JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A REVIEW: FEDERATED LEARNING FOR HEALTH CARE APPLICATION

Jyoti Bangare[1],Nilofer kittad[2]

Assistant Professor,Computer Department,

MKSSS's,Cummins College of Enginerring for Women, Pune, India.

Abstract

Data-driven machine learning (ML) has become a potential strategy for generating accurate, robust statistics models from medical data, gathered in huge amounts by modern healthcare systems, with the rapid development of computer software and hardware technologies. A large amount of healthcare data from clinicians, patients, insurance companies, the pharmaceutical industry, and others are openly accessible. This abundance of data and sophisticated access to it offered data science technology an unparalleled chance to obtain data-driven insights and improve the quality of healthcare provision.

Existing medical data are not properly used by ML models, mainly because it is contained in data silos. However, ML will not be able to achieve its full potential and ultimately migrate from research to medical care systems without access to complete data. This study reviews federated learning (FL), especially in the health care systems. We summarise in particular the broad solutions to federated learning's statistical challenges, system impediments, and privacy issues and identify the consequences and potentials for health care systems.

Keywords: Machine Learning, Healthcare, Federated Learning, Privacy

I. Introduction:

The interest in health care data analytics has been on the rise in recent years, since such data are increasing from numerous sources, among others, including clinical institutions, patients, insurance companies, and the pharmaceutical industry. In this way, computing technologies for data-driven insights to improve the quality of the care providers are developed without any previous occurrence. Health care data are frequently fragmented since the health care systems and processes are complicated. For example, the clinical records of their patients may only be accessible to separate hospitals [1,2].

Personalized medical techniques, strive to customize treatment to individual patients and profit from a method to cluster individuals and to make more educated assumptions about the needs of patients. For diverse study areas, ideally, sensitive existing databases can therefore be utilized without violations of privacy. One approach to achieve this is by pseudonymizing or de-identifying data, which is substituted by a pseudonym to safeguard personal privacy for specific identifiable data sections such as name, address, or social security number. There have been cases in which the pseudonymized data can be traced to individuals [3,4], this method is not safe. Federated Learning (FL) was presented, which arose out of these challenges and is used to divide the computing load of the training of the ML model.

The term FL was initially used by [5] and represents a strategy to train an ML model without the usage of private data by others, both dispersed and private. The FL uses shared model parameters, which can be aggregated to a common model, rather than share the data directly with non-reliant parties. The FL system uses a client-server model with one server to facilitate the training, create the model, and make it available to all customers that train the model on their datasets.

The paper is organized as follows. In the introduction section, we sketch the overview of FL which includes characteristics as well as describes why the FL is the right solution for health care systems. Section 2, summarizes the literature review. The scope of FL is illustrated in Section 3. Section 4, we point out three challenges in FL along with relative improvement. Furthermore, we conclude indirect information leakage in FL and the existing privacy-preserving method employed in FL. At last, concluded with some promising direction of FL to give guidance for future work.

1.1. Outline of Federated Learning:

Federated learning is a recent popular paradigm since it is highly promising to learn with fragmented, sensitive data. The system enables users to form a shared global model with a central server, rather than aggregate data from multiple locations together, or rely on traditional discovery and replication architecture while retaining the data in local institutions where the data originated.

Health data are frequently fragmented since the healthcare system and operations are sophisticated. For instance, various hospitals can only access their patient populations' clinical records. These records are extremely sensitive to individuals' protected health information (PHI). Strenuous rules have been devised to regulate the process of access and analysis of such data, such as the Health Insurance Portability and Accountability Act (HIPAA) [6]. It provides a great challenge for modern technology such as deep learning [7] in the field of data mining and machine learning. This generates a wide range of training data.

Multiple organizations or institutions work together, under the coordination of a central server or service provider, to solve a machine-learning problem. Therefore, a deep learning model within a central server is maintained and updated. The approach is developed by being distributed to faraway data centers such as hospitals or other medical establishments, enabling these places to maintain their data located. Never exchange or transfer data from each collaborator. The central server retains a globally shared model that is spread throughout all institutions instead of bringing data to a central server, as in conventional profound learning. Each organization then maintains a distinct model based on data from its patients. Each center will then supply the server with feedback, either through weight or the model's error gradient, based on the separately trained models. The central server adds feedback from all participants and updates the global model based on established criteria. The predetermined criteria permit the model to assess the feedback quality and consequently to include only value-added elements. Feedback can therefore be ignored from centers with unfavorable or odd findings. This approach is a federated learning process that is iterated until the global model is developed.

