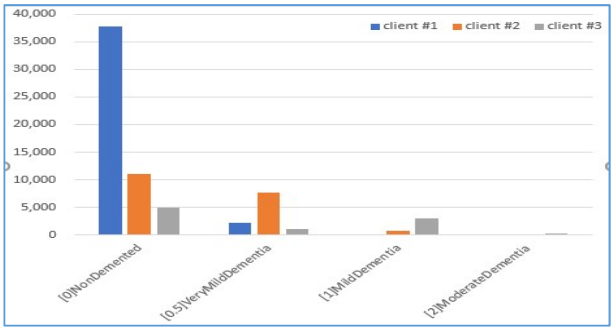

Fig. 5. Non-IID Data Distribution for FL

IV. Experimetal Result and Discussion

In this section, we provide the outcomes of our evaluations and discuss about findings from the experiments.

One of the primary objectives of the experiment case No.1 is to conduct a comprehensive comparative analysis between FL and centralized ML, focusing on accuracy and total duration. Table 1 provides a summary of the results on the imbalanced dataset, while Fig. 6 illustrates the progression of FL in terms of accuracy per federation round. The centralized ML approach demonstrates better accuracy along with a shorter execution time. Nevertheless, the performance of FL on an imbalanced dataset proves that it could potentially serve as a feasible alternative. Furthermore, an interesting observation emerges regarding accuracy not consistently improving with repeated FL rounds. For instance, as depicted in Fig 6, clients 1, 2, and 3 achieved their highest accuracy right after the first FL round.

Table 1. Result of Exp. No.1

Metrics	Federated Learning (FL Round = 20)		*Centralized Learning*	
	Epoch=2	*Epoch=20*	*Epoch=20*	*Epoch=40*
Accuracy	0.777	0.799	0.85	0.823
Loss	0.65	0.45	0.43	0.408
Durtion(sec)	31,610	104,296	24,001	47,728

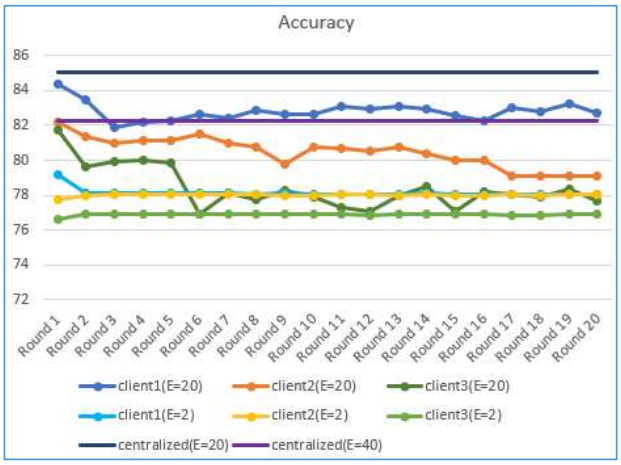

Fig. 6. Accuracy with Imbalanced Datasets

The outcomes of experiment case No.2 are visualized in Fig. 7.

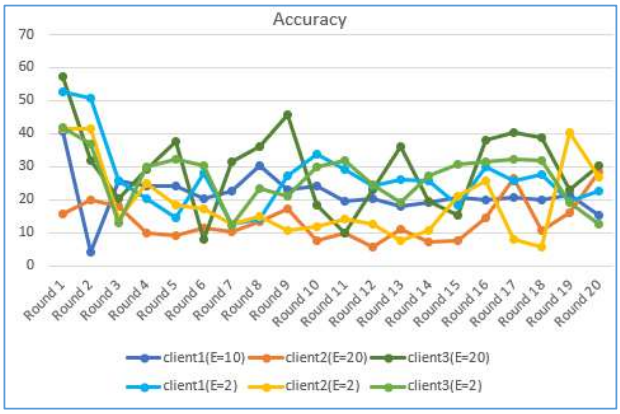

Fig. 7. Accuracy with Non-IID Datasets

As shown in the figure, during FL there were deep fluctuations in accuracy across federation rounds with Non-IID datasets. The disparity in accuracy among individual clients is more pronounced compared to the imbalanced dataset case. Interestingly, the impact of epoch variation is negligible in accuracy. Many previous studies indicated that degradation in accuracy of FL is inevitable and the observed divergences and decline in accuracy on Non-IID datasets found in our study align with the outcomes of prior research [5].

