PRIVATE CROSS-SILO FEDERATED LEARNING FOR EXTRACTING VACCINE ADVERSE EVENT MENTIONS

Pallika Kanani
Oracle Labs
pallika.kanani@oracle.com

Virendra J. Marathe
Oracle Labs
virendra.marathe@oracle.com

Daniel Peterson
Oracle Labs
daniel.peterson@oracle.com

Rave Harpaz
Oracle
rave.harpaz@oracle.com

Steve Bright
Oracle
steve.bright@oracle.com

ABSTRACT

Automatically extracting mentions of suspected drug or vaccine adverse events (potential side effects) from unstructured text is critical in the current pandemic, but small amounts of labeled training data remains silo-ed across organizations due to privacy concerns. Federated Learning (FL) is quickly becoming a goto distributed training paradigm for such users to jointly train a more accurate global model without physically sharing their data. However, literature on successful application of FL in real-world problem settings is somewhat sparse. In this paper, we describe our experience applying a FL based solution to the Named Entity Recognition (NER) task for an adverse event detection application in the context of mass scale vaccination programs. Furthermore, we show that *Differential Privacy (DP)*, which offers stronger privacy guarantees, but severely cripples the global model's prediction accuracy, thus dis-incentivizing users from participating in the federation. We demonstrate how recent innovation on *personalization* methods can help significantly recover the lost accuracy.

1 INTRODUCTION

As our world grapples with safely rolling out massive scale treatment and vaccination programs to end the COVID-19 pandemic, it is critical to understand adverse events (potential side effects) related to these drugs and vaccines. These adverse events are often expressed in free text form, such as social media posts and reports provided to health care agencies and pharmaceutical companies. Here's an example of such a report:

"Shortly after the patient was vaccinated, she started to feel an itching, tingling feeling in her throat. Fearing that it was an allergic reaction, I called 911."

Currently, mentions of specific adverse events are extracted and coded manually, which is a time consuming, expensive and non-scalable process. The use of and necessity of Machine Learning (ML) based automated methods for extracting such mentions from unstructured text is widely recognized in pharmacovigilance Harpaz et al. (2014). Several different genres of text are tackled in this line of research, including social media Gurulingappa et al. (2012); Korkontzelos et al. (2016), biomedical literature Leaman et al. (2010); Winnenburg et al. (2015), clinical narratives Haerian et al. (2012); LePendu et al. (2013) and drug labels Roberts et al. (2017). More recently, use of state of the art deep learning technology for Named Entity Recognition (NER) have been proposed Giorgi & Bader (2018).

Training these ML models requires data. Greater the amount of training data, the better the model performance. However, manually collecting and annotating this data is expensive, non-scalable, and particularly challenging, given the need to maintain privacy of health records. Though the resulting data scarcity problem can be addressed by sharing data amongst multiple institutions, privacy concerns, government regulations, and data use agreements may prohibit such data sharing. Federated Learning (FL) Bonawitz et al. (2019); Konecný et al. (2015), a distributed ML paradigm, may provide the perfect solution to this problem: Users can jointly train a ML model without sharing data with each other, offering advantages in both scale and privacy. Furthermore, much tighter privacy guarantees can be ensured via Differential Privacy (DP) enforcement mechanisms Dwork (2006);

Dwork & Roth (2014); Dwork et al. (2006); Abadi et al. (2016); Geyer et al. (2017); Konecný et al. (2016); McMahan et al. (2017).

Informally, DP forces a bound on the variation in the trained model's output based on the inclusion/exclusion of a single data point used in the training set. While DP enforces formally provable privacy guarantees, its employment, typically done by injecting noise in the training process, can lead to significant degradation in prediction accuracy of the resulting model, even making it worse than a user-resident model trained on just the user's data, which we call the *individual* model. This can dis-incentivize users from participating in the federation. However, recent work has shown that personalization can actually alleviate model degradation due to DP induced noise Peterson et al. (2019); Yu et al. (2020).