The federate learning process is illustrated in Figure 1 and the learning framework methods are summarised in Figure2. Federated learning enables individual hospitals without centralizing the data in one place to gain from rich data set from several non-affiliated hospitals. This approach solves key problems such as privacy, data security, rights of data access, and access to heterogeneous data. Federated learning hence enables several collaborators to develop a robust machine learning model through a huge dataset [8].

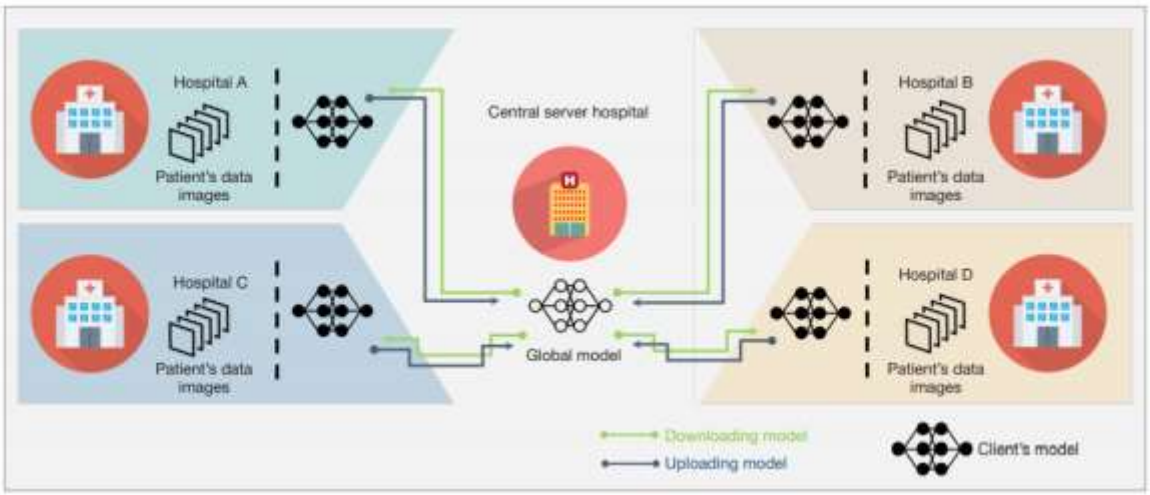

Figure 1: An overview of federated learning

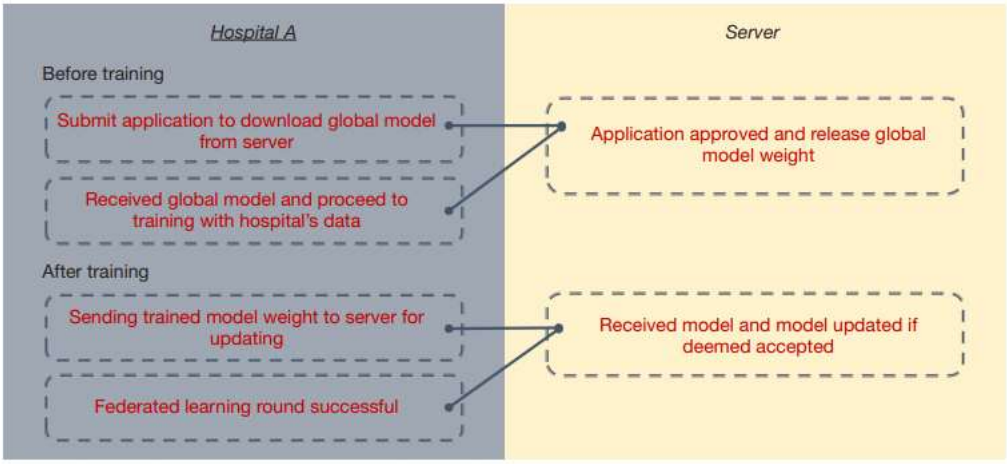

Figure 2: Summary of learning framework procedure

Federated learning is very promising in data analytics in healthcare. The sensitive patient data may be available in local institutions and with individual consumers without going out during the federated process for the multiple providers (e.g., establishing a model to predict the risk of hospital readmission with patients Electronic Health Records (EHR) [9]) and consumer applications [10]. The objective of this paper is to evaluate the establishment of federated learning, explore the overall solutions and obstacles, and examine its healthcare applications.

