Risk and Advantages of Federated Learning for Health Care Data Collaboration

Anna Bogdanova[1]; Nii Attoh-Okine, F.ASCE[2]; and Tetsuya Sakurai[3]

Abstract: This paper explores the problem of data collaboration in health care, which is the one of the critical infrastructure sectors designated by the Department of Home Security. Limitations to data sharing in health care obstruct the development of a new generation of medical technology powered by artificial intelligence (AI). Collaborative machine learning helps to overcome these limitations through training models on distributed data sets without data sharing. Among other approaches to collaborative machine learning, federated learning in recent years has demonstrated multiple advantages. However, it had been developed and tested in a highly distributed data environment, which is different from the typical cases of health care data collaboration. The objective of this paper is to validate the known advantages of federated learning and to assess possible risks in a small multiparty setting. The experiments show that federated learning can be successfully applied in a multiparty collaboration setting. However, with a small number of parties, it becomes easier to overfit to each local data so that the averaging steps have to occur more frequently. In addition, for the first time, the risks of a membership inference attack were assessed for different methods of collaborative machine learning. **DOI: 10.1061/AJRUA6.0001078.** © *2020 American Society of Civil Engineers.*

Introduction

The last decade has witnessed a rapid advancement of digital technology due to the data explosion and recent breakthroughs in machine learning. These breakthroughs continue to generate much value across different domains by enabling autonomous technology and getting insights from Big Data. Meanwhile, some fields are experiencing data-related barriers to leveraging new technology. The biggest problem arises from data privacy concerns in the fields related to public infrastructure and health care. In these domains, data cannot be easily shared; thus, gathering it in one place for machine learning applications is prohibitive. At the same time, training machine learning models in isolation has the dangers of introducing bias in critical decision making. Machine learning models trained on a data set absorb representations and relations reflected in that data set and are prone to make costly mistakes when faced with less-represented data of minorities or rare events.

These challenges are especially relevant to the health care field. An increasingly large amount of medical information becomes available in computable form but cannot leave hospital servers without threatening patient confidentiality. Meanwhile, there are strong incentives to combine data from several institutions to advance medical research and train more generalizable predictive algorithms.

A 2010 essay, *Achieving a Nationwide Learning Health System*, proposed a federated learning system among hospitals, government bodies, and research institutions where data are physically shared on-demand among the members. The authors envisioned multiple applications of such collaboration, such as to plan and design clinical trials of a new drug, track the spread of an infectious decease outbreak, monitor the safety of a new drug, or develop clinical decision-support systems (Friedman et al. 2010). However, there are several administrative and logistic barriers to the actual implementation of such a system. To state a few, partners have to implement shared standards of data accounting, establish a system of access clearances, and secure the entire infrastructure for data transfer from cyber attacks.

Recent developments in collaborative machine learning can help to overcome these limitations by allowing one to train machine learning models on distributed data sets without data sharing. This paper offers an overview of existing methods of collaborative machine learning, focusing on federated learning, which is gaining popularity in the field. Then, a discussion introduces of risks and benefits of such a method for the particular case of health care data collaboration, supporting the findings with experiments on a toy data set.

Methods of Collaborative Machine Learning

The idea behind collaborative machine learning is to train machine learning models locally and only share the results of this training, i.e., model parameters. The product of such collaboration is a model equally optimized for disjoint sets of members' data, which can then be used by all members to make predictions on new data.

There are several techniques of aggregating the parameters to produce a unified model developed for specific types of machine learning algorithms. Regardless of the algorithms used, there are three communication schemes that can be established between the collaborators: model averaging, weight transfer, and federated learning, schematically explained in Fig. 1.

In model averaging, individual models are trained on disjoint sets of data, and only aggregated model parameters are shared among the parties. One practical example of such system has

[1]Postdoctoral Researcher, Center for Artificial Intelligence Research, Univ. of Tsukuba, Tsukuba 3050005, Japan (corresponding author). ORCID: https://orcid.org/0000-0001-7468-882X. Email: anna.bogdanova.fw@u.tsukuba.ac.jp

[2]Interim Academic Director, UD Security Initiative, Univ. of Delaware, Newark, DE 19711; Professor, Dept. of Civil and Environmental Engineering, Univ. of Delaware, Newark, DE 19711.