In this paper, we case study application of FL to the problem of vaccine adverse event detection, the first of its kind to the best of our knowledge. We use data from Vaccine Adverse Event Reporting System (VAERS), which is the prominent surveillance system for vaccines in the U.S., managed by the U.S. Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA). The VAERS data is de-identified, publicly available, and contains both textual narrative and meta data around vaccines and patients. We annotated the narratives for adverse event mentions and partitioned them by vaccine manufacturers. Each vaccine manufacturer acts as a federation user whose dataset is siloed in its private sandbox; all these sandboxes participate in Federated Learning over multiple training rounds.

Our experiments on the (VAERS) dataset reveal several interesting insights including general effectiveness of FL on model performance, effects of local DP enforcement on model performance, and the value of personalization techniques to incentivize users to participate in FL. In particular, we show that FL improves average F1 value by 37.43% over the individual model, while enforcement of local DP (DP-FL) degrades the FL model's average F1 by 25.17%. For one of the users, this degradation is so severe that the private FL model F1 is worse by 45.55% when compared with the individual model F1. This clearly makes DP-FL a non-starter for some users to join the federation. We study FL with *Fine-Tuning* (FT-FL) Yu et al. (2020), a personalization approach that fine-tunes the global model at each user *after* the entire FL training process completes. Interestingly, contrary to prior work Yu et al. (2020), simply augmenting fine-tuning to FL does not result in prediction accuracy improvement for the federation users. However, somewhat surprisingly, fine-tuning in the presence of DP (FT-DP-FL) boosts user accuracy by 24.88%, compared to the individual model, to strongly incentivize users to join and stay with the federation.

2 Federated Learning with Differential Privacy

In FL, a federation server initializes a global model and ships it to all participating users thereby initiating distributed training. Training happens over multiple rounds. In each round, each user, on receiving the the global model re-trains the model on its private data and sends back the resulting parameter updates to the federation server. The federation server aggregates updates from all users applying them to the global model, and then ships the revised model back to the users. The most widely used method of aggregation is FedAvg Konecný et al. (2015); McMahan et al. (2016), where user parameters updates are averaged at the federation server and applied to the global model.

Noting privacy concerns, more recent work has proposed addition of differential privacy to FL Geyer et al. (2017); Konecný et al. (2016); McMahan et al. (2016). *Differential privacy* Dwork et al. (2006) is a mathematically quantifiable privacy guarantee for a data set used by a computation that analyzes it. While it originally emerged in the database and data mining communities, triggered by privacy concerns in Machine Learning (ML) Fredrikson et al. (2015; 2014); Hitaj et al. (2017); Korolova (2010); Shokri et al. (2017); Tramèr et al. (2016), DP has garnered enormous traction in the ML community over the last decade Abadi et al. (2016); Carlini et al. (2019); Chaudhuri et al. (2011); Differential Privacy Team (2017); Dimitrakakis et al. (2017); Fredrikson et al. (2014; 2015); Park et al. (2016b;a); Sarwate & Chaudhuri (2013).

In the FL context, one can enforce DP using two distinct approaches: (i) *Global DP*, also called *Central DP* in the literature McMahan et al. (2017); Zhu et al. (2020), where users fully trust the federation server to enforce DP. The server in turn enforces DP to obfuscate the *participation* of each user. (ii) *Local DP* Differential Privacy Team (2017); Duchi et al. (2013); Kasiviswanathan et al. (2008); Truex et al. (2020), where users do not trust the federation server, and enforce DP on the updates shipped back to the server. This method of DP enforcement typically guarantees privacy to a finer granularity of individual training data points Liu et al. (2020).

Vaccine Manufacturer	Merck Co. Inc.	Sanofi Pasteur	Pfizer-Wyeth	Glaxo Smithkline Biologicals	Novartis Vaccines &Diagnostics	CSL Ltd.	Medimmune Vaccines Inc.	Seqirus Inc	Emergent Bio-solutions	Berna Biotech Ltd.
Num Reports	7638	3352	2428	2289	1183	465	265	111	58	52

Table 1: VAERS Dataset.'Vaccine Manufacturer' is a field in the public VAERS database that identifies the manufacturer of the vaccine reported in the VAERS form. There is no relationship between this field and the reporter. 'Num Reports' does not represent the rate of adverse events associated with the manufacturer or its products and cannot be used to estimate such rates. The statistics are based on a sample of reports submitted to VAERS between 2015-2017 whose MedDra coded adverse events appeared in the narrative. Because the statistics are based on a carefully selected sample, the distribution of reports shown may not represent the true distribution of reports associated with different vaccine manufacturers.

