COVID-19: FEDERATED LEARNING FOR PRIVACY PRESERVING MULTI-INSTITUTIONAL COLLABORATION

We describe a study among 10 institutions using Federated Learning (FL), a novel paradigm for data-private multi-institutional collaborations, where model-learning leverages available data without sharing data between institutions. This approach resulted in models reaching 99% of the model quality achieved with centralized data. These encouraging results are driving a project in Spain, where Intel, local partners and three hospitals establish an institutional federation to develop an open source Artificial Intelligence (AI) model to diagnose COVID-19 with lung scans without sharing patient data.

The current pandemic has shown a light on issues from the lack of sharing of healthcare data for public health and medical research. Privacy is one of the issues raised by institutions as reasons they do not share data that would be valuable for that research.

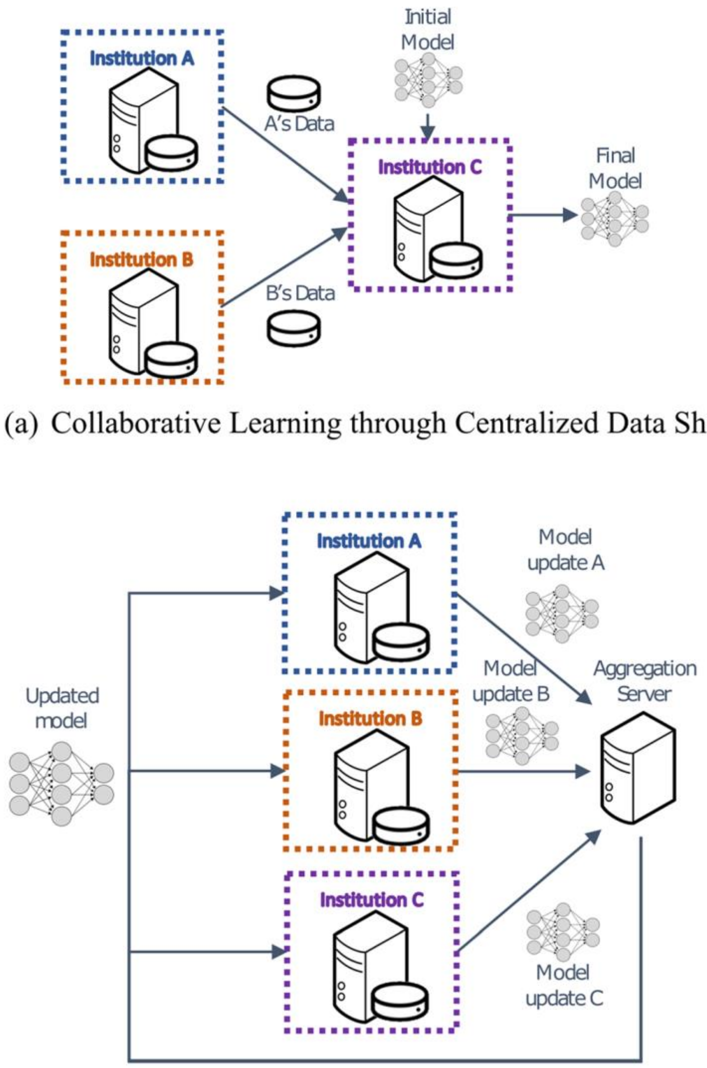
Intel believes that privacy and public health should not be a zero-sum game and we have been working on federated and privacy protective machine learning approaches which are now being used to develop COVID-19 diagnostics through data private multi-institutional collaborations.

Deep learning, a specific AI technique, shows promise in medical diagnosis and treatment but requires large amounts of data for training and effectiveness. A novel use case triggered by COVID-19 involves the analysis of lung scans to improve the diagnosis of pulmonary disease. However, identifying sufficiently large and diverse datasets is a significant challenge and can rarely be found in individual institutions.