[3]Director, Center for Artificial Intelligence Research, Univ. of Tsukuba, Tsukuba 3050005, Japan; Professor, Faculty of Engineering, Information and Systems, Univ. of Tsukuba, Tsukuba 3050005, Japan.

Note. This manuscript was submitted on July 1, 2019; approved on April 7, 2020; published online on June 23, 2020. Discussion period open until November 23, 2020; separate discussions must be submitted for individual papers. This paper is part of the ***ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering***, © ASCE, ISSN 2376-7642.

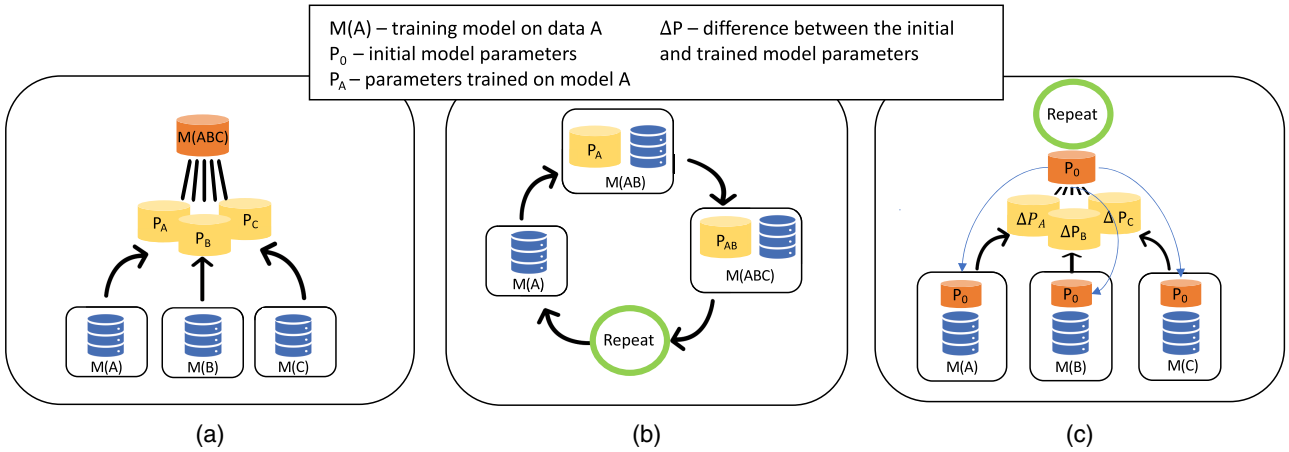

Fig. 1. Methods of collaborative machine learning: (a) naïve averaging; (b) weight transfer; and (c) federated learning.

been deployed by researchers from the MAASTRO clinic in the Netherlands. Their experiment with a distributed machine learning system among several oncology clinics in Europe produced a series of publications along with downloadable prediction models in radiation oncology. Researchers used a simple model averaging approach with a naive Bayes classifier and obtained a model that performed slightly better than the baseline. These experiments were conducted as a proof of concept that the distributed learning approach can be used to extract and employ knowledge from multiple hospitals while being compliant with the various national and European privacy laws (Jochems et al. 2017; Deist et al. 2017).

In the weight transfer approach, collaborators form a chain where they train a model locally for some number of epochs, then pass the parameters (weights) to the next member. For the model to converge in such manner, this process has to repeat for several cycles; thus, it is sometimes referred to in the literature as cyclical weight transfer. One significant advantage of such a method is that collaborators do not need a trusted third party to aggregate the parameters of their trained models. However, the absence of the third party makes it impossible to validate the unified model against unseen data, and the only performance measure available is the average of validation tests among the collaborators, which may not represent the actual model performance. Practical application of cyclical weight transfer had been explored by Chang et al. (2018). Those authors experimented with several methods of collaborative deep learning for the classification of retina images from the Kaggle Diabetic Retinopathy challenge. In particular, they compared ResNet deep learning models trained by four clients in isolation, in a centralized setting, and two collaborative settings: single weight transfer and cyclical weight transfer. They found cyclical weight transfer superior to other methods and observed the improvement of performance with the increase of the frequency of weight transfers.

