Differentially Private Federated Learning for Medical Image Analysis

Mohammed Adnan[1 2 4], **Jesse C. Cresswell**[3], **Shivam Kalra**[1 3 4], **Graham W. Taylor**[2 4], **H.R. Tizhoosh**[1 4]

[1]Kimia Lab, University of Waterloo
[2]University of Guelph
[3]Layer 6 AI
[4]Vector Institute

m7adnan@uwaterloo.ca, jesse@layer6.ai, shivam.kalra@uwaterloo.ca, gwtaylor@uoguelph.ca, tizhoosh@uwaterloo.ca

Abstract

The artificial intelligence revolution has been spurred forward by the availability of large-scale datasets. In contrast, the paucity of large-scale medical datasets hinders the application of machine learning in healthcare. The lack of multi-centric and diverse datasets mainly stems from confidentiality and privacy concerns around sharing medical data. To demonstrate a feasible path forward, we conduct a case study of applying a differentially private federated learning framework to histopathology images. We study the effects of IID and non-IID distributions along with the number of healthcare providers and the individual dataset sizes, using The Cancer Genome Atlas (TCGA) dataset to simulate a distributed environment. We empirically compare the performance of private, distributed training to conventional training and demonstrate that distributed training can achieve similar performance with strong privacy guarantees. Our work indicates that differentially private federated learning is a viable and reliable framework for the collaborative development of machine learning models in the healthcare domain.

1 Introduction

Deep neural networks have achieved state-of-the-art results in many domains. However, deep learning algorithms are data-intensive, i.e., they often require millions of training examples to learn effectively. Medical images contain confidential and sensitive information about patients that often cannot be shared outside the institutions of their origin. The European General Data Protection Regulation (GDPR) and the United States Health Insurance Portability and Accountability Act (HIPAA) enforce guidelines and regulations for storing and exchanging personally identifiable data and health data. Ethical guidelines also encourage respecting privacy, that is, the ability to retain complete control and secrecy about one's personal information (Kaissis et al. 2020). As a result, large archives of medical data from consortia remain untapped sources of information. For instance, histopathology images cannot be collected and shared in large quantities due to the aforementioned regulations, as well as data size constraints given their enormous resolution. Without sufficient and diverse data, deep learning models trained on histopathology images from one hospital fail to generalize well on data from other hospitals (Aggarwal et al. 2021). The lack of diverse images from a single institution brings about the need for a collaborative approach which does not require data centralization.

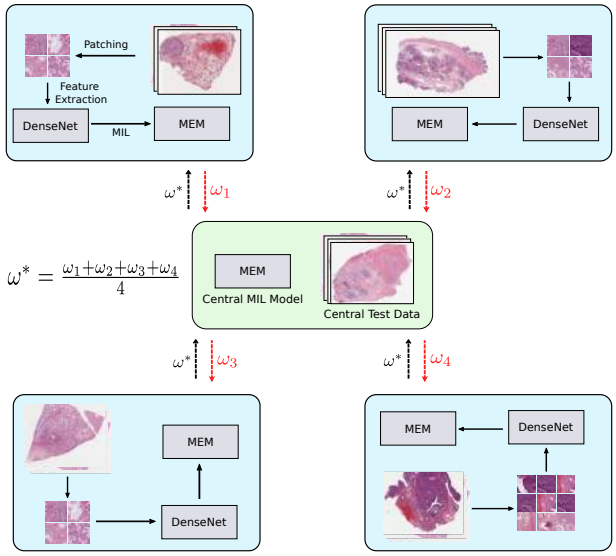


Figure 1: Our federated learning framework for histopathological image classification. The MEM model (Kalra et al. 2020a) is used for learning representations of sets of patches from the full resolution image.

In this paper, we explore differentially private federated learning as a collaborative learning paradigm where models are trained across several institutions without explicitly sharing patient data. We show that using federated learning with additional privacy preservation techniques can improve the performance of histopathological image classification compared to training without collaboration. We use lung cancer images from The Cancer Genome Atlas (TCGA) dataset (Weinstein et al. 2013) to construct a simulated environment of several institutions to validate our approach.