To enforce local DP, we use the algorithm proposed by Abadi et al. Abadi et al. (2016) that injects gaussian noise (calculated using their moments accountant algorithm) in parameter gradients during local training at each user. Noisy gradients lead to noisy parameter updates, which are eventually shipped from the user to the federation server. Since users can possess datasets with different sizes, the computed noise, which is a function of the dataset size, varies considerably from user to user.

Personalization through Fine Tuning The main allure of FL for a user is the promise of significant prediction accuracy improvements over a locally trained *individual* model. While parameter aggregation through FL can significantly improve accuracy of the global model, introduction of noise to enforce DP can severely compromise that improvement. The degradation can be severe enough to make users reconsider their decision to join the federation, and deter new users from joining the federation.

Researchers have recently proposed different forms of *personalization* approaches to remedy the problem of model degradation due to DP enforcement Peterson et al. (2019); Yu et al. (2020). Among the proposed personalization approaches, we focus on FL with *Fine Tuning* Yu et al. (2020): FT-FL for fine tuning on top of plain FL, and FT-DP-FL for fine tuning on top of FL with local DP enforcement. In this approach each user continues training, without noise, the local copy of the global differentially private model *after* the FL training process has completed.

The fine tuning based parameter updates are private to each user and are not shared with the federation. As a result, the fine tuned local models may diverge from the global model at varying degrees in order to better fit the users' private data. While endlessly fine tuning the global model can lead to the model converging to a locally trained individual model, standard hyperparameter tuning techniques can help ensure that the fine-tuned model does not deteriorate.

3 EXPERIMENTS

3.1 DATASET

The VAERS data (de-identified) is publicly available in structured format. Each VAERS report includes textual narrative, along with the name of (and additional information about) the administrated vaccine, a list of adverse events related to the vaccine, dates, and limited demographic information about the patient receiving the vaccine (e.g., age, gender).

We used a total of 17,841 narratives submitted to VAERS through the years 2015-2017 to form the NER data set used for this study. The narratives were automatically annotated for adverse event named entities using the list of adverse events supplied with each report. In total the NER data set used for this study comprised of 87,730 sentences and 39,139 annotated adverse event named entities. Table 1 describes the dataset. In our experiments, we split the data randomly into train, validation, tune and test sets in the proportion 60%, 10%, 10%, and 20% respectively. We used the validation set to decide early stopping in the fine tuning algorithm and tuned the rest of parameters on the tune set. We refer to "large manufacturers" as those with more than 1000 VAERS reports in this data and "small manufacturers" as those with fewer reports to reflect the availability of training data in each user's silo.

3.2 NER BASED ON RECURRENT NEURAL NETWORKS

The recurrent neural network (RNN) architecture we used to perform NER is based on a commonly applied BiLSTM architecture. The architecture consists of three major components: (1) a word representation layer made of word embeddings, (2) two stacked layers of bidirectional long short-term memory (LSTM) cells, and (3) a feedforward layer that performs the final BIO sequence labeling. We use pre-trained word embeddings to seed the network's word embedding layer. These were generated using Word2Vec applied to the sentences comprising the VAERS NER dataset. The network

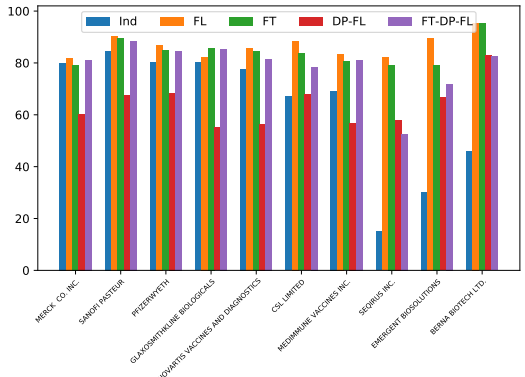

Figure 1: F1 per manufacturer for different methods for $\epsilon = 2.0$

was implemented on PyTorch6 and trained using stochastic mini-batch gradient descent with the Adam optimizer for a pre-defined number of iterations. Each iteration processed a batch of 256 randomly selected sentences. The network was trained for a total of 20 epochs, each epoch consisting of number of sentences in the training set / batch size iterations. Dropout regularization was implemented between each of the three major network components, with the drop rate of 0.4.