1.2. How federated learning for healthcare AI is the perfect option

The solution to healthcare AI is federated learning. It allows data from various data silos to connect without any patient data migration. No big transferring and saving files. Above all, there is no risk to the privacy of patients [11].

Three things are obvious while conversing with AI innovators who address some of the most complicated healthcare challenges:

1. Developing improved solutions which can increase results for the enormous heterogeneous community of patients today is slowed down by poor access to actual data in the world;
2. The issues connected with the management of different data sets across many sites and systems are hindered by validating new AI-based health care solutions; and
3. Maintenance of models during the life of an AI solution is not yet flawless and requires not only continuous access to vast amounts of data but also to regulatory compliance infrastructure and tools as a model evolves.

The problems of data sharing and patient privacy are central to these dilemmas. It is clearly of the utmost significance to protect Personally Identifiable Information (PII). This does not mean, however, that the development of AI needs to be slow. The future of healthcare necessitates advanced solutions to be developed that encompass data privacy.

The field of health data aims to have a wealth of on-site and cloud data silos. Some of these data silos come from a virtual private cloud of hospitals with a shared data lake and some from one hospital on-prem. There will be numerous separate silos of health data all around the world, though.

To make data relevant in these silos, a data federation is required. We must be able to use data from these disparate data silos to design sophisticated healthcare solutions, which operate effectively with various patient populations. This needs to be done in a way that safeguards the privacy of patient data. For this reason, FL is the appropriate choice for AI in medicine.

2. Related Work

FL offers significant prospects in health care as a disturbing means of preserving data privacy. There could be many patient data in each medical organization, but that could be far enough to train their prediction models [12]. Combining FL and illness prediction is one of the best methods in different hospitals to break down the obstacles to analysis.

In paper [13] the Electronic health records (EMR) stated that they include many significant clinical ideas and tried to utilize tensor factorization models to analyze phenotypes to extract information contained in medical records without sharing patient-level information. The first attempt to apply FL in the medical field might be considered. In the paper [14], private learning for EMR in a federated context was examined separately. Moreover, the performance of training in a centralizing environment was found to be comparable. Moreover, the paper [15] also uses EMRs to assess if an FL method termed a cluster primal-dual splitting (cPDS) hospitalizes a cardiac illness patient. Either health monitoring systems or hospitals keeping these health data without leaking information can execute this prediction task.

The authors [16] proposed a federated patient hashing architecture using health records to recognize identical patients in multiple hospitals without the sharing of patient data. This patient matching strategy can assist doctors in synthesise the general character and lead them to treat more experienced patients. Federate averaging algorithms on drug use taken from the MIMIC-III Data Base to estimate patient death rate are also levied in the paper [17], Loss-based adaptive boosting. The research involved computing complexity and cost of transmission as well as precision, which outperforms the baselines for every customer. The paper [18] emphasizes the need to handle clinical notes using non-structured data. This was the first FL-based NLP attempt. They carried out a two-stage federal training model that included a pre-processing stage for each patient to forecast a model of representation and a phenotyping stage to analyze each type of disease.

In addition to the above domains, it could have very broad possibilities for popularisation and application in fields sensitive to data, as best we know from the increase and maturation of FL. The comparison of various review documents is shown in Table 1. It is therefore confident that in the future, FL has significant potential. Currently, FL contributes mostly to horizontally collaborating landing application training, meaning that each data's feature dimensions are similar. Medical data could be used at hospitals in the future to obtain reasonable pricing with additional institutions, such as insurance agents. Hence, FL is a viable way to be examined vertically. In addition, existing federal training mostly is based on small organizations and cannot cover large numbers of devices or institutions with joint training.