2 Background

Federated learning. Federated learning (FL) algorithms learn from decentralized data distributed across various client devices, in contrast to conventional learning algorithms. In most examples of FL, there is a centralized server which facilitates training a shared model and addresses critical issues such as data privacy, security, access rights, and heterogeneity (Kairouz et al. 2021). In FL, every client locally trains a copy of the centralized model, represented by the model weights ω, and reports its updates back to the server for aggregation across clients. Mathematically, FL can be formulated as:

$$\min_{\omega \in \mathbb{R}^d} f(\omega) \qquad \text{where} \qquad f(\omega) = \frac{1}{n} \sum_{i=1}^{n} f_i(\omega), \qquad (1)$$

where $f(\omega)$ represents the total loss function over n clients, and $f_i(\omega)$ represents the loss function with respect to client i's local data. The objective is to find weights ω that minimize the overall loss. McMahan et al. (2017) introduced federated averaging, or *FedAvg*, in which each client receives the current model ω_t from the server, and computes $\nabla f_i(\omega_t)$, the average gradient of the loss function over its local data. The gradients are used to update each client's model weights using stochastic gradient descent (SGD) as $\omega_{t+1}^i \leftarrow \omega_t - \eta \nabla f_i(\omega_t)$ according to the learning rate η. Next, the central server receives the updated weights ω_{t+1}^i from all participating clients and averages them to update the model, $\omega_{t+1} \leftarrow \sum_{i=1}^{n} \frac{n_i}{n} \omega_{t+1}^i$, where n_i is the number of data points used by client i.

Differential privacy. While FL attempts to provide privacy by keeping private data on client devices, it does not provide a meaningful privacy guarantee. Updated model parameters are still sent from the clients to a centralized server, and these can contain private information (Bhowmick et al. 2018), such that even individual data points can be reconstructed (Melis et al. 2019). *Differential privacy* (DP) is a formal framework for quantifying the privacy that a protocol provides (Dwork et al. 2006b). Consider a *database* $\mathcal{D}$, which is simply a set of datapoints, and a probabilistic function M acting on databases, called a *mechanism*. The mechanism is said to be (ϵ, δ)-*differentially private* if for all subsets of possible outputs $\mathcal{S} \subset \text{Range}(M)$, and for all pairs of databases $\mathcal{D}$ and $\mathcal{D}'$ that differ by one element,

$$\Pr[M(\mathcal{D}) \in \mathcal{S}] \leq \exp(\epsilon) \Pr[M(\mathcal{D}') \in \mathcal{S}] + \delta. \qquad (2)$$

When ϵ and δ are small positive numbers, Eq. (2) implies that the outcomes of M will be almost unchanged in distribution if one datapoint is changed in the database. In other words, adding one patient's data to a differentially private study will not affect the outcomes, with high probability.

The advantage of DP is that it is quantitative. It yields a numerical guarantee on the amount of privacy that can be expected, in the stochastic sense, where lower ϵ and δ implies that the mechanism preserves more privacy. The framework also satisfies several useful properties, including privacy under composition of mechanisms (Dwork, Rothblum, and Vadhan 2010) and under post-processing (Dwork and Roth 2014). Hence, while FL alone does not guarantee privacy, we can apply FL in conjunction with DP to give rigorous bounds on the amount of privacy afforded to clients and patients who participate in the collaboration. The simplest way to create a DP mechanism is by adding Gaussian noise to the outcomes of a deterministic function with bounded sensitivity (Dwork et al. 2006a). This method can be used in the context of training a machine learning model by clipping the norm of gradients to bound them, then adding noise, a process called differentially private stochastic gradient descent (DP-SGD) (Abadi et al. 2016).