As our experimental results present, it becomes even more evident that discovering effective approaches to handle Non-IIDness within FL, particularly in the realm of medical images, is a significant challenge. We leave this for our future work.

V. CONCLUSION

In this paper, we have presented our experiments involving FL applied to medical images. Through our experiments, we conducted a thorough comparison between FL and centralized ML methodologies. Acknowledging the prevalence of Non-IIDness in FL and the challenges it brings, we also conducted a detailed exploration of the effects of Non-IID data on the classification of Alzheimer's disease stages. Our experimental results show the promising potential of FL, yet the performance degradation caused from the weight divergence in the context of Non-IID data.

ACKNOWLEDGMENT

This work was supported by Institute of Information &Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT)(2022-0-01032, Development of Collective Collaboration Intelligence Framework for Internet of Autonomous Things.

REFERENCES

[1] O. Aouedi, A. Sacco, K. Piamrat and G. Marchetto, "Handling Privacy-Sensitive Medical Data With Federated Learning: Challenges and Future Directions," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 2, pp. 790-803, Feb. 2023.

[2] E. Darzidehkalani, M. Ghasemi-Rad, P. Ooijen, ″Federated learning in medical imaging: Part I: Toward multicentral health care ecosystems,” J. Am. Coll. Radiol., pp. 969–974, 2022.

[3] H. Wang, Z. Kaplan, D. Niu and B. Li, "Optimizing federated learning on Non-IID data with reinforcement learning," IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, Toronto, ON, Canada, pp. 1698-1707, 2020

[4] J. Clerk Maxwell, “Federated learning from pre-trained models:A contrastive learning approach”, arXiv preprint arXiv:2209.10083, 2022.

[5] H. Zhu, J. Xu, S. Liu and Y. Jin, “Federated learning on non-IID data: A survey, Neurocomputing, vol. 465,pp. 371-390, 2021.

[6] Hiep Nguyen, Lam Phan, Harikrishna Warrier, Yogesh Gupta, “Federated Learning for Non-IID Data via Client Variance Reduction and Adaptive Server Update”, arXiv preprint arXiv:2207.08391, 2022.

[7] I. S. Jacobs and C. P. Bean, “Federated Learning in Medical Imaging: Part I: Toward Multicentral Health Care Ecosystems Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, pp. 271–350, 1963.

[8] Marcus, Daniel S., Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner, “Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults.” Journal of Cognitive Neuroscience 19 (9) (September): 1498–1507, 2007.

[9] M. Orouskhani, C. Zhu, S. Rostamian, F. Zadeh, M. Shafiei, and Y. Orouskhani,″Alzheimer's disease detection from structural MRI using conditional deep triplet network,″ Neuroscience Informatics, vol. 2, Issue 4, 2022.

[10] Jo T, Nho K, Saykin AJ. “Deep learning in Alzheimer's Disease: Diagnostic classification and prognostic prediction using neuroimaging data,” Front Aging Neurosci. Aug, 2019

[11] B. Khagi, B. Lee, J. -Y. Pyun and G. -R. Kwon, "CNN Models Performance Analysis on MRI images of OASIS dataset for distinction between Healthy and Alzheimer's patient," 2019 International Conference on Electronics, Information, and Communication (ICEIC), Auckland, New Zealand, pp. 1-4, 2019.

[12] Flower: A Friendly Federated Learning Framework

[13] https://www.kaggle.com/datasets/ninadaithal/imagesoasis