Federated learning was introduced by researchers at Google in 2016 for learning models on mobile devices in a highly distributed setting (McMahan et al. 2016). In this framework, a central server creates an initial model and sends it to selected clients for training. After some training time, clients call back to the server with deltas of model parameters, which are then aggregated and used for updating the central model. The process repeats many times, each time with a different set of clients. This framework offers multiple advantages for a distributed mobile network, such as communication efficiency and robustness to clients dropping out. The federated learning algorithm was, in particular, designed to overcome the following difficulties: (1) clients have an uneven amount of data; (2) distribution of data values differs with each client (non-IID data, i.e., not independent and identically distributed); (3) the number of clients is large, but participation of every client is not guaranteed; and (4) the communication is limited, so minimization of the number of communication rounds is often an objective.

In addition to the expanding adoption of federated learning for highly distributed use cases, some researchers already considered it for solving the data immobility problem in medical data analysis, although no practical deployment exists at this moment to the best of the authors' knowledge. For instance, Sheller et al. (2018) applied federated deep learning for the open Brain Tumor Segmentation Challenge (BRaTS). They tested the performance of a federated algorithm against other methods of collaborative learning and in various experimental configurations including non-partitioned (centralized) data and a different number of data partitions (from 4 to 32) as well as the real (unbalanced) distribution of images among contributors to BRaTS data set. Their results demonstrated the superiority of federated learning (FL) over other methods over other methods in all distributed experimental settings and competitive performance with the centralized model. Brisimi et al. (2018) deployed federated learning to predict hospitalizations for cardiac events from electronic health records. They designed an original algorithm for federated optimization of support vector machine (SVM) classification model and tested it on different graph topologies for 5 and 10 data partitions. They experimentally demonstrated faster convergence and smaller communication overhead compared with alternative methods while maintaining lossless performance compared with a centralized model.

Federated learning, adjusted for to a multiparty setting, could be a feasible solution for the data immobility problem in health care. However, given the sensitive nature of medical data, the framework has to be thoroughly tested, and data privacy guarantees provided before it can be adopted for real health care applications.

Federated Learning Performance in a Multiparty Setting

Unlike the typical federated learning setting where one service provider connects to the loose federation of clients, in a multiparty setting, several entities accumulate data of their clients and seek to enhance their data analysis through collaboration. Thus, the following disparities from the fully distributed setting are expected: (1) data are much closer to being IID-distributed; (2) the number

of participants is limited, but the participation of each member can be guaranteed; (3) there are no strict constraints on communication resource; and (4) with just a few participants, there is a bigger danger of overfitting of the global model to local data sets. Hence, some of the known advantages of federated learning might not be applicable for the multiparty setting, and other alternatives should be carefully considered.

In order to identify the advantages of federated learning for health care data collaboration, first, a comparison of different collaborative machine learning schemes is conducted, including the alternatives to collaborative learning such as training models individually or centralizing the data for model training. Then, an exploration is made of functionality that is only provided by the federated learning, and the experiments of McMahan et al. (2016) are replicated to find out conditions in which the distributed model converges to the global minimum for a small number of collaborating parties, and in which it can reach the level of the nonfederated performance.

Experimental Setting

All of the experiments were conducted on the CIFAR10 data set (Krizhevsky 2009). It consists of 60,000 small-resolution color images in 10 classes: airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks; it is often used for benchmarking image classification algorithms.

A share of this data set consisting of 15,000 images was evenly and randomly split among three users for training in a federated framework. The validation set consisted of holdout data of 10,000 images.

The same neural network architecture was used for all experiments: two convolutional 5×5 layers with 32 and 64 channels, each followed by a 2×2 MaxPool layer, then a fully connected layer with 512 units, and a softmax output layer. This particular convolutional neural network (CNN) architecture was chosen because it was used in the experiments of the original federated averaging algorithm (McMahan et al. 2016), and because it offers good performance on the CIFAR10 data set with a minimum amount of trainable parameters.