Differential privacy for medical imaging. Past works have noted the potential solution DP provides for machine learning in healthcare. Kaissis et al. (2020) surveyed privacy-preservation techniques to be used in conjunction with machine learning, which were then implemented for classifying chest X-rays and segmenting CT scans (Kaissis et al. 2021; Ziller et al. 2021). In histopathology, Lu et al. (2020) reported DP guarantees for a neural network classifier trained with FL, following Li et al. (2020). Their treatment involved adding Gaussian noise to trained model weights, however, neural networks weights do not have bounded sensitivity making their DP guarantee vacuous. A meaningful guarantee would require clipping the model weights before adding noise, but this tends to destroy the finely tuned activations of the network. We propose the more standard approach of DP-SGD, which clips gradient updates leaving the gradient direction unchanged, and then adds noise for use in histopathology.

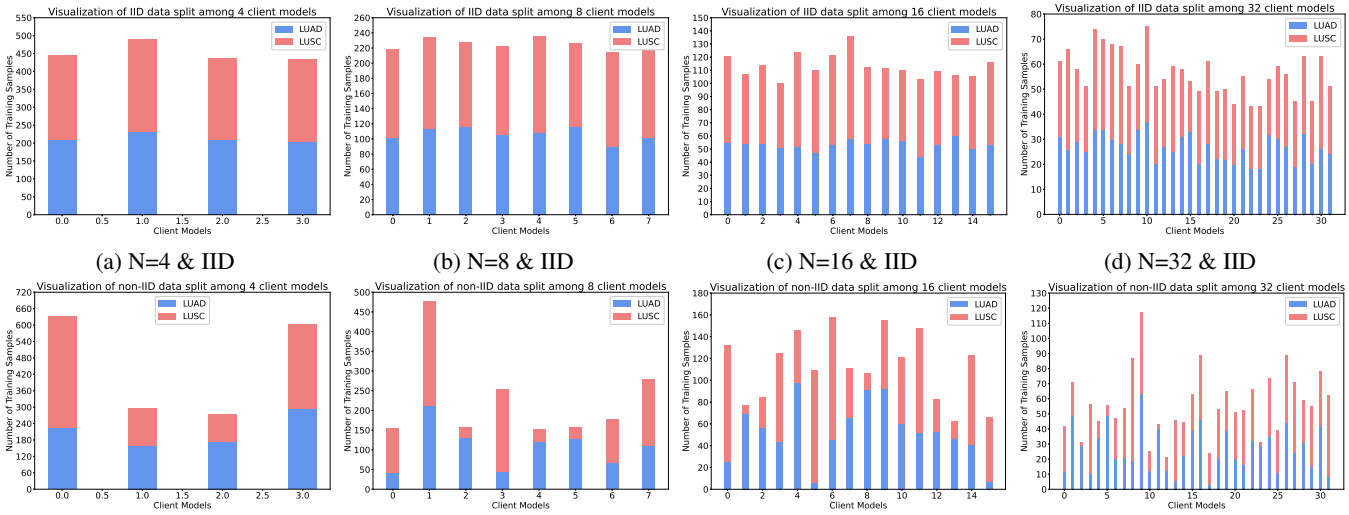
3 Experiments

We validated the performance of FL for the classification of histopathology images using a simulated distributed environment and also using real-world hospital data. Previous studies have mostly experimented with a fixed number of clients having similar distributions of data (Kaissis et al. 2020; Lu et al. 2020; Chang et al. 2018).

Since real-world data is not necessarily IID, it is important to study the effect of non-IID data on the performance of *FedAvg*. In the first experiment, we vary the number of clients, with each client representing one hospital. To make our simulated environment better approach the non-IID real-world data, each client can have a different number of images and a different distribution of cancer sub-types.

In the second experiment, we considered the effect of distributional differences from different source hospitals, and a requirement to preserve privacy. Histopathology images can differ greatly, depending on the staining and imaging protocols of the source hospital among other factors. We used the available attributes in TCGA to create four client datasets divided by the tissue origin site (hospital). Then we test on an external test dataset comprised of images from hospitals distinct from those participating in the FL scheme. Finally, to preserve privacy these models were trained using DP-SGD, and we calculate the privacy guarantee obtained from *differentially private* FL.

