Medical Dataset Preparation and Privacy Preservation for Improving the healthcare facilities using Federated Learning Approach

1st Bagesh Kumar
Manipal University Jaipur
bagesh.kumar@jaipur.manipal.edu

2nd Prakhar Shukla
Indian Institute of Information Technology
Allahabad, India
prakharshukla165@gmail.com

3rd Krishna Mohan
Jaipur Engineering college and Research Centre
Jaipur, India
krishanmohank974@gmail.com

4th Akash Bharadwaj
Indian Institute of Information Technology
Allahabad, India
iit2021019@iiita.ac.in

5th Yuvraj Shivam
Indian Institute of Information Technology
Allahabad, India
iib2021006@iiita.ac.in

6th Chandan Kumar
Indian Institute of Information Technology
Allahabad, India
iit2021209@iiita.ac.in

***Abstract*—As a result of recent advances in deep learning, several breakthrough stories in modern medical diagnostics with data-driven insights into improving healthcare facilities' quality of treatment have arisen. Deep learning methods that perform well are significantly data-driven. As more data is trained, the deep learning model's performance becomes much more robust and generalizable. On the other hand, collecting medical data in a central storage system to train effective deep learning models raises concerns about privacy, ownership, and regulatory compliance. Federated learning overcomes the previous difficulties by using a deep learning model which is shared and a centralized aggregating platform. On the other side, patient data resides with the local party, assuring data confidentiality and data security. First, we give a comprehensive, up-to-date survey of federated learning research in healthcare applications. Next, we propose a solution for preparation of the medical dataset for federated learning approach from publicly available medical repositories and then apply federated averaging(FedAvg) and FedProx algorithm for aggregating across clients without accessing local private data.**

***Index Terms*—Federated learning, Deep learning, Healthcare, Medical imaging, Privacy preserving.**

I. INTRODUCTION

Artificial intelligence and deep neural networks are at the forefront of many areas, and the availability of large datasets has revolutionized this area. Deep learning models typically require millions of training datasets. Deep learning methods cannot be successfully generalized without a large and diverse training dataset. On the other hand, the lack of large datasets in medicine often interferes with machine learning applications in medicine. The lack of freely accessible large and diverse datasets was primarily due to confidentiality and privacy concerns when sharing medical history data.

979-8-3503-0511-1/23/$31.00 © 2023 IEEE

Medical images may contain personal or sensitive information about patients that cannot be shared outside the original facility, particularly if complete anonymization is not possible. The General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States, respectively, lay forth rules and limitations on the storage and sharing of personal and health information. Ethics also supports the right to complete secrecy and control over one's personal information. Large databases of medical data from many sources are therefore still mostly underutilised.

If the data from a single institution is not biased or diverse, collaboration techniques that do not require data centralization are proposed. Collaborative data sharing (CDS) or federated learning between different institutions is a way to solve this challenge. Federated learning is a machine learning methodology that overcomes the need for a single data pool to train models when coupled with a client-server architecture. Instead, you can train the model locally on the device and then transfer the output from many devices to the central model. This system facilitates data to be shared among multiple institutions and researchers while preserving patient privacy.

In this paper, we aim to investigate federated learning (FL) as a collaborative learning paradigm in medical field where privacy is the utmost concern that need to be addressed. We will use public data repository in the medical field such as breast cancer imaging, malaria infected cells etc. to simulate distributed environment for our project.

II. LITERATURE SURVEY

Leal et al.,2022 [1] propose federated learning in multi-task and multi-domain settings utilizing two different experiments.

In Experiment 1, they executed organ localization using the U-Net model for segmentation and localization; in [1] this model is shown for localization of the kidneys . In Experiment 2, they used a federated learning setup with two client nodes and one central server to perform multi-organ tumor segmentation. The conclusion of an experiment is a [1] similarity of 65% for organ segmenting and 0.79 for organ localisation.