Alternatives to Federated Learning

In order to be able to weight the benefits of federated learning deployment against its costs and possible risks, it is important to assess the viable alternatives. Apart from other methods of collaborative learning described in the previous section, choices can be made to centralize the data despite the privacy constrains, or to refuse collaboration and train an individual model instead.

Fig. 2 shows the convergence rates and validation accuracy of alternative methods of collaborative machine learning. A centralized model was trained on 15,000 samples for 36 epochs with batch size of 32. It fully converged to the training data (reached zero loss), and peaked at 63% performance on the test data around Epoch 5, after which the validation performance was decreasing. This signals overfitting to the training data, which can be treated by one of a many available techniques of regularization, for example by decreasing the number of trainable parameters or adding dropout layers to model architecture. A regularized version of the centralized learning is shown by a dotted line, where each layer of trainable parameters was followed by a dropout layer randomly blocking of 25% of nodes.

Similarly, unregularized and regularized versions of the CNN model were trained by the three clients separately, on 5,000 samples each and validated on the centralized set of 10,000 samples of test data. The average performance of the three models is shown. One can see that the rate of overfitting is much bigger even for the regularized model, which often happens for the small training samples. When the parameters of separately trained models were averaged, 61% validation accuracy was obtained, and a 0.44 loss on trained data, the worst among all methods.

The weight transfer was implemented in a cyclical manner, with each client training the model for two epochs and then sending the trained parameters to the next client. The next client fine-tunes these parameters to their data for the next two epochs, and this process continues for the total of six full cycles. The results show that this model converges much slower than the centralized version while not reaching its validation accuracy benchmark. It is also clear that the training process is rather unstable, with training loss jumping each time the model parameters are transferred to the next client. This effect can be even more pronounced in cases where the data distribution between clients is not even, as seen by Sheller et al. (2018).

This study's implementation of federated learning followed the FedAvg algorithm (McMahan et al. 2016). First, the initial model parameters are randomly initialized and sent to each client. Then, each client in parallel trains the model for one epoch with a batch size of 32. After that, the differences between the initial and resulting model parameters from each training are averaged and applied to update the initial model. This cycle is repeated for 36 epochs to match the centralized model training. The results showed gradual and steady convergence of the federated learning training, similar to the regularized version of the centralized model. The federated

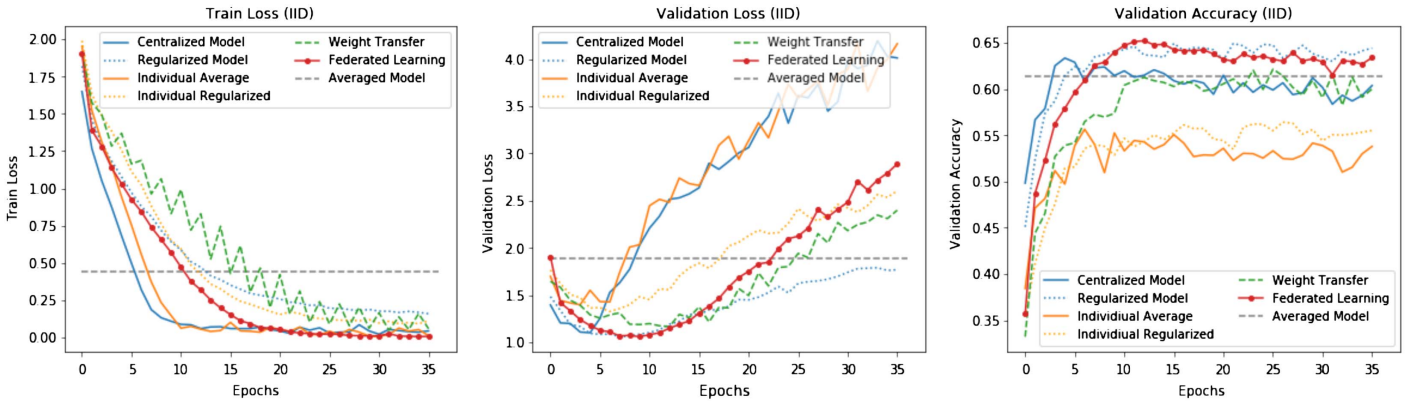

Fig. 2. Alternative methods of collaborative machine learning.