Datasets. We evaluated the FL framework for lung cancer subtype classification; Lung Adenocarcinoma (LUAD) and



(a) N=4 & IID (b) N=8 & IID (c) N=16 & IID (d) N=32 & IID

(e) N=4 & non-IID (f) N=8 & non-IID (g) N=16 & non-IID (h) N=32 & non-IID

Figure 2: Visualisation of IID and non-IID distribution of data among client models

Lung Squamous Cell Carcinoma (LUSC). We used TCGA data (Weinstein et al. 2013) and obtained 2,580 hematoxylin and eosin stained Whole Slide Images (WSIs) of lung cancer, comprising about two terabytes of data.

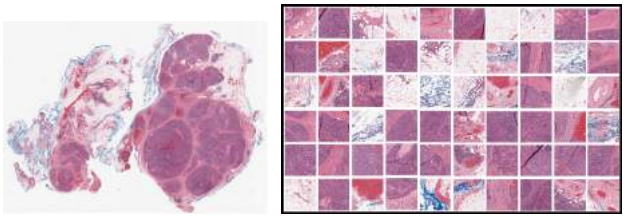
For the first experiment, the images were split into two groups of 1,806 training, and 774 testing samples WSIs. These sets were further divided amongst varying numbers of clients in IID and non-IID fashions. Data distributions for the IID and non-IID splits are visualized in Fig. 2.

For the second experiment, we selected images from four hospitals to create four clients with training and validation datasets randomly sampled in an 80:20 ratio. Then we combined images from three different hospitals for the external test dataset. The data split is summarized in Table 1.

Table 1: Source hospitals for training datasets. The top four hospitals represent clients and each splits their data into a training and validation set. The bottom three hospitals are combined into an external test set.

	Images		
Source Hospitals	LUAD	LUSC	Total
Int. Genomics Consortium	189	78	267
Indivumed	94	117	211
Asterand	90	117	207
John Hopkins	121	78	199
Christiana Healthcare	169	54	223
Roswell Park	35	75	110
Princess Margaret Hospital	0	52	52

Classification of histopathology images. To process the gigapixel histopathology images, we converted each full-resolution image into a set of representative patches, called a mosaic (Kalra et al. 2020b). An example WSI and mosaic are shown in Fig. 3. We extract a feature vector for each patch in the mosaic to create a set of features for each histopathology image, and classify the sets using a set-based model called MEM (Kalra et al. 2020a).



(a) A sample WSI (b) Extracted mosaic of 1000 × 1000 pixel tiles

Figure 3: Illustration of a sample WSI and its mosaic extracted using the approach in Kalra et al. (Kalra et al. 2020b)

MEM is composed of four sequentially connected components: a feature extraction model, memory units, memory blocks, and fully connected layers to predict the output. Memory units transform a given input sequence to an attention vector. A higher attention value represents a higher "importance" of the corresponding element of the input sequence. Essentially, the memory unit captures the relationships among different elements of the input. Memory blocks are the main component of MEM and learn a permutation invariant representation of a given input sequence. It is a sequence-to-sequence model, i.e., it transforms a given input sequence $X = x_1, \ldots, x_n$ to another representative sequence $\hat{X} = \hat{x}_1, \ldots, \hat{x}_m$, with the output sequence invariant to element-wise permutations of the input. A memory block contains several memory units, outputting several attention vectors which the memory block uses to compute the final output sequence. Multiple memory blocks are stacked for modeling complex relationships in exchangeable data.

Source Hospital	Without FL		With FL		Combined Training	
	Validation	External	Validation	External	Validation	External
Int. Genomics Consortium	0.654	0.631	0.823 ± 0.01	0.707 ± 0.01	0.839 ± 0.01	0.768 ± 0.003
Indivumed	0.648	0.556				
Asterand	0.709	0.701				
John Hopkins	0.681	0.600				

Table 2: Evaluation of collaborative and non-collaborative learning on validation and external test datasets using DP-SGD, achieving privacy parameter $\epsilon = 2.90$ for $\delta = 10^{-4}$. For FL and Combined training we report the mean accuracy and standard deviation across the clients' test datasets. On the external dataset we ran the experiments using three random initializations, and report the mean accuracy and standard deviation across them.