In[2] Zeyuan Allen- Zhu shows that With running times exponential in n and L, SGD achieves 100% training accuracy in classification problems or minimises regression loss in linear convergence speed. All smooth and potentially non-convex loss functions, as well as the well-known but non-smooth ReLU activation, may be solved using our method. Our hypothesis, at least in terms of network design, holds for fully connected neural networks (FCNN), convolutional neural networks (CNN), and residual neural networks (ResNet)

In[3] the authors of this study put up an ethical preposition for exploiting and sharing medical/patients' data in the creation of applications for artificial intelligence (AI). The main reason for collecting the data is satisfied when it is used to offer care, according to the philosophical basis. Pateints data should then be viewed as a type of public good, to be exploited for the advantage of upcoming patients.

In their essay from 2013, Faden et al. made the case that everyone involved in the healthcare/Medicine system, including patients, has a moral duty to help make it better. The authors create the method to get over challenges with the secondary usage of clinical data for AI applications. The authors specifically advise that everyone who has access to clinical data and who serves as a data steward has a duty of care to patients to protect their privacy as well as a duty of care to the general public to make sure that the data is accessible for the creation of knowledge and tools that will benefit both current and future patients. According to the authors, it is immoral for doctors to "give/sell" their patients' data to outside parties by granting them access to their medical records, especially when exclusion agreements are in place, in exchange for a payment in cash or in-kind that exceeds the costs incurred. The authors further suggest that in cases where getting patient approval would be excessively expensive or difficult, as long as systems are in place to ensure that ethical norms are carefully observed, patient consent should not be necessary before the data are used for secondary purposes.Instead of debating who "owns" the data—patients or provider organizations—the authors suggest that all individuals who contact with or have control over the clinical data have an obligation to use it for the benefit of present and future patients as well as society. In a very non-IID FL situation, the newly developed technique remarkably stabilises model training for the diagnosis of many chest illnesses in chest X-ray pictures. According to the outcomes of the federated trainings for people with intellectual disabilities (IID) and non-IID using it, the proposed method may encourage organisations or researchers to develop better systems to extract value from data with regard to data privacy, not only for the healthcare industry but also for other fields.

Adnan and kalra et al [3] studied the differential private federated learning framework methods and its effects of IID and non-IID distribution. In[3] they have evaluated the performance of different histopathology images and its effect in source domains. Adnan and kalra et al have used FedAvg(or federated averaging) algorithms for the problem of differential privacy.The result concluded with efficient use of FL for real -world data for using both ID and non-IID data distribution by Adnan and kalra et al in [3].

Ming y.Lu et al.,2021 [4] presented a computation pathology algorithm with privacy FL on medical images. In [4] they used federated learning to improve accuracy with a weakly supervised deep learning model. For this in [4] instance learning framework they have used for the privacy federated learning on data medical images. Ming y.Lu et al have applied the algorithm based on multiple instance learning for histology base classifications and prediction. The dataset for evaluating proposed FL framework on weakly supervised learning, they have used RCC and BRCA dataset. Results of [4] provide an opportunity to integrate WSI datasets and train a more robust model that gives institutions control over data while maintaining privacy of data.

Apler and Murat et al.,2021[5] proposed an improved performance-based analysis on medical images in non-IID settings using image augmentation in federated learning. This paper[5] presents a naive-based approved for Non-IID FL data in image augmenting. The method provided by the paper[5] assists systems in obtaining values for data privacy in healthcare. Apler and Murat et al. proposed a novel method for mitigating the performance degradation caused by nonidentical data implementation in FL. FedAvg is used for model augmentation in the image augmentation process, which artificially trains images. So, using the method described in the paper, the client sends a number of label samples to the server, after which the server selects a maximum number of samples from the clients for labeling and sends this information to each client. Following that, the client applies a transformation to the local dataset and increases the number of samples to send during the augmentation phase. The results of IID and non-IID are presented in [5]paper with a model accuracy of 83.22 percent. The models we implemented are inspired by these papers.