ASCE-ASME J. Risk Uncertainty Eng. Syst., Part A: Civ. Eng., 2020, 6(3): 04020031

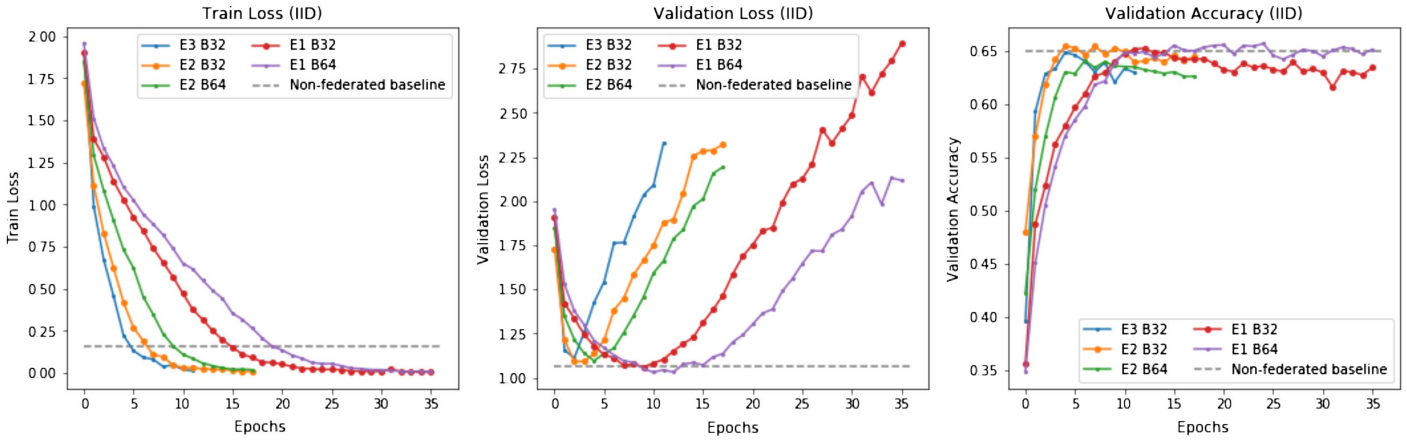

Fig. 3. Federated learning parameters in a multiparty setting.

learning approach in this study's experiments achieved 65% validation accuracy, reaching the baseline of the centralized and regularized model. Thus, the performance of federated learning in this experiment may be considered lossless, i.e., there is no drop in performance compared with the centralized model.

However, depending on the specifics of the data, its distribution among the clients, and complexity of the model, the results of such analysis might be different, and other methods can appear more beneficial than federated learning. The goal for this experiment was to show that federated learning can achieve results comparable to the centralized setting, and thus can be considered even outside the traditional highly distributed federated setting.

Federated Learning Parameters in a Multiparty Setting

One important advantage of the federated learning approach is its communication efficiency (McMahan et al. 2016). Because the clients are training their models in parallel, the number of training epochs and batch sizes can be tuned so that less communication between the clients and the server is necessary for the global model convergence.

Fig. 3 shows the results of federated training with different amounts of computation per client, in comparison with the baseline model performance (best metrics achieved by the regularized nonfederated model). The amount of computation was controlled by the parameter B, indicating a batch size (number of data records to process before the model parameters get updated locally), and parameter E, indicating the number of epochs (complete passes through all data) each user makes before the communication round. A bigger batch size indicates a smaller amount of gradient descent steps that each client makes before the communication round.

In general, it was observed that increasing the amount of computation per client causes faster convergence on the training data, thus leading to overfitting and lower results on validation data. This is contrary to the behavior of fully distributed data found by McMahan et al. (2016), where the training runs with more local updates converged to higher values of validation accuracy. At the same time, decreasing the amount of computation per round, taking a batch size of 64 instead of 32 with one epoch step (E1 B32) had a notable regularization effect, and this run achieved higher validation accuracy.

