Patient Privacy in Differentially Private Diagnostic Models

Priit Järv
priit.jarv@utu.fi
Department of Computing
University of Turku
Turku, Finland

Tapio Pahikkala
aatapa@utu.fi
Department of Computing
University of Turku
Turku, Finland

Antti Airola
ajairo@utu.fi
Department of Computing
University of Turku
Turku, Finland

Abstract

Differential privacy gives a theoretical guarantee against inferring the presence of a patient in the training data of a model. Patients can have multiple data records that are used as separate training inputs, but to apply differential privacy we must view the contribution of a patient as a single privacy unit. The main approaches are user level privacy that uses all records of the patient as the privacy unit, group privacy which simplifies computation by allowing each patient to have at most k records, and sample level privacy that requires each patient to contribute one record. Of these, sample level privacy is well supported by efficient deep learning libraries and can be adapted to group privacy. User level differential privacy in the medical domain remains poorly supported by software and, as a consequence, unexplored. We compare these three approaches in experiments with cardiovascular disease and melanoma datasets. We found that in case most patients contribute only a single data record, sample level privacy suffices. With all patients contributing multiple records, user level privacy performed best, because it avoids the limitations of the alternative approaches: restricting the number of examples per patient, or inaccurately assuming all patients have the same number of records in privacy accounting.

CCS Concepts

• **Applied computing** → *Health informatics*; • **Computing methodologies** → *Neural networks*; • **Security and privacy**;

Keywords

Differential privacy, Neural networks, Health informatics

ACM Reference Format:
Priit Järv, Tapio Pahikkala, and Antti Airola. 2026. Patient Privacy in Differentially Private Diagnostic Models. In *Proceedings of the 12th ACM International Workshop on Security and Privacy Analytics (IWSPA '26), June 23–25, 2026, Frankfurt am Main, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3806007.3810963

1 Introduction

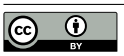
Deep learning models are vulnerable to membership inference and gradient inversion attacks [22, 28]. These attacks may determine that a particular data record was part of model training, or even partially reconstruct training inputs. For diagnostic models, training data comes from the clinical records of patients. Minimizing the risk of successful attacks in diagnostic models is both an important research problem and an ethical responsibility.

Differential privacy can quantify and limit the probability of inferring if an individual patient contributed to model training. The privacy protection formally applies to a part of the training data called the privacy unit. The term sample level privacy is used when the unit is one training example, and user level privacy is used when the privacy unit is all of the training examples belonging to a patient. Group privacy is a method of extending the privacy protection over a fixed number of units. The strength of the privacy protection w.r.t. the privacy unit is determined by the (ε, δ)-privacy budget, which we formally define in Section 2.

Although user level privacy may seem like the principled choice, there are practical difficulties with efficient implementation, which we will explain shortly. We must therefore also consider the following alternatives that avoid user level privacy. In cases where most patients have only one data record, it is possible to limit the training examples to one per patient and use the simpler sample level privacy approach. Group privacy can be implemented by subsampling the training dataset such that each patient has at most k records. In this paper we investigate, whether user level privacy is really required and how it compares to simpler approaches of sample level and group privacy. We experiment with training models for multilabel ECG classification and melanoma classification from dermoscopy images.

Any differentially private output must have limited sensitivity to privacy unit sized changes in data. Abadi et al. solved this for deep learning with the differentially private stochastic gradient descent algorithm (DP-SGD), which treats the gradient computation as a differentially private query [1]. In theory, obtaining user level privacy using this principle only requires limiting the gradient over all training examples of one individual to some norm C. In practice, deep learning frameworks rely on batch processing for efficient training and accessing individual sample gradients is slow and complicated [14]. With user level privacy, further complication to efficient, parallelized training is that computations are required over variable-sized sets of records.

Extensive literature exists on user level differential privacy (ULDP) in cross-device federated learning, following the landmark paper of McMahan et al. [16]. In the cross device setting, all training examples on one device belong to one user and computations on individual example level are unnecessary. ULDP is a natural choice for federated learning and straightforward to implement using application frameworks like Flower [3]. The algorithms developed in this setting are not applicable to a typical medical dataset, where the patients do not contribute enough records to obtain viable models separately [12] and their records must be mixed together in training.