Table 1: Comparison of review papers [19]

Authors & Year	Application domain	Descriptions	Advantages	Limitations
Kim et al., 2017 [13]	Analysis of computational phenotypes	Federated privacy-preserving tensor factorization for computational phenotyping	Summarized information does not reveal the patient data	Accurate with small or skewed data only
Pfohl et al., 2019 [14]	Clinical prediction	Set the effectiveness of FL through centralized and local learning	Perform FL in a distinctly private way	Data protection costs are underestimated
Brisimi et al., 2018 [15]	Predict future patient hospitalizations	Cluster Primal-Dual Splitting algorithm	Yield classifiers using relatively few features	Need more iterations for convergence
Lee et al., 2018 [16]	Similar patient matching	Federated patient hashing framework	Avoid reverse engineering security attacks	Inevitable computational complexity
Huang Yin et al., 2019 [17]	Mortality prediction over drug utilization data	Adaptive boosting method	Alleviate non-iid by introducing data-sharing technology	Training on iid data outperform non-iid data
Liu et al., 2019 [18]	Extraction of clinical notes	Two-stage federated NLP method	Adding a pre-processing step to improve accuracy	Not suited for tiny questionable instances

3. Benefits of Federated Learning:

If a medical organization seeks to maximize the usage of a specific department or unit such as a specialist health center, the technology may be used for the use of clinical data as a working machine. Federated learning may also be employed by physicians from a transactional standpoint when interacting with electronic record systems.

Regardless of their application, federated learning systems provide healthcare institutions with various benefits, data possessed including:

- **Privacy and Compliance:** Federated learning complies with privacy guarantees based on approval requirements and legislation such as HIPAA (Health Insurance Portability and Accountability Act) and the EU's (European Unions) General Data Protection Regulation, by maintaining all data in networks of health care organizations at all times. Alternative techniques to data security should be taken into consideration in prior attempts to protect healthcare data utilizing de-identification. Federated learning can help resolve the renowned de-identification conundrum of not sufficiently de-identified data to re-identify patients and too de-identified AI algorithm training.
- **Security and accessibility:** Federated learning makes the transfer of high-risk and slow physical data an important thing of the past by keeping all data contained in the health care networks. Once they have finished working with it, it's also necessary to trust data users to appropriately return or dispose of data. Those permitted to access data can access them safely and quickly using federated learning systems regardless of where they are.
- **Organization and integration:** There's a never-ending range of health data to confront between X-ray and MRI images, blood pressure readings, and physician handwritten notes. Federated study can aid by smoothly integrating and formatting all data into existing health databases or EHRs, thus making it easy and searchable.
- **De-risked correlation:** For clinical research, the capacity to correlate disparate data sets is crucial, as local data sets are often too limited or biased. This, however, leads to a difficulty of connecting and eventually de-anonymizing data sets. Although the interaction between privacy and utility is likely to always be established, federated learning can help to achieve secure correlation by enabling varied privacy, allowing the sharing of data by characterizing group patterns in the data sets while retaining information about individuals.

4. Federated Learning Challenges:

Although FL has advantages, it does not overcome all problems inherent in medical data learning. A successful model training depends still on aspects such as data quality, prejudice, and standardization [20]. These problems must also, through appropriate means such as cautious learner conception, common information collecting protocols, structured reporting, and advanced methodology for the discovery of bias and concealed stratification, be resolved for both federated and non-federated learning endeavors. In the following, we touch on crucial characteristics of FL which have to be taken into account in constructing FL [21, 22, 23] when it is applied to digital health.

1. Data heterogeneity: Medical data is very diversified, not only because of the variety of modalities, size, and features in general but also because of factors like acquisition disparities, medically supported branding, and local demographics in the specific protocol. FL can assist tackle some kinds of bias through the possible diversity of the data sources, but the uniform distribution of data offers challenges for FL algorithms and methods, with many assuming that data are independently and identically dispersed (IID) throughout the participants. Recent results show, however, that training in FL remains doable [24], but medical information is not delivered uniformly among institutions [25, 26] or has local orientations [27]. Research to tackle these challenges include, for example, FedProx [28], the Domain Adaptation with FL, and Part Data Sharing Strategy [29]. One further problem is that the variety of data can lead to a situation where an individual local participant does not have an ideal global solution.

2. Confidentiality and security. Medical data are extremely sensitive and must, in line with acceptable confidentiality rules, be protected accordingly. Consequently, trade-offs, methods and remaining concerns related to FL's privacy potential are some of the essential factors.

    - Privacy vs performance: FL does not resolve any possible privacy problems and, in general, will always entail some risks similarly to ML algorithms. Protection levels that are above today's commercially available ML models can be found in FL privacy-preserving approaches [30]. There is, however, a performance trade-off and these strategies can influence, for instance: the correctness of the last model [31].