Data Privacy Concerns in Collaborative Machine Learning

One necessary caution to acknowledge is that even without data sharing, there is a possibility of unintended information leakage through the shared model. For instance, it has been shown that given only the outputs of a classifier model, is it possible to reconstruct model parameters and generate an image representative of each class (Hitaj et al. 2017). This is known as the model inversion attack, and as long as classes contain data of more than one individual, it does not pose severe concerns from a data privacy perspective.

Another type of attack is a property inference attack, where the adversary can infer some additional properties of the training data unrelated to the intent of the classifier, for example, if there were more men or women in a training set for predicting hospital readmissions. Melis et al. (2019) demonstrated that this type of attack poses a valid concern for collaborative learning because the properties can be easily inferred from the model parameters that are exchanged between the collaborators. The present study does not focus on this type of attack because it is directed against the data holder rather than an individual represented in data. Moreover, federated learning offers a defense against the attacks on shared model parameters in form of secure aggregation (Bonawitz et al. 2016). With secure aggregation, individual parameters cannot be accessed, so that the server or eavesdropper can only view average parameter values across all clients.

Finally, there is a membership inference attack that poses serious privacy concerns for individuals (Shokri et al. 2017). In this attack, an adversary given only the outputs of a model and some data record can establish with a high degree of certainty whether this data record was used to train the model. Previous research demonstrated that federated learning is susceptible to insider membership inference attacks, and in the case of the small number of clients, conventional protection mechanisms such as differential privacy cannot be applied (Truex et al. 2019). In the original federated learning framework, membership inference is not a concern because the knowledge of individual membership in a mobile network has little information value. In a medical field, however, merely knowing if someone's data appeared in the model training for predicting the progression of some disease is a serious privacy breach.

This paper explores the relationship between different methods of collaborative learning and different federated learning parameters on the success of the membership inference attack. The authors argue that a thorough analysis of the risks of membership inference attack should be a major part of the decision-making process in health care data collaboration.

Membership Inference Attack on Federated Learning

To stage the membership inference attack on collaborative learning, this study followed the framework proposed by Shokri et al. (2017).

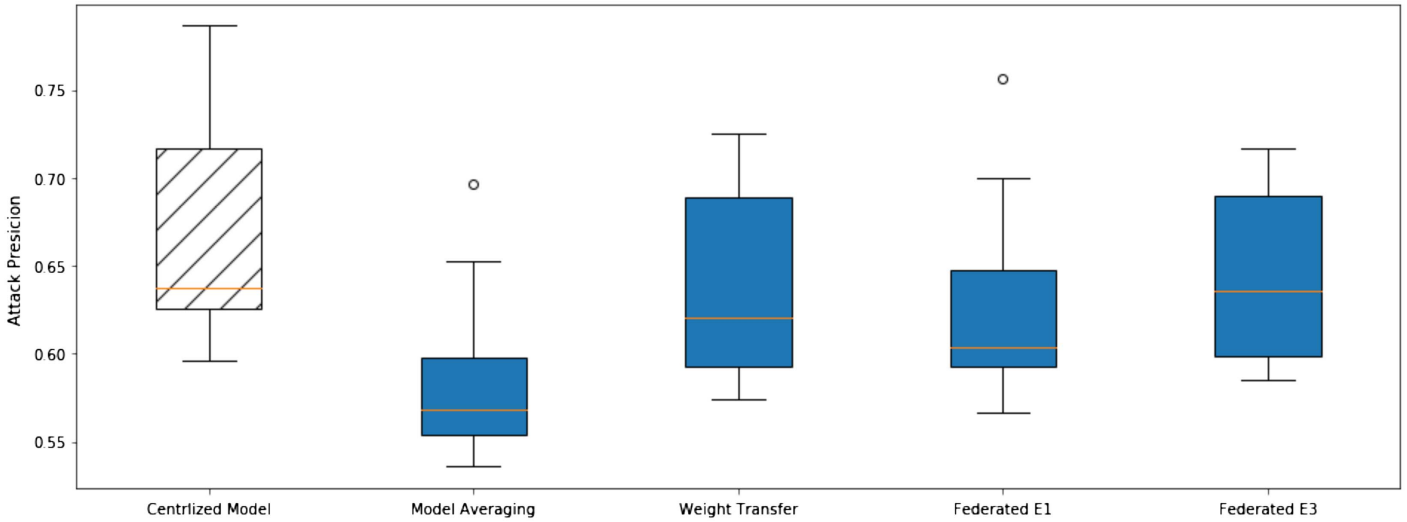

Fig. 4. Precision of a membership inference attack on collaborative machine learning.