3.1 Results

Effect of number of clients and data distributions. We studied the effect of IID and non-IID distributions on the performance of FL while varying the number of clients (n), but with the total number of images fixed. We compared the performance with and without FedAvg for each of the 16 experimental settings in Table 3.

In each experiment, the server model trained using FedAvg outperformed the models trained using local client datasets, showing the advantage of collaboration. As the total dataset was divided into smaller partitions for more clients, both client and server model performances deteriorate. FL achieved superior performance for both IID and non-IID distributions of data compared to non-collaborative training. FL performed comparably to centralized training for reasonably sized datasets $(n = 4, 8)$. Interestingly, FL achieved better accuracy when trained on a non-IID data distribution for some values of n.

Table 3: Evaluation on different data distributions. Combined accuracy denotes the accuracy when the data is centralized and combined into one training set. The accuracy without FL is mean and standard deviation of accuracy values across multiple clients without any collaboration. The accuracy with FL/Combined is the accuracy of the central model over client test datasets.

Data	n	Accuracy		
		Without FL	With FL	Combined
IID	4	0.731 ± 0.03	0.824 ± 0.02	0.848 ± 0.02
	8	0.620 ± 0.06	0.780 ± 0.05	
	16	0.570 ± 0.03	0.726 ± 0.06	
	32	0.527 ± 0.02	0.641 ± 0.09	
Non IID	4	0.682 ± 0.10	0.824 ± 0.01	0.848 ± 0.02
	8	0.561 ± 0.08	0.823 ± 0.05	
	16	0.524 ± 0.03	0.750 ± 0.06	
	32	0.520 ± 0.03	0.550 ± 0.20	

External validation and privacy guarantees. We compared the performance of privacy-preserving FL training with both centralized training and non-collaborative training. In FL training, the four hospitals act as clients collaborating to train one central model. Performance is evaluated on each client's internal validation set, as well as the external test set. For comparison, we train a single model on the combined training datasets which gives an upper bound on what could be achieved in the absence of privacy regulations. Finally, in the non-collaborative setting each client hospital trains their own model on only their own training dataset. We used DP-SGD to train the FL and combined models and computed the privacy guarantees (ϵ, δ) using a Rényi DP accountant (Mironov 2017). We used a vectorized Adam optimizer (Subramani, Vadivelu, and Kamath 2020) with the following hyper-parameter values (Abadi et al. 2016): epochs = 180, training set size = 705, batch size = 32, gradient clipping norm = 1.0, Gaussian noise standard deviation = 4.0, number of microbatches = 32, learning rate = 2×10^{-5}. δ is fixed at 10^{-4}, significantly smaller than the inverse of the number of training data points.

As shown in Table 2, FL training achieves strong privacy bounds ($\epsilon = 2.90$ at $\delta = 10^{-4}$) with better performance than non-collaborative training, comparable to centralized training. This demonstrates that FL could be effectively used in clinical settings to ensure data privacy with minor degradation in performance.

4 Conclusions

There is a vast reserve of knowledge held by hospitals which remains mostly untapped due to many confidentiality and privacy concerns. In this case study, we explored differentially private federated learning as a potential method for learning from decentralized medical data such as histopathology images. Federated learning allows training models without explicitly sharing data and thus mitigates some confidentiality and privacy issues associated with medical data. Differential privacy supplements this with quantitative bounds on the amount of privacy provided. We demonstrated the efficacy of federated learning (FedAvg) with simulated real-world data, using both IID and non-IID data distributions. Private federated learning achieves a comparable result compared to conventional centralized training, and hence it could be considered for distributed training on medical data.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*.

Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D. S.; Karthikesalingam, A.; King, D.; Ashrafian, H.; and Darzi, A. 2021. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine*, 4(1): 1–23.

Bhowmick, A.; Duchi, J.; Freudiger, J.; Kapoor, G.; and Rogers, R. 2018. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*.