With elastic averaging in deep learning, sgd.Sixin Zhang, Anna E. Choromanska, and Yann LeCun investigated the issue we encountered when doing stochastic optimisation with communication restrictions in the parallel computing environment for deep learning in their article Advances in Neural Information Processing System. Here, a unique method for task coordination and communication amongst concurrent processes (local workers) is provided, where task coordination and communication are based on an elastic force linking the parameters they compute with a crucial variable recorded by the parameter server (master). The method enables the local variables to diverge further from the central variable, allowing the local workers to perform more exploration, by restricting communication between the local employees and the master. They assess the robustness of the asynchronous round-robin

method and contrast it with the more popular parallelized Alternating Direction Method of Multipliers (ADMM). They demonstrate that, in contrast to ADMM, Elastic Averaging SGD (EASGD) stability is assured when a simple stability requirement is met. Additionally, they suggest the asynchronous and synchronous implementations of our technique based on momentum. Convolutional neural networks are trained using an asynchronous version of the technique on the CIFAR and ImageNet datasets for image categorization.

Federated learning has the potential to transform how AI models are taught, with the advantages spreading throughout the healthcare ecosystem.

Access to secure, inter-institutional data would improve collaboration and benefit bigger hospital networks. High-level AI algorithms would be available to smaller hospitals in rural and community settings. Clinicians would gain access to more powerful AI algorithms based on information from a broader patient demographic for a certain therapeutic area as well as from unusual circumstances that they would not have seen locally. If they did not agree with the outcomes, they would also be permitted to contribute to the continued training of these algorithms.Research institutions would be able to focus their efforts on really clinical requirements, based on a wide range of real-world data, as opposed to depending on a finite supply of free datasets.

III. PROBLEM STATEMENT AND OBJECTIVE

A. Problem Statement

Given a Breast Cancer Wisconsin (Diagnostic) dataset, build a federated learning based generalizable model eliminating privacy concerns.

B. Objective

Building a federated learning based machine learning model for the given medical dataset which can generalize well across client's data. We will use Breast Cancer Wisconsin (Diagnostic) dataset, available publicly on UCI Machine Learning Repository, to simulate distributed environment. We use FedAvg and FedProx federated learning algorithms to eliminate confidentiality and privacy concerns of patients.

IV. DATASET

Breast Cancer Wisconsin (Diagnostic) Dataset: This dataset is a classification dataset, which records the measurements for breast cancer cases and is available publicly on UCI Machine Learning Repository.The features of the dataset are derived from fine needle aspirate(FNA) of a breast mass's digitized and purified images.They describe the characteristics of the observable cell nuclei in the image.

Attribute Information:

- Id-No.
- Diagnosis 3-32)

Ten real-valued features are computed for each cell nucleus:

- radius
- perimeter
- area
- texture
- smoothness
- concavity
- symmetry
- fractal dimension
- concave points (number of concave portions of the contour)
- compactness ($perimeter^2$ / area - 1.0)

A total of 30 features were created by calculating the average, standard deviation, and "worst" feature—the mean of the three highest values—for each image. For instance, fields 3 and 13 and 23, respectively, provide the Mean Radius and Radius SE. Four significant digits are used for all feature values.

V. PROPOSED METHODOLOGY

Our proposed method includes preparation of medical dataset for federated learning and then using Federated Averaging(FedAvg) and FedProx algorithm for averaging across various clients while securing their data privacy locally. Every client trains a copy of local centralized model and reports back the weight updates to the server which aggregates them to train a generalizable model.

Preparation of medical dataset for federated learning involves taking publicly available normal medical dataset such as Breast Cancer Wisconsin (Diagnostic) Data and then randomly partitioning the dataset based on number of clients in the model which is then fed into the model after some preprocessing. Random partitioning of dataset allows different distribution of data for each client where one client may have scarcity of training example for a particular class and other clients may have enough training example for that class which simulates real environments.

Federated Averaging(FedAvg) : Federated learning can be used to train models across several institutions without exposing the patient data to the centralized server. In federated learning every client trains the local copy of centralized model and then shares the update of model parameters with the centralized server for aggregation across clients to build a generalizable model. Mathematical formulation of federated learning is :

$$min_{\omega \in R} f(\omega) where f(\omega) = \frac{1}{n} \sum_{i=1}^{n} f_i(\omega)$$

The loss function of the i-th client with regard to local data is represented by $fi(\omega)$, while the loss function overall over n clients is represented by $f(\omega)$. To minimise the overall loss is the goal. In each iteration, the client receives the most recent model parameters from a central server and uses stochastic gradient descent to compute the gradient over its local data to update the model parameters. The central server then averages the model parameters from clients' updates and updates the central model.