3.3 Experimental Setup

We have implemented our own FL simulation framework, on PyTorch6, that hosts the federation server and users on the same computer.

As the first baseline for our experiments, we train Individual models (*Ind*), i.e. assume that each manufacturer only uses their own training set, and test on their respective test set. This baseline represents the case in which the manufacturer chooses not to participate in the federation at all. *FL* is the federated learning model trained in a collaborative fashion across users using the FedAvg algorithm. This model is then fine tuned for each user using the protocol described in section 2, which yield a set of models, one per manufacturer, that we call *FT*. Next, we introduce local differential privacy to the *FL* model, as described in section 2. We use $\epsilon = 2.0$ for this first set of experiments as it is considered a fairly conservative privacy setting in the literature Abadi et al. (2016) and calculate the sigma values suitable per user. We call this private federated learning variant *DP-FL*. Finally, we fine tune this private FL model and call it *FT-DP-FL*.

The training parameters for all of these algorithms were tuned using a separate tuning dataset. We use a learning rate of 0.01 and train all the federated models for 20 rounds of FedAvg, with additional 20 epochs for the fine tuning variants at each manufacturer. For evaluation, we compute the precision, recall, and F1 of each token label on a 1-vs-all basis. The values reported are the mean F1 score (henceforth called F1) for the labels at the beginning or inside of an adverse event mention.

We ask the following questions as part of this study. Does *FL* perform better than *Ind* models across users? What happens when differential privacy is introduced? Does personalization help improve accuracy over *FL* and mitigate *DP-FL*'s accuracy loss enough to re-incentivize users to participate in the federation? In the appendix, we also study robustness to varying parameters of DP and stability against uncertainties of real world, such as users dropping out.

3.4 Private Federated Learning with Personalization

Figure 1 shows the F1 values for each of the described models on the individual users' test sets. Note that the manufacturers on the x-axis are sorted based on the size of their training sets. As we can see, the FL model consistently outperforms *Ind* models for each of the users, including large manufacturers with a lot of training data. Contrary to findings by Yu et. al. Yu et al. (2020), in our case, personalization based on fine tuning *FT-FL* performs worse than *FL* in most cases. As we add noise related to differential privacy to the federated learning model, F1 values drop significantly across the board. This makes participation for larger manufacturers in the federation unattractive, since the *DP-FL* model ends up performing worse than their *Ind* models. However, applying fine tuning in this case helps bring it back up to the point, where it is again advantageous for each party to participate in the federation. This shows that personalization based approach can help mitigate the loss of accuracy from introducing differential privacy.

It is interesting to note that for small manufacturers, with an exception of one with very small amount of evaluation data, it is always beneficial to participate in the federation, even for *DP-FL*, with or

without personalization. For large manufacturers however, the DP is only attractive in the presence of the mitigation offered by fine-tuning based personalization (*FT-DP-FL*).

4 CONCLUSION

Extracting mentions of vaccine adverse events using machine learning methods is an extremely urgent task right now. Federated Learning is a promising approach for breaking down organizational and geographical barriers to collaboration on building very effective models to solve this problem. Our work demonstrates that manufacturers with dataset of all different sizes can benefit from participating in such a federation, and that the loss of accuracy incurred through adding additional layers of privacy can be mitigated by introducing personalization.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konecný, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. *CoRR*, abs/1902.01046, 2019.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, pp. 267–284, 2019.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *The Journal of Machine Learning Research*, 12:1069–1109, July 2011.

Differential Privacy Team. Learning with Privacy at Scale, *https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html*, 2017.

Apple Differential Privacy Team. Learning with privacy at scale. *Machine Learning, Journal*, 1(8): 1–25, 2017.