The current paradigm for multi-institutional collaborations require sharing patient data to a centralized location for model training (Fig. 1a). We refer to this as *Collaborative Data Sharing (CDS)*. However, CDS does not scale well to large numbers of collaborators, given privacy, technical, and data ownership concerns. Consequently, knowledge remains scattered across institutions, raising a need to seek alternative solutions. Recent collaborative learning approaches enable training models across institutions without sharing patient data. We define such approaches as *Data Private Collaborative Learning (DPCL)*.

Privacy-Preserving Artificial Intelligence (PPAI) techniques can overcome some of the current limitations to share health data across institutions. Moreover, PPAI may address principles such as data minimization, purpose limitation, security safeguards and data quality that have inspired privacy frameworks worldwide.

Federated learning is a novel PPAI paradigm for DPCL where multiple collaborators train a machine learning model simultaneously (i.e., each on their own data, in parallel) and then send model updates to a central server for aggregation into a consensus model (Fig. 1b). The aggregation server then sends the consensus model to all collaborating institutions for use and/or further training.



(a) Collaborative Learning through Centralized Data Sharing

(b) Data-private Collaborative Learning using Federated Learning

Figure 1 – The current paradigm for multi-institutional collaborations, based on Centralized Data Sharing, is shown in (a), whereas (b) denotes Federated Learning.

Data Private Collaborative Learning introduces additional restrictions to the training process over that of data sharing (e.g., not shuffling data across participants) as the computational process is not identical. For any potential collaboration, a crucial question is whether the increased access to data from DPCL improves model accuracy more than these restrictions hamper model accuracy.

Our study addressed brain cancer as an example and performed a quantitative evaluation of DPCL to distinguish healthy from cancerous brain tissue using magnetic resonance imaging scans. We reconstituted the original 10 institutional contributions to the data of the largest manually annotated publicly available medical imaging dataset (i.e., BraTS), to form the *Original Institution* group, such that dataset assignments matched the real-world configuration. We quantitatively compared models trained by (1) single institutions, (2) using the DPCL methods Federated Learning, Cyclic Institutional Incremental Learning, and Institutional Incremental Learning, and (3) using CDS, by evaluating their performance on both data from the *Original Institution* group, and data collected at institutions outside of that group.

These evaluations revealed that *the loss relative to CDS in final model quality for FL is considerably less than the benefits the group's data brings over single institution training.*

Implications for privacy and security

While DPCL methods keep patient records confidential and allow multi-institutional training without sharing patient data, we caution that privacy risks remain. Protecting patient data goes beyond just preventing direct access, i.e., not sharing data.

Data Private Collaborative Learning exposes the AI model to several potential vulnerabilities:

- Machine learning models leak some amount of training data information, enabling model inversion attacks, wherein attackers partially reconstruct training data.
- Model poisoning attacks, where the attackers alter the model weights, enable malicious collaborators to bias models (such as over-recommending certain treatments), install backdoors, and generally degrade a model's performance.
- In data-private collaborative learning, the model parameters are necessarily shared with the collaborators for training. In the case of proprietary models, this exposes the model to theft via local system attacks such as cold boot attacks, where the attackers read data directly from memory.

But proper use of **Trusted Execution Environments** (TEEs) may mitigate FL threats:

- High-end TEEs provide an environment on a computer where computation and memory are hidden from view or influence of even the host operating system, while providing cryptographic assurances of exactly what code is running inside the environment.
- This provides hardware-based assurances between collaborators that they are running the correct code on the correct platforms, and that those platforms mitigate attacks from local adversaries, i.e., the owner of the platform or a rogue actor with physical access.

Conclusion

- We found that Data Private Collaborative Learning approaches, and particularly FL, can achieve the full learning capacity of the data while removing the need to share patient data, thus facilitating large-scale multi-institutional collaborations. This approach is being conducted in Spain to help diagnosing COVID-19.
- Data Private Collaborative Learning removes the need to trust anyone with the data, but to various degrees, requires trusting every participant with the model. In institutional collaborations, these trust relationships may be based on the identities of the institutions, i.e., all institutions know and trust each other. Some TEEs provide cryptographic mechanisms for remotely attesting to the hardware capabilities and the software being run. In such cases, collaborators may base their trust primarily in the TEE.