Algorithm 1 $FederatedAveraging$. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and n is the learning rate.

Server executes:
initialize w_0
for each round t = 1 , 2, ... **do**
$m \leftarrow max(C \cdot K, 1)$
$S_t \leftarrow$ (random set of m clients)
for each client $k \in S_t$ **do**
$w^{k}_{t+1} \leftarrow ClientUpdate(k, w_t)$
$w^{k}_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w^{k}_{t+1}$
end for
end for
ClientUpdate: (k, w) It Run on client k
B $\leftarrow$ ($split$P$_k$ into batches of size B)
for each local epoch i from 1 to E **do**
for batch b $\in$ B **do**
w $\leftarrow w - n$(w;b)
end for
end for
=0

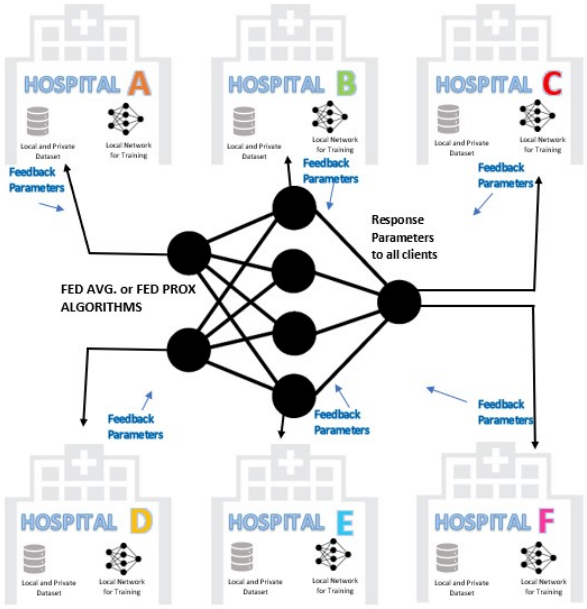

Fig. 1. Federated Learning Architecture

FedProx : FedProx can be seen as generalization and re-parameterization of well known FedAvg technique, and provides more stable and accurate convergence behaviour in federated networks which helps us to tackle heterogeneous setting. Heterogeneity can be expressed both in terms of systems heterogeneity which indicates that each client may have different systems resource as well as statistical heterogeneity which indicates that clients may have non-identical heterogeneous data. FedProx contains additional proximal term in the loss function which contribute to model's stability.

$$min_{\omega} h_k(\omega; \omega^t) = F_k(\omega) + \frac{\mu}{2}||\omega - \omega^t||^2$$

Here $F_k(\omega)$ represents the loss function of device k and $\frac{\mu}{2}||\omega - \omega^t||^2$ represents the proximal term which helps in two aspects such that it prevents the local updates on each client to diverge from the global model due to heterogeneous data and also allows to account for variable amount of work done on each client because of systems heterogeneity. The objective of the client is to minimize the total loss function along with proximal term with its local solver and then update the central server of the local model parameters which then aggregates them to update the central model.

Experiment 1: In this Experiment, there are 13 hospitals each with 35 randomly selected samples. So in order to simulate the process, we start by dividing the dataset in 13 parts, randomly and then we send each part to each hospital. Each part is composed of the features, the diagnosis for disease. We are using the FedAvg algorithm here to test accuracy of the final model with the testing dataset we separated from the beginning. The outputs are shown below:

Observation: In the above results, we can see the training accuracy for 13 hospitals each in different colour. We are achieving 92.98% accuracy on test set after 100 epochs containing 114 samples. FedAvg achieves very high accuracy because dataset is normalized and local optima for each clients is same as global optima hence after each round of aggregation each clients have more or less same accuracy.