Christos Dimitrakakis, Blaine Nelson, Zuhe Zhang, Aikaterini Mitrokotsa, and Benjamin I. P. Rubinstein. Differential privacy for bayesian inference through posterior sampling. *The Journal of Machine Learning Research*, 18(1):343–381, January 2017.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. *CoRR*, abs/1302.3203, 2013. URL `http://arxiv.org/abs/1302.3203`.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP*, pp. 1–12, 2006.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, August 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pp. 265–284, 2006.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333, 2015.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, pp. 17–32, 2014.

Robin C. Geyer, Tassilo Klein, and Moin Nabi. Differentially Private Federated Learning: A Client Level Perspective. *CoRR*, abs/1712.07557, 2017.

John M. Giorgi and Gary D. Bader. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, 34(23):4087–4094, 2018.

Ian J. Goodfellow. Efficient per-example gradient computations. *CoRR*, abs/1510.01799, 2015. URL http://arxiv.org/abs/1510.01799.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45 (5):885–892, 2012.

K. Haerian, D. Varn, S. Vaidya, L. Ena, H.S. Chase, and C. Friedman. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clinical Pharmacology and Therapeutics*, 92(2):228–234, 2012.

Rave Harpaz, Alison Callahan, Suzanne Tamang, Yen Low, David Odgers, Sam Finlayson, Kenneth Jung, Paea LePendu, and Nigam Shah. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug safety : an international journal of medical toxicology and drug experience*, 37, 08 2014.

Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the gan: Information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 603–618, 2017.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? *CoRR*, abs/0803.0924, 2008. URL http://arxiv.org/abs/0803.0924.

Jakub Konecný, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. *CoRR*, abs/1511.03575, 2015. URL http://arxiv.org/abs/1511.03575.

Jakub Konecný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527, 2016.

Ioannis Korkontzelos, Azadeh Nikfarjam, Matthew Shardlow, Abeed Sarker, Sophia Ananiadou, and Graciela H. Gonzalez. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics*, 62:148–158, 2016. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2016.06.007. URL https://www.sciencedirect.com/science/article/pii/S1532046416300508.

A. Korolova. Privacy violations using microtargeted ads: A case study. In *2010 IEEE International Conference on Data Mining Workshops*, pp. 474–482, 2010.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts in health-related social networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, BioNLP@ACL 2010, Uppsala, Sweden, July 15, 2010*, pp. 117–125. Association for Computational Linguistics, 2010.

P. LePendu, S.V. Iyer, A. Bauer-Mehren, R. Harpaz, J.M. Mortensen, T. Podchiyska, T.A. Ferris, and N.H. Shah. Pharmacovigilance using clinical notes. *Clinical Pharmacology and Therapeutics*, 93:547–555, 2013.

Yuhan Liu, Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, and Michael Riley. Learning discrete distributions: user vs item-level privacy. *CoRR*, abs/2007.13660, 2020. URL https://arxiv.org/abs/2007.13660.

H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private language models without losing accuracy. *CoRR*, abs/1710.06963, 2017. URL http://arxiv.org/abs/1710.06963.

Mijung Park, James R. Foulds, Kamalika Chaudhuri, and Max Welling. Private Topic Modeling. *CoRR*, abs/1609.04120, 2016a.

Mijung Park, Jimmy Foulds, Kamalika Chaudhuri, and Max Welling. Practical privacy for expectation maximization. *CoRR*, abs/1605.06995, 2016b.

Daniel W. Peterson, Pallika Kanani, and Virendra J. Marathe. Private federated learning with domain adaptation. *CoRR*, abs/1912.06733, 2019. URL http://arxiv.org/abs/1912.06733.

Kirk Roberts, Dina Demner-Fushman, and Joseph M. Tonning. Overview of the TAC 2017 adverse reaction extraction from drug labels track. In *Proceedings of the 2017 Text Analysis Conference, TAC 2017, Gaithersburg, Maryland, USA, November 13-14, 2017*. NIST, 2017.

A. D. Sarwate and K. Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Processing Magazine*, 30(5):86–94, 2013.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *Proceedings of the 25th USENIX Conference on Security Symposium*, pp. 601–618, 2016.