    - Confidence level: Participating parties can often enter two kinds of FL collaboration: (1) Trusted: many of the more harmful reasons, such as deliberate attempts to obtain sensitive information or to deliberately damage the model, maybe eliminated for FL's consortia in which all parties are regarded to be trustable and bound by an enforceable collaborative agreement. (2) Untrusted: it may be impractical to establish an enforceable collaboration arrangement in FL systems operating on bigger sizes. Certain customers may attempt to impair performance purposefully, reduce the system, or extract information from others. Security strategies to mitigate such risks will therefore be required, such as advanced model submission encryption, secure authentication of all parties, action traceability, differential

confidentiality, verification systems, the integrity of execution and model confidentiality, and safeguards from adverse events.

- Informational leakage: FL systems are developed to prevent data sharing between entities involved in healthcare. The shared information, however, can still indirectly disclose the private data used for local training, e.g. model reversal of the model updates, the gradients, or adverse attacks [32,33]. If the training process is exposed to many parties, FL differs from traditional training and thus increases the risk of leaking from reverse engineering, if opponents can observe model changes over time, observe particular model updates or manipulate the model. It may take and remains an ongoing topic of research to develop countermeasures such as reducing the granularity of the updates and adding noise and providing acceptable differential privacy.

3. **Traceability and accountability:** The reproducibility of the system is essential for FL in health care, as in much safety-critical application. FL requires multi-part computers in situations which include significant variability concerning software, networking, and hardware, as opposed to centralized training. The traceability of all system assets throughout the training procedures, including the history of data access, training settings, and tweaking hyperparameters is thus required. Traceability and accountability mechanisms, particularly in non-trusted federations, require comprehensive implementation [34]. Once the course process has met the optimal criteria mutually agreed, the contribution of each participant can also be measured, for example, the usage of computational resources, the quality of information utilized to provide local training etc. Through these measures, relevant remuneration could be identified and a revenue model for participants established. One of the consequences of FL is that researchers cannot study data that train models for unforeseen effects.

5. Conclusion:

In the field of digital health care ML, and in particular DL, has led to a broad spectrum of advancements. Since all ML methods strongly benefit from the capacity to access data that matches genuine worldwide distribution, FL is a potential way to achieve strong, accurate, secure, robust, and unbiased models. By letting several parties work together without exchanging or centralizing data sets, FL tackles challenges relating to the extent to which sensitive data are collected. This study helps to finish health systems application and summarises the FL review, however,, it is confined to applications. This is the first time that we have summarised FL's development outlook in the health system to our best knowledge. We have concluded literary masses that FL remains the correct choice and challenge for the health care AI. We also provide the major path to optimization to clarify several solutions, including data protection considerations, that researchers have taken to optimize FL. During the following decade, FL is surely an active field of research. However, we believe it can have a highly positive impact on health care systems and ultimately improvement in medical treatment. Researchers profit from this study to tackle FL's remaining problems.

References:

[1]. Miotto R, Wang F, Wang S, Jiang X, Dudley JT (2018) Deep learning for healthcare: review, opportunities and challenges. Brief Bioinformatics 19(6):1236–1246

[2]. Wang F, Preininger A (2019) Ai in health: state of the art, challenges, and future directions. Yearb Med Inform 28(01):016–026

[3]. Chris Culnane, Benjamin I. P. Rubinstein, and Vanessa Teague. 2017. Health data in an open world. CoRR abs/1712.05627 (2017). arXiv:1712.05627 http://arxiv.org/abs/ 1712.05627.

[4]. Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. Nat. Commun. 10, 1 (2019), 3069

[5]. H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. 2016. Federated learning of deep networks using model averaging. CoRR abs/1602.05629 (2016). arXiv:1602.05629 http://arxiv.org/abs/ 1602.05629.

[6]. Gostin LO (2001) National health information privacy: regulations under the health insurance portability and accountability act. JAMA 285(23):3015–3021

[7]. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

[8[. Ng D, Lan X, Yao MM, Chan WP and Feng M. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. Quant Imaging Med Surg, vol. 11, no. 2, pp. 852-857, 2021.