Algorithm 2 $FedProx$. (Proposed Framework)

input: $K, T, \mu, w^0, N, p_k, k = 1, ..., N$
for t=0, ..., T-1 **do**
Server selects a subset S_t of K devices at random (each device k is chosen with probability p_k)
Server sends W^t to all chosen devices
Each chosen deivce k $\in$S$_t$ finds a w_k^{t+1} which is a γ_k^t inexact minimizer of : $w_k^{t+1} \approx argmin_w\ h_k(w; w^t) = F_k(w_+ \frac{\mu}{2}||w - w^t||^2$
Each device k $\in S_t sends$w$_k^{t+1}$ back to the server
Server aggregates the w's as w$^{k+1} = \frac{1}{k}\sum_{k_t}\ w_k^{t+1}$
end for=0

Training/Validation Loss:

Experiment 2: In experiment 2, to obtain a better result and to improve significantly more stable and accurate convergence behavior relative to FedAvg, we will be using FedProx algorithm without system heterogeneity. Similar to previous experiment we again divide the datatset into 13 parts and send each part to a corresponding hospital. The outputs are shown below:

Observation: In the above results, we can see that accuracy is 92.11% for the test set after training for 100 epochs. As the dataset is same for each client so there is neither data

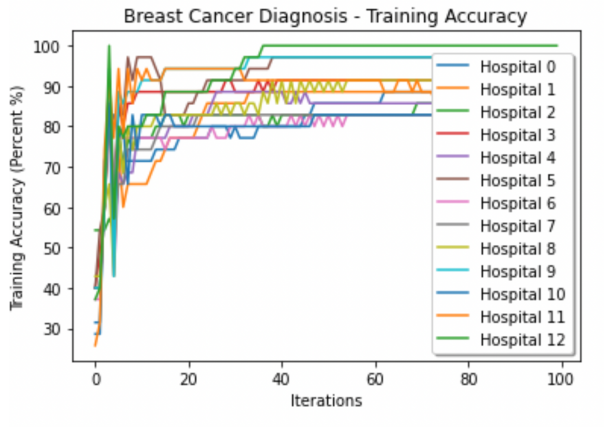

Fig. 2. Experiment 1: Training Accuracy graph

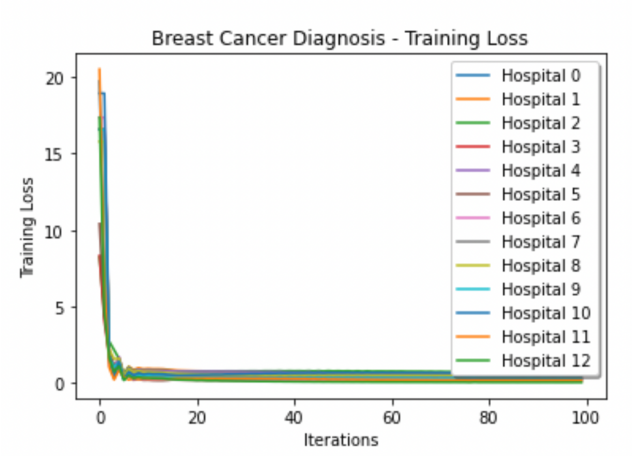

Fig. 3. Experiment 1: Training Loss graph

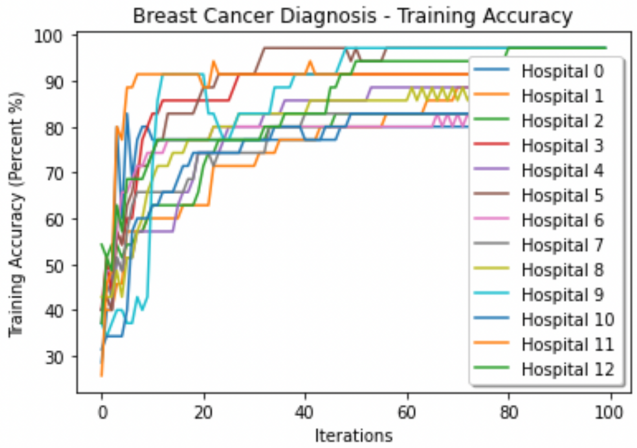

Fig. 4. Experiment 2: Training Accuracy graph

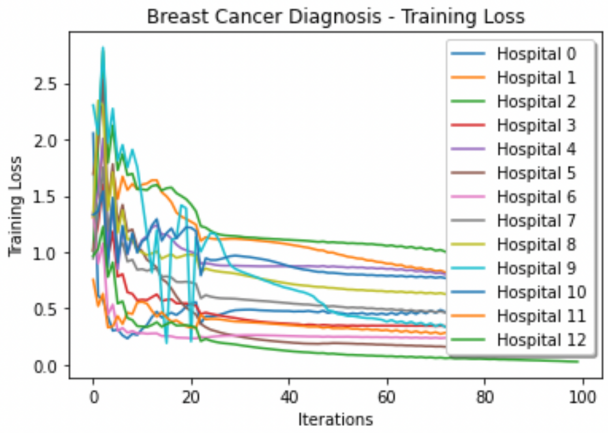

Fig. 5. Experiment 2: Training Loss graph

heterogeneity nor systems heterogeneity for this experiment. But this algorithm improves performance considerably when data is heterogeneous. Extra proximal term in the loss function penalizes the local model updates for diverging far from global model to maximise their local optima and ignoring the global optima and hence builds the better generalizable model.

Training/Validation Loss:

Experiment 3: In this experiment 3, we have used FedProx Algorithm along with systems heterogeneity where each hospitals performs variable number of local epochs (depending on hospital's system capability) before aggregation on central server. Similar to previous experiments we again divide the datatset into 13 parts and send each part to a corresponding hospital. The outputs are given below:

Observation: In this experiment, we can see that accuracy is 91.23% on the test set after training for 100 epochs. We have used heterogeneous setting for system so convergence for each client is different as we can see in Fig.7. Clients which perform more round of local epochs converge faster to their local optimum whereas clients which performs less round of local epochs before aggregation converge slowly to their local optimum.

Training/Validation Loss:

VI. RESULTS

We train our simple two layered neural network model based on Federated setting for 100 epochs on 'Breast Cancer (Diagnostic) dataset from Kaggle'. We have used three different experiment to show training accuracy and training loss of model in which we have used FedAvg and FedProx Algorithm, In order to simulate federated setting, we start by dividing the dataset in 13 parts randomly, and then we send each part to each corresponding hospital. Each part is composed of the features and the diagnosis for disease. Final highest Testing Accuracy is 92.98% which is produced by FedAvg algorithm. FedProx algorithm tends to produce better results in more heterogeneous setting.

