Adaptive Federated Healthcare System with Model Compression and Lifelong Learning

By: Moronfoluwa Akintola

https://github.com/Foluwa/Adaptive_Federated_Healthcare_System

This work was supported by the Engineering and Physical Sciences Research Council [Grant EP/X031012/1]

INTRODUCTION

Healthcare datasets are fragmented and highly sensitive. Centralised machine learning poses risks of privacy breaches. Federated Learning (FL) enables training across hospitals and devices without data sharing. However, models are often too large for wearables or low-power hospital servers, and data evolves. This project integrates Federated Learning with model compression and lifelong learning to ensure scalable, adaptive, and secure healthcare AI.

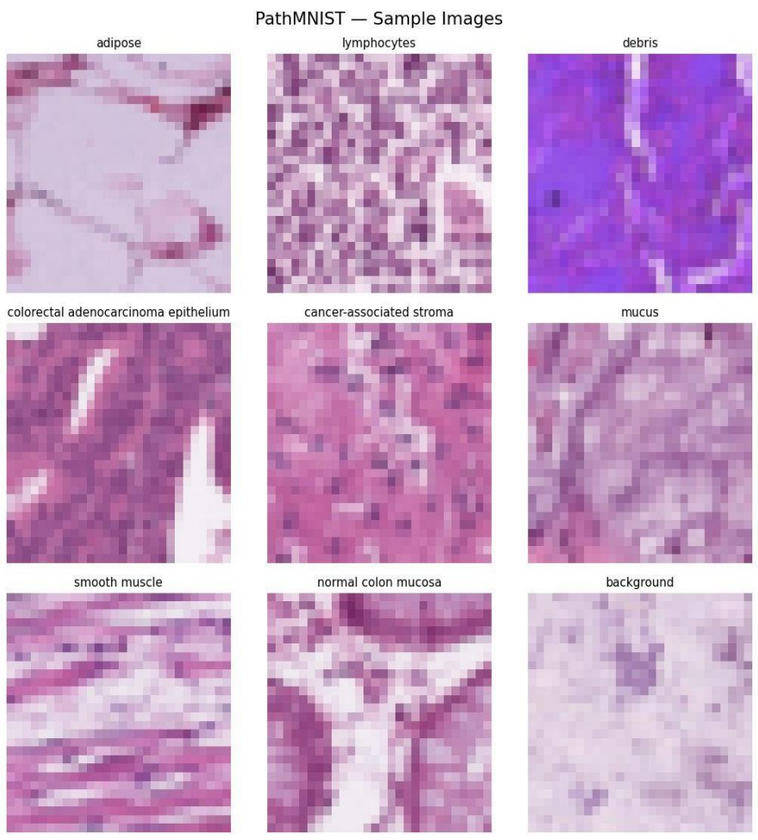

Fig. 1: PathMNIST Sample histopathology tiles

Sample histopathology from PathMNIST, showing diverse tissue types such as adipose, smooth muscle, lymphocytes, and colorectal adenocarcinoma epithelium used for federated model training.

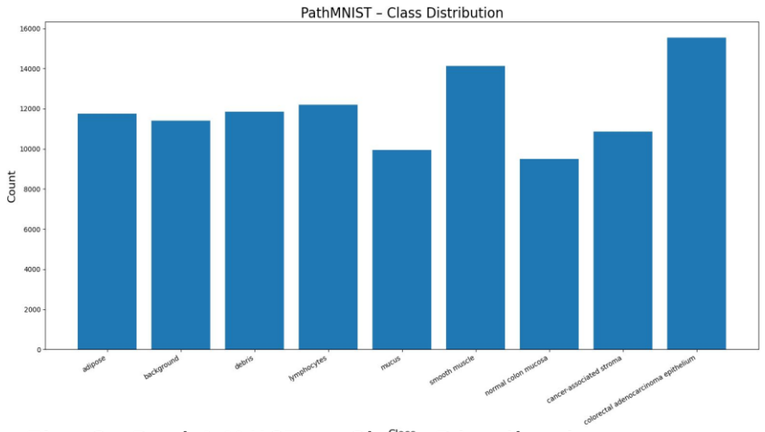

Fig. 2: PathMNIST – Class Distribution

Class distribution of the PathMNIST dataset, showing variation in sample counts across nine tissue types, with colorectal adenocarcinoma epithelium and smooth muscle being the most represented

METHODS

- Federated Learning: Clients train CNNs on PathMNIST locally and server aggregates with FedAvg (Flower).
- Model Compression: Transfer learning + pruning + quantisation to produce client-specific lightweight models.
- Lifelong Learning: Elastic Weight Consolidation (EWC) reduces catastrophic forgetting across sequential client tasks.
- Evaluation: Accuracy, F1, AUC and (when implemented) latency on simulated device