[9]. Min X, Yu B, Wang F (2019) Predictive modelling of the hospital readmission risk from patients' claims data using machine learning: A case study on COPD. Sci Rep 9(1):2362

[10]. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Balasubramanian V, Russo AM, Rajmane A, Cheung L, et al. (2019) Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med 381(20):1909–1917

[11]. https://medcitynews.com/2021/05/why-federated-learning-is-the-right-solution-for-healthcare-ai/

[12] Szegedi, G., Kiss, P., & Horv´ ath, T. (2019). Evolutionary federated learning on EEG-data. In ITAT 2019-Information technologies – Applications and Theory (pp. 71–78).

[13] Kim, Y., Sun, J., Yu, H., & Jiang, X. (2017). Federated tensor factorization for computational phenotyping. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 887–895). https://doi.org/ 10.1145/3097983.3098118.

[14] Pfohl, S. R., Dai, A. M., & Heller, K. (2019). Federated and differentially private learning for electronic health records. ArXiv:1911.05861 [Cs, Stat]. Retrieved from htt p://arxiv.org/abs/1911.05861.

[15] Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. C., & Shi, W. (2018). Federated learning of predictive models from federated Electronic Health Records. International Journal of Medical Informatics, 112, 59–67. https://doi.org/10.1016/j. ijmedinf.2018.01.007.

[16] Lee, J., Sun, J., Wang, F., Wang, S., Jun, C.-H., & Jiang, X. (2018). Privacy-preserving patient similarity learning in a federated environment: development and analysis. JMIR Medical Informatics, 6(2), Article e20. https://doi.org/10.2196/ medinform.7744.

[17] Huang, L., Shea, A. L., Qian, H., Masurkar, A., Deng, H., & Liu, D. (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. Journal of Biomedical Informatics, 99, Article 103291. https://doi.org/10.1016/j. jbi.2019.103291.

[18] Liu, Z., Li, T., Smith, V., & Sekar, V. (2019). Enhancing the privacy of federated learning with sketching. ArXiv:1911.01812 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1 911.01812.

[19] Li Li, Yuxi Fan, Mike Tse and Kuo-Yi Lin, A Review of Applications in Federated Learning. Computers and Industrial Engineering, vol. 149, pp. 106854, 2020.

[20].Wang, F., Casalino, L. P. & Khullar, D. Deep learning in medicine—promise, progress, and challenges. JAMA Intern. Med. 179, 293–294 (2019).

[21]. Yang, Q., Liu, Y., Chen, T. & Tong, Y. Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. (TIST) 10, 12 (2019).

[22]. Kairouz, P. et al. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977 (2019).

[23]. Xu, J. & Wang, F. Federated learning for healthcare informatics. arXiv preprint arXiv:1911.06270 (2019).

[24]. Li, X., Huang, K., Yang, W., Wang, S. & Zhang, Z. On the convergence of fedavg on non-IID data. https://openreview.net/forum?id=HJxNAnVtDS (2020).

[25]. Li, W. et al. Privacy-preserving federated brain tumour segmentation. In International Workshop on Machine Learning in Medical Imaging, 133–141 (Springer, 2019).

[26]. Sheller, M. J., Reina, G. A., Edwards, B., Martin, J. & Bakas, S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. In International MICCAI Brainlesion Workshop, 92–104 (Springer, 2018).

[27]. Medical institutions collaborate to improve mammogram assessment ai. https:// blogs.nvidia.com/blog/2020/04/15/federated-learning-mammogramassessment/ (2020) (Accessed 28 May 2020)

[28]. Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A. & Smith, V. Federated optimization in heterogeneous networks. arXiv preprint arXiv:1812.06127 (2018).

[29]. Zhao, Y. et al. Federated learning with non-iid data. arxivabs/1806.00582 (2018).

[30]. Kairouz, P. et al. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977 (2019).

[31]. Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine 37, 50–60 (IEEE, 2020).

[32]. Wang, Z. et al. Beyond inferring class representatives: user-level privacy leakage from federated learning. In 2019 {IEEE} Conference on Computer Communications, {INFOCOM} 2512–2520. https://doi.org/10.1109/INFOCOM.2019.8737416 (IEEE, Paris, France, 2019).

[33]. Hitaj, B., Ateniese, G. & Perez-Cruz, F. Deep models under the gan: information leakage from collaborative deep learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS'17, 603–618 (Association for Computing Machinery, New York, NY, USA, 2017).

[34]. Nocola Rieke, Jonny Hancox, Wenqi Li et al., The future of digital health with federated learning. Npj Digital Medicine, vol. 3, no. 119, pp. 1-7, 2020.