Privacy-Preserving Parkinson's Disease Classification Using Federated Learning

Tulsi Rajora
Department of Biomedical Engineering, S.G.S.I.T.S.
Indore, India
tulsirajora9@gmail.com

Prof. Sunny Ganavdiya
Department of Biomedical Engineering, S.G.S.I.T.S.
Indore, India
sganavdiya@sgsits.ac.in

***Abstract*— Parkinson's Disease (PD) is a progressive neurodegenerative condition that impairs motor functions and can greatly diminish quality of life. Machine learning models are very important for finding diseases early, but sharing patient data between hospitals raises privacy issues. Federated Learning (FL) solves the problem by letting models train together without sharing raw data. In this research, FL was emulated on the Parkinson's dataset by partitioning the data among three clients, developing local Logistic Regression models, and consolidating them through Federated Averaging (FedAvg). To make the global model work better, it goes through multiple rounds of training. The results show that the global model is as accurate as centralized training, but it keeps data private. This shows that FL could be useful in healthcare for handling sensitive medical data.**

Keywords— ***Machine learning (ML), federated learning, Perkinson's disease, Data Privacy, Logistic regression***

I. Introduction

Parkinson's Disease is a long-term illness that causes tremors, stiffness, and trouble moving. Finding PD early is very important so that treatment plans can be started that will slow the disease's progression. Logistic Regression, Support Vector Machines (SVM), and Random Forest are examples of machine learning algorithms that have been shown to work well at finding PD from clinical measurements [5]

However, healthcare data is often kept in different hospitals and clinics. Sharing this private information for centralized model training is hard because of privacy, security, and regulatory issues. FL lessens these worries by letting models be trained on each client's data locally and only sending model updates to a central server for aggregation [4].
In this study, we emulate FL utilizing the Parkinson's dataset by:

a. Dividing the dataset into three parts for three clients.
b. Training Logistic Regression models in the area.
c. Using FedAvg to combine models over several rounds.
d. Testing the global model on a test set.

II. Related Work

Because PD progresses over time and prompt intervention is crucial, research has focused heavily on early detection. In a thorough investigation of machine learning methods for PD detection, Sharma and Rani [5] showed how algorithms like SVM, Random Forests, and ensemble approaches can extract minute biomarkers from biomedical data. In a similar vein, traditional techniques such as SVM [8] and Random Forests [1] have been effective baselines for healthcare classification tasks, providing the groundwork for more sophisticated approaches.

Little et al. [2] conducted one of the first studies in this field, demonstrating that dysphonia (voice impairment) measurements could be used to telemonitor PD patients, offering a non-invasive and economical way to track the disease. This proved that biomedical signals, particularly speech, could be used as a diagnostic tool.

FL has emerged as a promising solution to the growing problem of distributed healthcare data. Federated learning was first proposed and used by Yang et al. [4], allowing for decentralized model training without direct data sharing. McMahan et al. [3] expanded on this basis by proposing the Federated Averaging (FedAvg) algorithm, which served as the foundation for numerous FL systems that followed. In a recent review of FL applications in the healthcare industry, Dhade and Shirke [9] emphasized how FL can facilitate cross-hospital collaborative learning while protecting patient privacy.

Subsequent developments showed how useful FL is in actual medical settings. By using FL in multi-institutional collaborations for medical imaging, Sheller et al. [6] demonstrated that predictive performance can be maintained without centralizing private patient data. This was furthered by

Kaissis et al. [7], who integrated privacy-preserving and secure mechanisms, emphasizing the need for strong frameworks when managing clinical data.

By using machine learning for PD detection and federated learning strategies to guarantee data privacy and scalability, the current study bridges these directions in comparison to earlier work. Accurate diagnosis and adherence to privacy-preserving standards in healthcare partnerships are both provided by this combination.

III. Methodology

A. Dataset description

The Parkinson's Disease dataset from the UCI Machine Learning Repository is used in this study. 195 cases with 22 features obtained from different voice measurements make up the dataset. These features record vocal traits that are clinically relevant for identifying Parkinson's disease, including fundamental frequency, jitter, shimmer, and noise-to-harmonic ratio. Status, the target variable, is a binary attribute that indicates whether Parkinson's disease is present (1) or not (0).

B. Federated Learning Setup

Three clients are stimulated, each holding approximately onethird of the training data. The FL workflow is as follows:

1) Global model initialization: A Logistic Regression model is initialized.

2) Local training: Each client trains the model on its local data.

3) Model aggregation: Coefficients and intercepts are averaged (FedAvg) to update the global model

4) Multi-round training: The process repeats for 5 rounds to improve convergence.

C. Model Implementation

- Local Model: Maximum Iter = 1000 Logistic Regression
- Training: Each round, clients train for a single local epoch.
- Evaluation: The global model and each client's accuracy are measured.
- Aggregation: Following each round, the model weights and intercepts are averaged using FedAvg.