The goal of this attack is to infer if particular data entry was used for training the target model. This type of attack uses target model outputs, which are represented by a vector of probabilities that a particular data entry belongs to each class.

The first stage of the attack is to assemble several shadow training sets with data that resemble the original training data but comes from a disjoint set. In our case, the validation set of the 10,000 images of CIFAR10 was used. The shadow training set is then divided into two parts: shadow training and shadow test. The shadow training set is then used to run several imitations of the target model and collect the prediction vectors that they produced. As a result, the adversary obtains a set of prediction vectors for the data used in training and pairs it with prediction vectors from the shadow test, not used in training. A binary classifier is then trained to learn the distinction between the two sorts of prediction vectors. Such a distinction is possible because of the different level of confidence that the model reveals through prediction vectors when it is queried with the data that it has seen versus the data it has not seen in training. Finally, having the attack model, the adversary can get the output from the target model on any data entry and classify this output as being in or out of the target training set.

Previous research showed that several factors influence the success of this kind of attack: overfitting of the target model, the number of classes, and the similarity of data within one class (Truex et al. 2019). When the models are overfitting, their output classifies the inputs from the training set with much bigger confidence than it classifies the new entries, and the attack model can learn from this difference. Similarly, more classification labels mean more features in the attackers' set; consequently, the attack model becomes more sensitive to slight differences.

Fig. 4 presents the results of membership inference attacks on the centralized baseline model and the different methods of collaborative learning. Following the original method, the success of the attack is evaluated with the precision metrics, and the attack is conducted separately for each of the 10 class labels.

As previously observed in the literature, attack precision varies greatly according to the class of the data capturing the amount of data-specific variation within each class. The present study also found that all of the collaborative learning methods had lower attack precision than the centralized model, with the model averaging method the least susceptible to the attack. This effect can be explained by bigger disparities between the imitation model trained by an attacker and the actual target model trained in a collaborative manner. Surprisingly, federated learning runs with larger amount of computation per client, which were shown to be overfitting to the training data, did not expose more susceptibility to the attack.

Although on average, the precision of the membership inference was not exceeding 65% in all cases, the occasional high precision of 80% for certain data classes can be a concern.

Conclusion

This work reviewed several variants of collaborative machine learning for health care data collaboration. Among other methods, federated learning is a fast-developing technology pioneered by Google, which has demonstrated promising results in previous research. However, federated learning was developed for a highly distributed environment, where there are thousands of clients, each holding a single sample of data. In contrast, data collaboration in the medical field involves combining large sets of precollected patient data. Therefore, there is a need to experimentally test the performance of federated learning in the environment more similar to cases of medical data collaboration.

This study's experiments showed that federated learning can be successfully applied in a multiparty collaboration setting. It not only reached the baseline of a centralized model but also had a regularization effect on the training process. However, with a small number of parties, it becomes easier to overfit to each local data so that the averaging steps have to occur more frequently. In the conducted experiments, configurations where the global model was averaged at each local epoch of the training performed best. Therefore, it was concluded that one of the main advantages of federated learning, namely achieving communication-efficient training, cannot be easily gained in multiparty settings.

Finally, to explore the privacy guarantees of the federated learning, the precision of membership inference attacks against the centralized baseline model were compared with different methods of collaborative machine learning. It was observed that all of the collaborative learning methods had lower attack precision than the centralized model due to the regularization effect of model averaging and the difficulty to imitate a collaboratively trained model.