Chang, K.; Balachandar, N.; Lam, C.; Yi, D.; Brown, J.; Beers, A.; Rosen, B.; Rubin, D. L.; and Kalpathy-Cramer, J. 2018. Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association*, 25(8): 945–954.

Dwork, C.; Kenthapadi, K.; McSherry, F.; Mironov, I.; and Naor, M. 2006a. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In *Advances in Cryptology (EUROCRYPT 2006)*, volume 4004 of *Lecture Notes in Computer Science*, 486–503. Springer Verlag.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006b. Calibrating Noise to Sensitivity in Private Data Analysis. In Halevi, S.; and Rabin, T., eds., *Theory of Cryptography*, 265–284. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4): 211–407.

Dwork, C.; Rothblum, G. N.; and Vadhan, S. 2010. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, 51–60. IEEE.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; D'Oliveira, R. G. L.; Eichner, H.; Rouayheb, S. E.; Evans, D.; Gardner, J.; Garrett, Z.; Gascón, A.; Ghazi, B.; Gibbons, P. B.; Gruteser, M.; Harchaoui, Z.; He, C.; He, L.; Huo, Z.; Hutchinson, B.; Hsu, J.; Jaggi, M.; Javidi, T.; Joshi, G.; Khodak, M.; Konecný, J.; Korolova, A.; Koushanfar, F.; Koyejo, S.; Lepoint, T.; Liu, Y.; Mittal, P.; Mohri, M.; Nock, R.; Özgür, A.; Pagh, R.; Qi, H.; Ramage, D.; Raskar, R.; Raykova, M.; Song, D.; Song, W.; Stich, S. U.; Sun, Z.; Suresh, A. T.; Tramèr, F.; Vepakomma, P.; Wang, J.; Xiong, L.; Xu, Z.; Yang, Q.; Yu, F. X.; Yu, H.; and Zhao, S. 2021. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210.

Kaissis, G.; Ziller, A.; Passerat-Palmbach, J.; Ryffel, T.; Usynin, D.; Trask, A.; Lima, I.; Mancuso, J.; Jungmann, F.; Steinborn, M.-M.; et al. 2021. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, 3(6): 473–484.

Kaissis, G. A.; Makowski, M. R.; Rückert, D.; and Braren, R. F. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6): 305–311. Number: 6 Publisher: Nature Publishing Group.

Kalra, S.; Adnan, M.; Taylor, G.; and Tizhoosh, H. R. 2020a. Learning permutation invariant representations using memory networks. In *European Conference on Computer Vision*, 677–693. Springer.

Kalra, S.; Tizhoosh, H.; Choi, C.; Shah, S.; Diamandis, P.; Campbell, C. J.; and Pantanowitz, L. 2020b. Yottixel – An Image Search Engine for Large Archives of Histopathology Whole Slide Images. *Medical Image Analysis*, 65: 101757.

Li, X.; Gu, Y.; Dvornek, N.; Staib, L. H.; Ventola, P.; and Duncan, J. S. 2020. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Medical Image Analysis*, 65: 101765.

Lu, M. Y.; Kong, D.; Lipkova, J.; Chen, R. J.; Singh, R.; Williamson, D. F.; Chen, T. Y.; and Mahmood, F. 2020. Federated learning for computational pathology on gigapixel whole slide images. *arXiv preprint arXiv:2009.10190*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Melis, L.; Song, C.; De Cristofaro, E.; and Shmatikov, V. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, 691–706. IEEE.

Mironov, I. 2017. Rényi Differential Privacy. *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*.

Subramani, P.; Vadivelu, N.; and Kamath, G. 2020. Enabling Fast Differentially Private SGD via Just-in-Time Compilation and Vectorization. arXiv:2010.09063.

Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R. M.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; and Stuart, J. M. 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10): 1113–1120.

Ziller, A.; Usynin, D.; Braren, R.; Makowski, M.; Rueckert, D.; and Kaissis, G. 2021. Medical imaging deep learning with differential privacy. *Scientific Reports*, 11(1): 1–8.