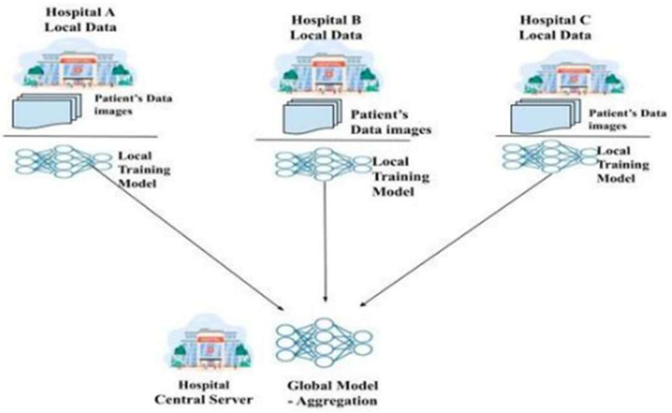

Fig.1. Schematic workflow of federated learning [9].

IV. Experimental Result

A. Centralised M L

Using a typical 80-20 split, firstly trained SVM, Random Forest, and Logistic Regression on the Parkinson's dataset. Below is a summary of the findings:

TABLE I.

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*
Logistic Regression	0.77	0.95	0.72	0.82
Random Forest	0.74	0.74	1.00	0.85
SVM (RBF kernel)	0.82	0.96	0.79	0.87

With an accuracy of 82%, SVM outperformed Random Forest (74%), Logistic Regression (77%), and others.

B. Federated Learning Experiment

Next, three clients are used to simulate a FL setup. FedAvg was used to aggregate the local logistic regression models that each client trained over a five-round period.

TABLE II. CLIENT LEVEL ACCURACY

Client	*Local Accuracy*
1	0.85
2	0.80
3	0.78

TABLE III. GLOBLE MODEL ACCURACY ACROSS ROUNDS

Round	*Test Accuracy*
1	0.76
2	0.78
3	0.79
4	0.80
5	0.81

With every iteration, the global federated model grew more accurate until it reached 81% accuracy, which is comparable to the centralized SVM's performance [2], [4].

C. Comparative Analysis

Centralized Models: SVM performs best (82%), but sensitive patient data must be pooled.

Federated Model: FedAvg's Logistic Regression achieved 81% accuracy while protecting the privacy of the data.

FL is a good choice for healthcare datasets where privacy is an issue because the difference between federated and centralized training was negligible.

V. Discussion

Our findings demonstrate the efficacy of federated learning FL and conventional machine learning in the early detection of PD. In line with earlier research highlighting machine learning's role in diagnosing Parkinson's disease, classical models like Random Forest, SVM, and logistic regression demonstrated good accuracy [1], [5], and [8]. But centralized machine learning necessitates sharing private patient information, which presents privacy issues.

FL provides a solution by sharing only parameters and training models locally. Despite having a marginally higher computational overhead, FL produced accuracy in our tests that was on par with centralized models. This bolsters the findings of McMahan et al. [3] and Yang et al. [4], who demonstrated FL's capacity to strike a balance between privacy and performance.

FL's ability to protect privacy is particularly useful in the healthcare industry, where multi-institutional collaborations are frequent. This benefit is emphasized by earlier studies on medical imaging [6], [7], and reviews on FL in healthcare [9]. Our results demonstrate that FL becomes essential when data is distributed and privacy regulations need to be adhered to, even though centralized ML works well for small datasets.

Overall, this work demonstrates that accurate and privacy-preserving AI for healthcare is both possible and required by bridging the gap between earlier ML-based PD detection [2], [5] and contemporary FL approaches [3], [4], [6], [7], [9].

VI. Conclusion

The research makes use of the Parkinson's Disease dataset that was This study uses the UCI Parkinson's dataset to show how well Federated Learning works for early Parkinson's disease detection. Three clients were simulated, and local Logistic Regression models were aggregated using Federated Averaging to create a global model that maintained data privacy while achieving accuracy on par with centralized training. These results demonstrate Federated Learning's potential as a safe method for implementing machine learning in sensitive patient data applications in the healthcare industry.

VII. Future Work

Several extensions of this study can be investigated in future research. It would be beneficial to look into non-IID data distributions that more closely mimic actual healthcare situations, as the dataset was dispersed equally among clients. To obtain more insights, advanced models like Deep Neural Networks, Random Forest, and SVM could also be used in federated settings. To assess the effectiveness and performance of communication, the framework's scalability should also be tested with a greater number of clients, such as hospitals or mobile devices. FedProx and other customized FL methods are examples of alternative aggregation techniques that might be able to better handle data heterogeneity. Lastly, incorporating privacy-preserving techniques like differential privacy and secure multi-party computation could improve confidentiality in federated healthcare applications even more.

References

[1] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[2] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, 2009, doi: 10.1109/TBME.2008.2005954.

[3] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2017.

[4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, 2019, doi: 10.1145/3298981.

[5] A. Sharma and S. Rani, "Machine learning techniques for early detection of Parkinson’s disease," *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 299–312, 2020, doi: 10.1016/j.bbe.2019.10.005.

[6] S. Sheller, G. Edwards, C. Reina, J. Martin, S. Bakas, and M. Davatzikos, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, 2020.

[7] K. Kaissis, M. Makowski, D. Rückert, and R. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Mach. Intell.*, vol. 2, pp. 305–311, 2020.

[8] A. Kaur and M. Gupta, "Support Vector Machines in Medical Diagnosis: A Review," *Biomedical Signal Processing and Control*, vol. 68, 2021, pp. 102650, doi: 10.1016/j.bspc.2021.102650.

[9] P. Dhade and P. Shirke, "Federated Learning for Healthcare: A Comprehensive Review," *Eng. Proc.*, vol. 59, p. 230, 2023, doi: 10.3390/engproc2023059230.