Combining experimental results with the literature research, the following advantages of federated learning for health care data

collaboration have been identified: (1) it is possible to achieve lossless model convergence rates and validation accuracy; (2) it is robust to clients dropping out or highly imbalanced data shares; and (3) it offers better privacy guarantees through secure aggregation of model parameters. At the same time, it was found that not all known advantages of federated learning can be gained in the small multiparty setting. For instance, attempts to make communication rounds less frequent can lead to overfitting to local data and should be avoided.

Perhaps the biggest risk of federated learning for health care data collaboration is the susceptibility to a membership inference attack. This attack can be attempted by a central server, an eavesdropper, or an insider, and it explores the difference in model responses to known and unknown data samples. Because there is currently no defense against such attack, the risks should be assessed according to the properties of the training data. Higher risks of the attack are associated with a bigger number of classification labels and the higher in-class variation. In addition, the damage from a potential membership inference should be properly weighted.

Data Availability Statement

All data, models, and code generated or used during the study as well as experiment instances are available online at https://github.com/abogdanova/FL-MIA.

Acknowledgments

This study is funded by NEDO (New Energy and Industrial Technology Development Organization), the funding agency of the Japan Ministry of Economy, Trade and Industry (METI) for US–Japan Collaborative Research and Development of Next Generation AI Technology.

References

Bonawitz, K., V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth. 2016. "Practical secure aggregation for federated learning on user-held data." In *Proc., NIPS Workshop on Private Multi-Party Machine Learning*. Ithaca, NY: Cornell Univ.

Brisimi, T. S., R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi. 2018. "Federated learning of predictive models from federated electronic health records." *Int. J. Med. Inf.* 112 (Apr): 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007.

Chang, K., N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D. L. Rubin, and J. Kalpathy-Cramer. 2018. "Distributed deep learning networks among institutions for medical imaging." *J. Am. Med. Inf. Assoc.* 25 (8): 945–954. https://doi.org/10.1093/jamia/ocy017.

Deist, T. M., et al. 2017. "Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: Eurocat." *Clin. Transl. Radiat. Oncol.* 4 (Jun): 24–31. https://doi.org/10.1016/j.ctro.2016.12.004.

Friedman, C. P., A. K. Wong, and D. Blumenthal. 2010. "Achieving a nationwide learning health system." *Sci. Transl. Med.* 2 (57): 29–57. https://doi.org/10.1126/scitranslmed.3001456.

Hitaj, B., G. Ateniese, and F. Pérez-Cruz. 2017. "Deep models under the GAN: Information leakage from collaborative deep learning." In *Proc., 2017 ACM SIGSAC Conf. on Computer and Communications Security*, 603–618. New York: Association for Computing Machinery.

Jochems, A., et al. 2017. "Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries." *Int. J. Radiat. Oncol. Biol. Phys.* 99 (2): 344–352. https://doi.org/10.1016/j.ijrobp.2017.04.021.

Krizhevsky, A. 2009. "Learning multiple layers of features from tiny images." M.S. thesis, Dept. of Computer Science, Univ. of Toronto.

McMahan, H., E. Moore, D. Ramage, and S. Hampson. 2016. "Communication-efficient learning of deep networks from decentralized data." Preprint, submitted February 17, 2016. http://arxiv.org/abs/1602.05629.

Melis, L., C. Song, E. De Cristofaro, and V. Shmatikov. 2019. "Exploiting unintended feature leakage in collaborative learning." In *Proc., 2019 IEEE Symp. on Security and Privacy*, 691–706. New York: IEEE.

Sheller, M. J., G. A. Reina, B. Edwards, J. Martin, and S. Bakas. 2018. "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation." In *Proc., Int. MICCAI Brainlesion Workshop*, 92–104. Berlin: Springer.

Shokri, R., M. Stronati, C. Song, and V. Shmatikov. 2017. "Membership inference attacks against machine learning models." In *Proc., 2017 IEEE Symp. on Security and Privacy*, 3–18. New York: IEEE.

Truex, S., L. Liu, M. E. Gursoy, L. Yu, and W. Wei. 2019. "Demystifying membership inference attacks in machine learning as a service." In *Proc., 2017 IEEE Transactions on Services Computing*. New York: IEEE.