IWSPA '26, Frankfurt am Main, Germany

ACM ISBN 979-8-4007-2609-5/2026/06
https://doi.org/10.1145/3806007.3810963

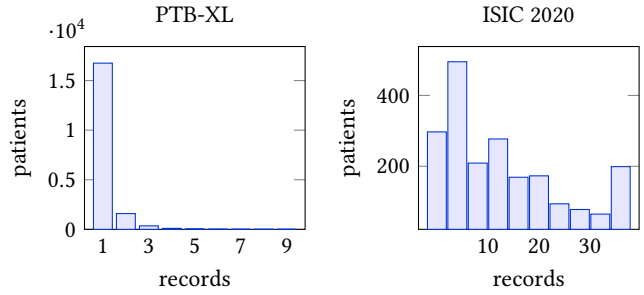


Figure 1: Distribution of records per patient in datasets.

We will not discuss federated learning further, because in this paper we consider the problem of training a model from a single dataset. In this setting, user level privacy has been considered in the context of training large models in a data center. Empirical work has so far restricted the contribution of one user to some limit [7], controlling sensitivity at batch level for computational efficiency [26], or sampling a fixed number of records from each user [8]. To the best of our knowledge, user level privacy algorithms that do not restrict or place any assumptions on the number of records per user have been mostly covered in theoretical papers [12, 15].

In contrast, group privacy is not an algorithmic approach but a method of privacy accounting. Privacy protection can be extended to a fixed k examples by adjusting the sample level (ε, δ) privacy bound to $(k\varepsilon, ke^{(k-1)\varepsilon}\delta)$ [11]. For DP-SGD a more precise bound has been proven [7].

Figure 1 depicts the distribution of training examples per patient in two medical datasets: PTB-XL ECG data for cardiovascular disease classification, and ISIC 2020 dermoscopy images for melanoma classification. The distribution of the number of ECG records (Figure 1, left) is what we would expect, if most patients have only one measurement taken for a specific medical diagnosis. The number of dermoscopy images (Figure 1, right) shows that either due to data collection methodology or the nature of the medical problem addressed, distributions can also be shifted towards higher number of records per patient. This already suggests that there is no "one size fits all" approach for providing patient privacy, as the effect of limiting training examples to one would vary from negligible to severe. At the same time, we do not know what costs to the utility are incurred by approaches that attempt to use more examples. All approaches considered in this paper have some drawback that negatively affects predictive performance of trained models, either by reducing available training data or requiring increased privacy preserving perturbation in training.

The goal of this paper is to shed some light on which differentially private approaches for preserving patient privacy in diagnostic models are best at maintaining prediction accuracy. We provide new evidence for the following questions:

- RQ1: What is the comparative prediction accuracy of sample level privacy, group privacy and user level privacy on different datasets, given equal privacy budget constraints?
- RQ2: What factors can be associated with the differences in predictive performance between compared approaches?
- RQ3: When is using approaches other than sample level privacy justified?

In particular, our contribution is to analyze, implement and evaluate the full user level privacy training algorithm, with no restriction on the number of training examples. This paper represents early results which we intend to expand on during ongoing work. We also provide a PyTorch implementation for ULDP training under open source license, available from https://github.com/priitj/sulp. To the best of our knowledge, there is only one empirical paper using this privacy model outside the federated learning setting by Chua et al. [8]. Their user-wise DP-SGD algorithm differs from ours mainly by taking a fixed size subsample of user records in each step.

The paper is organized as follows. In Section 2 we give the mathematical background and describe the algorithm for training with ULDP. In Section 3 we describe the experimental setup and in Section 4 the results. We summarize the findings in Section 5. The source code of the experiments and data access instructions are available at https://github.com/priitj/iwspa26.