Stacey Truex, Ling Liu, Ka Ho Chow, Mehmet Emre Gursoy, and Wenqi Wei. Ldp-fed: federated learning with local differential privacy. In *Proceedings of the 3rd International Workshop on Edge Systems, Analytics and Networking, EdgeSys@EuroSys 2020, Heraklion, Greece, April 27, 2020*, pp. 61–66. ACM, 2020.

Rainer Winnenburg, Alfred Sorbello, Anna Ripple, Rave Harpaz, Joseph Tonning, Ana Szarfman, Henry Francis, and Olivier Bodenreider. Leveraging medline indexing for pharmacovigilance - inherent limitations and mitigation strategies. *Journal of Biomedial Informatics*, 2015.

Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *CoRR*, abs/2002.04758, 2020. URL https://arxiv.org/abs/2002.04758.

Wennan Zhu, Peter Kairouz, Brendan McMahan, Haicheng Sun, and Wei Li. Federated heavy hitters discovery with differential privacy. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3837–3847. PMLR, 2020.

A Appendix

A.1 Robustness to Differential Privacy Noise

Next, we study the effectiveness of personalization in recovering from the accuracy loss resulting from differential privacy noise. We vary the parameter ϵ and measure F1 averaged across users for two of the algorithm variants: differentially private federated learning (DP-FL) and the fine tuned differentially private federated learning (FT-DP-FL). As we can see from Figure 2, average F1 for DP-FL deteriorates significantly for values of ϵ less than 2. However, even in these cases, the personalized version, FT-DP-FL manages to retain its performance. We believe this is an important finding that provides significant latitude to differentially private FL frameworks to further tighten the privacy budget of ϵ without compromising utility.

A.2 Stability of Federation against Users Leaving

Building a federation across organizations can be challenging in the real world due to a variety of factors. For instance, users may discontinue their participation in the federation. We simulate this scenario and study the effect of one of the manufacturers leaving the federation. As we can see from Tables 2 and 3, both federated learning and private federated learning with fine tuning are fairly stable against such a change, with the exception of a few manufacturers with very small amount of

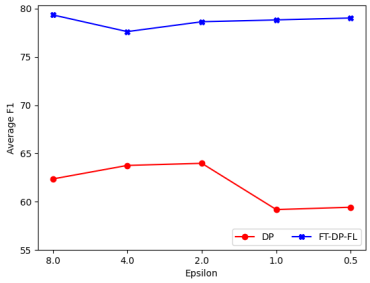

Figure 2: Average F1 across users for the two differentially private FL variants.

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0	0.9	1.8	0.4	1.0	2.1	1.8	0.4	1.0	0.0
M2	-0.3	0.0	0.4	0.5	1.4	1.6	1.6	-0.4	3.2	-1.5
M3	-0.1	0.5	0.0	0.1	0.1	0.9	1.4	1.9	1.0	-1.5
M4	-0.6	0.8	0.2	0.0	2.6	-0.2	3.5	1.3	1.0	0.0
M5	-0.5	-0.1	-0.1	2.9	0.0	0.6	0.6	-1.9	1.0	0.0
M6	-0.8	0.0	0.2	-0.5	-0.4	0.0	1.6	-1.1	2.1	0.0
M7	-0.5	0.5	-0.3	-0.5	0.1	0.7	0.0	0.4	1.0	-1.5
M8	-0.7	0.3	0.3	-0.1	-0.5	0.0	-0.5	0.0	0.8	0.0
M9	-0.4	0.1	0.2	0.0	0.4	0.1	0.9	0.9	0.0	4.5
M10	-1.0	0.0	-0.2	-0.2	-0.2	0.3	-1.3	-1.1	0.0	0.0

Table 2: Stability of FL performance when a single user leaves. M1-M10 are manufacturers sorted in descending order by size. Each row represents a manufacturer that is leaving the federation. Each Column represents the difference between F1 values under full federation and this reduced federation for that manufacturer.

training and test data. In other words, no single manufacturer has disproportionally large impact on the overall accuracy gains from participating in the federation.