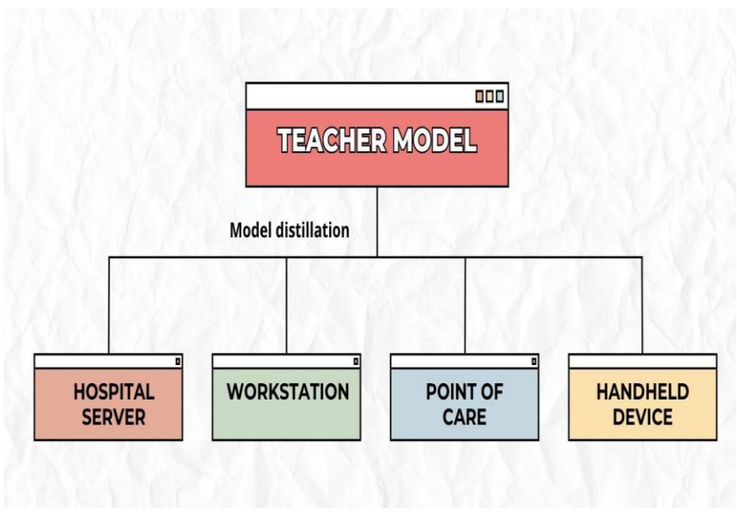

Fig. 3. Knowledge distillation from a central teacher model to resource-specific student models

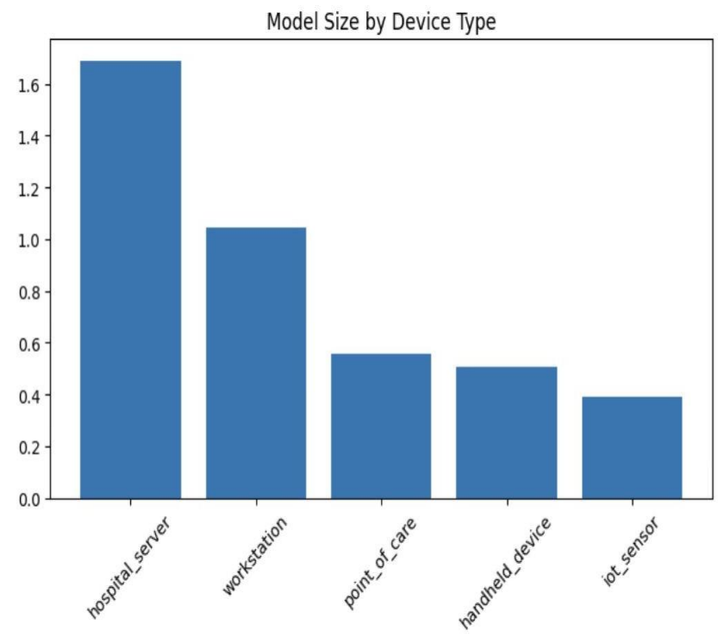

Fig. 4. Model size reduction across device types

FedAvg achieved teacher-level accuracy (0.804) on PathMNIST, while adapter fine-tuning with EWC maintained stable per-client performance (mean ≈ 0.629) across specialty and transfer tasks. Compression further reduced model size and latency, enabling deployment on devices from hospital servers to handhelds without major accuracy loss.

RESULTS

Ten virtual clients (non-IID partitions) train locally for 10 epochs per round with AdamW, and the server aggregates via FedAvg. We evaluate with Accuracy/F1/AUC, sequential forgetting metrics, and latency/size for device profiles (hospital serve, workstation, point-of-care, handheld/IoT).

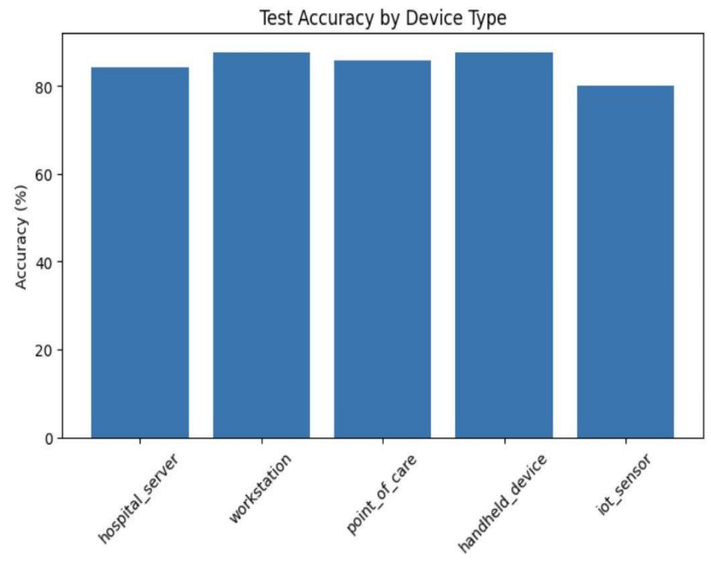

Fig. 5. Test accuracy of compressed student models across different device types

Model Type	Total Parameters	Model Size (MB)	Accuracy (%)	F1 Score	Inference Time (ms)
Hospital Server	442,729	1.69	89.54	0.8925	54.8
Workstation	274,353	1.05	88.38	0.8811	38.0
Point of Care	146,371	0.56	86.53	0.8618	26.1
Handheld Device	132,717	0.51	86.56	0.8625	29.9

Table. 1. Performance comparison of compressed models across device types,

The pipeline keeps data on-device, reducing privacy risk, and produces fast, small models suitable for each device or institution.

REFERENCES

1. Flower. Flower: A friendly federated learning framework [Internet]. Available from: https://flower.ai/ (accessed 2025 Aug 20).

2. PyTorch. PyTorch: An open source machine learning framework [Internet]. Available from: https://pytorch.org/ (accessed 2025 Aug 20).

3. Xu Q, Lim GL, Liu P, Wu M, Chen Z. A knowledge distillation-based deep learning model compression approach for human activity recognition