2 Preliminaries

In this section we describe our algorithm of training models with user level privacy and briefly review the mathematical background. We begin by restating the definitions of differential privacy and the models of privacy compared in this paper. In definitions and when discussing algorithms, we follow the convention in differential privacy literature, where the "user" is an individual who owns one or multiple training examples in a dataset. When discussing application of algorithms in the medical setting, we use "patient".

Definition 2.1 (Differential Privacy [1]). A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with a domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ε, δ)-differential privacy if for any two adjacent datasets $d, d' \in \mathcal{D}$ and for any subset of outputs $S \subseteq \mathcal{R}$ it holds that $Pr[\mathcal{M}(d) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(d') \in S] + \delta$.

Definition 2.2 (Adjacency). Let d and d' be two datasets of training examples. Then, d and d' are k-adjacent if d' can be formed by adding $1 \leq n \leq k$ examples to, or removing $1 \leq n \leq k$ from d. If $k = 1$, d and d' are sample-adjacent. d and d' are user-adjacent, if d' can be formed by adding all examples of one user to, or removing all examples of one user from d.

We have sample level differential privacy if in Definition 2.1 d and d' are sample-adjacent. With $k > 1$-adjacency and user-adjacency, we have group privacy and user level privacy, respectively.

For sample level privacy and group privacy, we use the standard DP-SGD algorithm [1]. To obtain (ε, δ)-differential privacy for a patient this way, we train on a subsample of the dataset where the number of samples per patient is restricted to k. If $k = 1$, the privacy protection for the patient is equivalent to sample level privacy. If $k > 1$, we adjust the privacy budget according to the improved group privacy bound derived by Charles et al. [7], using the Google differential privacy library[1].

For user level privacy, we use Algorithm 1. We train a model θ for K iterations using $x_1, \ldots, x_n$ training examples, each belonging

[1] https://github.com/google/differential-privacy

Algorithm 1 Training with user level differential privacy

Require:
$x_1, x_2, \ldots, x_n$ – n training examples, U – set of patients, K – steps, q – sampling rate, $\mathcal{L}$ – loss function, C – max. gradient norm, z – noise ratio

Initialize θ_0 randomly
for $i \leftarrow 1 \ldots K$ **do**
  $U_i \leftarrow \{u \mid u$ selected from U with probability $q\}$
  for $u \in U_i$ **do**
    $B_u \leftarrow \{x_j \mid x_j$ belongs to $u\}$
    $\Delta_{u,i} \leftarrow \sum_{x_j \in B_u} \nabla_{\theta_{i-1}} \mathcal{L}(\theta_{i-1}, x_j)$
    $\bar{\Delta}_{u,i} \leftarrow \Delta_{u,i} \min(1, \frac{C}{||\Delta_{u,i}||_2})$ ▷ clipping
  end for
  $\tilde{\Delta}_i \leftarrow \frac{1}{qn} \sum_{u \in U_i} \bar{\Delta}_{u,i} + \mathcal{N}(0, (\frac{zC}{qn})^2 \mathbb{I})$
  $\theta_i \leftarrow$ OptFunc$(\theta_{i-1}, \tilde{\Delta}_i)$
end for
θ_K is the differentially private model

to one user $u \in U$. Each iteration i of the algorithm takes a sample of users by including each user with probability q. This process is known in differential privacy literature as Poisson sampling with rate q. We compute the sum of per-sample gradients over all training examples of each user. If the L_2 norm of the sum $\Delta_{u,i}$ exceeds the maximum norm C, the clipping operation scales the gradient such that the norm is at most C. We estimate the mean gradient by scaling the sum of clipped gradients $\bar{\Delta}_{u,i}$ over all selected users by $\frac{1}{qn}$, where qn is the expectation of number of samples per iteration. The normally distributed noise scaled to the ratio z provides the mechanism of privacy protection.