A.3 Federation of Small Manufacturers

Another scenario that we simulate is the one where only participants with small amount of training data agree to collaborate. In this case, we do not have the advantage of the large amount of training data from any of the larger manufacturers. To better understand if such a federation is still advantageous, we compare the F1 values for small manufacturers in two different scenarios: one, in which they are a part of a large federation with all manufacturers, and second, in which they are a part of a federation with only the small manufacturers. Figures 3 and 4 show these comparisons for FL and FT-DP-FL respectively. As is clear from the bar chart, even in the case of a federation with just the small manufacturers, most of the manufacturers benefit significantly from participating. In fact,

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0	0.1	0.4	1.9	-2.4	1.4	2.9	-8.3	0.3	15.8
M2	-0.1	0.0	0.6	1.6	-1.5	-1.6	0.5	-2.5	1.4	22.5
M3	0.5	0.5	0.0	2.1	-1.7	0.2	-1.3	-1.2	-1.2	3.7
M4	-0.3	-0.3	0.2	0.0	-0.1	-4.3	0.7	-1.3	-0.4	18.7
M5	-0.1	0.0	-0.3	1.0	0.0	-0.3	-0.3	-1.9	-0.8	0.5
M6	-0.2	-0.5	0.3	1.6	-1.9	0.0	-1.5	-0.3	-0.5	4.2
M7	-0.5	0.1	0.3	2.2	-1.2	-2.8	0.0	-0.5	0.9	28.9
M8	0.5	-0.5	0.8	0.6	0.0	-4.0	-0.9	0.0	5.2	15.8
M9	-0.5	-0.5	0.3	1.0	-2.5	-3.3	-3.5	-2.4	0.0	4.1
M10	-0.1	-0.2	1.0	0.9	-1.8	-3.2	-0.1	-1.4	2.2	0.0

Table 3: Stability of Private FL with Fine Tuning performance when a single user leaves. M1-M10 are manufacturers sorted in descending order by size. Each row represents a manufacturer that is leaving the federation. Each Column represents the difference between F1 values under full federation and this reduced federation for that manufacturer.

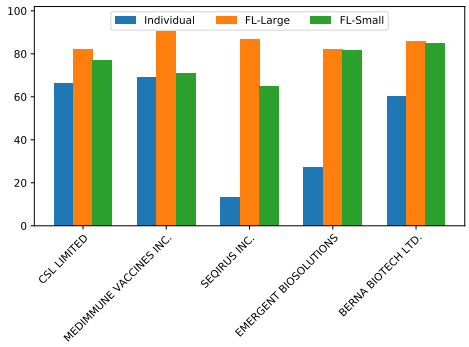

Figure 3: Comparison of FL F1 for small manufacturers when they are a part of a larger federation vs. a federation of only small manufacturers.

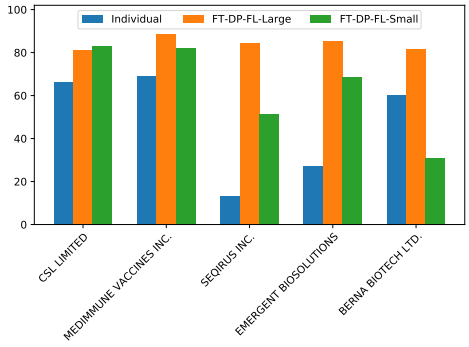

Figure 4: Comparison of FT-DP-FL F1 for small manufacturers when they are a part of a larger federation vs. a federation of only small manufacturers.

the performance of all manufacturers in the small federation closely tracks their performance in the large federation, with one exception.

A.4 Training Time

All experiments were run on the Oracle Cloud Infrastructure cluster of Tesla V100 GPUs running a job scheduling software. The GPUs were either 1, 2 or 8 core, with 90G, 180G, 768G memory respectively. Here we report the actual wall clock time for training different variants of federated learning. Training the FL model took 7.95 minutes, while training it and tuning it for each of the users in a serial fashion took a total of 17.10 minutes. The DP-FL model took 505.84 minutes to train by itself and 559.34 minutes with fine tuning. The DP models took over an order of magnitude of training time because during training the DP noise injection code path computes and clips gradients of individual data points in a training mini-batch before applying gaussian noise to the averaged mini-batch gradients. This is necessary to ensure that the training algorithm respects the allotted ϵ privacy budget over the training process. Parallelization of this component of our system using Goodfellow's technique Goodfellow (2015) is the subject of future work.