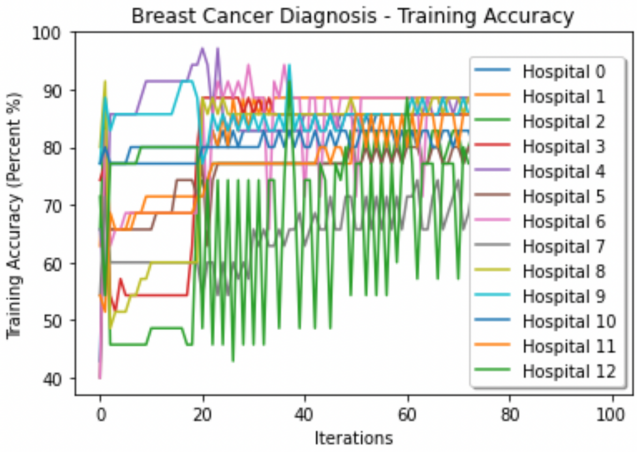

Fig. 6. Experiment 3: Training Accuracy graph

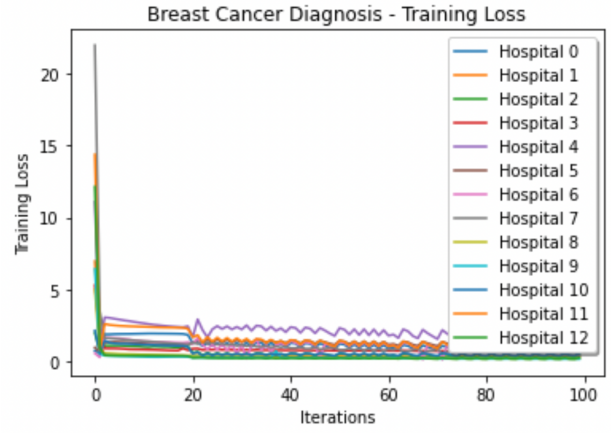

Fig. 7. Experiment 3: Training Loss graph

VII. CONCLUSION

In this paper, we have applied Federated learning on medical dataset because medical field suffers from lack of publicly available diverse datasets because of confidentiality and privacy concerns. The medical datasets are usually not available for public use which makes it difficult for the training of various medical models. Therefore, we adopted the method of federated learning which helps the clients to train the model on their particular local machine without publically disclosing the datasets and hence maintaining the privacy and confidentiality.First, we give a comprehensive, up-to-date survey of federated learning research in healthcare applications. Then we start from preparation of medical dataset for federated learning algorithm and we use public repository of Breast cancer Wisconsin (diagnostic) medical data and random partitioning to simulate distributed and unbalanced environment that is different for every client representing different systems the model is trained on, then we apply Federated Averaging(FedAvg) and FedProx algorithms for averaging across clients where model is locally trained for each client and updates are transferred to server for building a better generalizable model eliminating the privacy concerns. FedProx algorithm performed better than the FedAvg as the dataset was quite heterogeneous and FedProx also allows clients to optimize a regularized loss with a proximal term which helps the client models to not divert immensely from the main model. This way we were able to come up with a better of the two algorithms and eliminated privacy concerns for the clients.

REFERENCES

[1] Vishwa Parekh ,Shuhao Lai ,Vladimir Braverman , Jeff Leal, Steven Rowe , Jay J. Pillai ,Michael A Jacobs ,(2021) Cross-Domain Federated Learning in Medical Imaging .Proceedings of Machine Learning Research, 2022. arxiv.org/pdf/2112.10001v1.pdf

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via overparameterization. arXiv preprint arXiv:1811.03962, 2018.

[3] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas. (2017). Communication-Efficient Learning of ¨ Deep Networks from Decentralized Data. arxiv.org/pdf/1602.05629.pdf

[4] Larson, D.B., Magnus, D.C., Lungren, M.P., Shah, N.H., Langlotz, C.P.: Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. Radiology 295, 192536 (2020)

[5] Mohammed Adnan ,Shivam Kalra , Jesse C. Cresswell , GrahamW.Taylor Hamid R.Tizhoosh ,(2012 -2022) .Federated learning and diferential privacy for medical image analysis .

[6] Lu, M.Y., Chen, R.J., Kong, D., Lipkova, J., Singh, R., Williamson, D.F.K., Chen, T.Y. and Mahmood, F.Medical Image Analysis ,Federated learning for computational pathology on gigapixel whole slide images (2022).

[7] Alper Emin Cetinkaya ,Murat Akin,Seref Sagiroglu, Improving Performance of Federated Learning based Medical Image Analysis in Non-IID Settings using Image Augmentation (2021,)Published in: 2021 International Conference on Information Security and Cryptology (ISCTURKEY).

[8] Li, T., Sahu, A., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V..(2018). Federated Optimization in Heterogeneous Networks.

[9] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al. Large scale distributed deep networks. In Advances in neural information processing systems, pp. 1223–1231, 2012.

[10] Sixin Zhang, Anna E Choromanska, and Yann LeCun. Deep learning with elastic averaging sgd. In Advances in Neural Information Processing Systems, pp. 685–693, 2015.

[11] Blake Woodworth, Jialei Wang, Brendan McMahan, and Nathan Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. arXiv preprint arXiv:1805.10222, 2018.

[12] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. arXiv preprint arXiv:1808.07576, 2018.

[13] Fei Chen, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning for recommendation. arXiv preprint arXiv:1802.07876, 2018.

[14] Andrew Hard, Chloé M Kiddon, Daniel Ramage, Francoise Beaufays, Hubert Eichner, Kanishka Rao, Rajiv Mathews, and Sean Augenstein. Federated learning for mobile keyboard prediction, 2018.

[15] Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. Advances in Neural Information Processing Systems, 33, 2020.

[16] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In Advances in Neural Information Processing Systems, pp. 4424–4434, 2017.

[17] 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, 1–7 (2020)