The privacy bound of Algorithm 1 is based on the following definitions:

Definition 2.3 (Sensitivity). The L_2–sensitivity of a function f is

$$S_f = \max_{d,d' \in \mathcal{D} \wedge d,d' \text{ are adjacent}} ||f(d) - f(d')||_2$$

Definition 2.4 (Sampled Gaussian Mechanism [18]). Let u denote an element of $d \in \mathcal{D}$ such that d and $d \setminus \{u\}$ are adjacent. Let $f : \mathcal{D} \to \mathbb{R}^n$. The Sampled Gaussian Mechanism (SGM) with sampling rate $0 < q \leq 1$ and noise standard deviation σ is

$$\mathcal{M}_{q,\sigma}(d) = f(\{u : u \in d \text{ selected with probability } q\}) + \mathcal{N}(0, \sigma^2 \mathbb{I})$$

Definition 2.5 (Adaptive sequential composition [17]). Given randomized mechanisms $\mathcal{M}_1 : \mathcal{D} \to \mathcal{R}$ and $\mathcal{M}_2 : \mathcal{R} \times \mathcal{D} \to \mathcal{R}$, adaptive sequential composition is the release of $X \sim \mathcal{M}_1(\mathcal{D})$ and $Y \sim \mathcal{M}_2(X, \mathcal{D})$.

We show that Algorithm 1 satisfies (ε, δ)–differential privacy. Assuming (a.) $\tilde{\Delta}_1 \ldots \tilde{\Delta}_i$ are (ε, δ)–differentially private, θ_i are obtained by postprocessing and are differentially private. $\tilde{\Delta}_1 \ldots \tilde{\Delta}_i$ are obtained by adaptive sequential composition of the SGM $\mathcal{M}_{q,\sigma}(U)$ where $f(U_i \subset U) = \frac{1}{qn} \sum_{u \in U_i} \bar{\Delta}_{u,i}$. If $S_f = 1$, prior theory provides the mathematical tools to compute the noise ratio z from the number of composition steps K and sampling rate q, such that setting $\sigma = z$ makes $\tilde{\Delta}_1 \ldots \tilde{\Delta}_K$ (ε, δ)–differentially private. We will explain the computation of z below. For our assumption (a.) to hold, it is sufficient to set $\sigma = \frac{zC}{qn}$ and show that $S_f \leq \frac{C}{qn}$.

Proposition 2.6. *The sensitivity of* $f(U_i) = \frac{1}{qn} \sum_{u \in U_i} \bar{\Delta}_{u,i}$ *in Algorithm 1 is* $S_f \leq \frac{C}{qn}$.

Proof. For $S_f \leq \frac{C}{qn}$ to hold, the following must hold for any adjacent subset $U_i' = U_i \cup \{u'\}$:

$$\left\| \frac{1}{qn} \sum_{u \in U_i'} \bar{\Delta}_{u,i} - \frac{1}{qn} \sum_{u \in U_i} \bar{\Delta}_{u,i} \right\|_2 \leq \frac{C}{qn} \tag{1}$$

$$\left\| \sum_{u \in U_i'} \bar{\Delta}_{u,i} - \sum_{u \in U_i} \bar{\Delta}_{u,i} \right\|_2 \leq C \tag{2}$$

$$||\bar{\Delta}_{u',i}||_2 \leq C \tag{3}$$

The requirement (3) is fulfilled by the clipping operation in Algorithm 1. □

K, q and C can be treated as hyperparameters. In principle, for a given choice of K and $0 < q \leq 1$ and a budget (ε, δ), it is possible to calculate the noise ratio z using the classical Gaussian mechanism and sequential composition [11]. Because our training process requires hundreds or thousands of model updates, we use the refined privacy bounds for the sequential composition of the SGM [2, 17, 18] that have a software implementation in the Opacus[2] differential privacy library. Opacus performs a search for the least value of parameter z that satisfies some (ε, δ) privacy budget, with given K and q. The choice of z is independent of the maximum norm C.

3 Methods

In this section, we describe our experimental setup. To answer our research questions, we train models that give same privacy protection to the patient, but differ by the approach: restricting the number of training examples, or using full user level privacy. We then compare the model performance in cardiovascular disease and melanoma classification tasks.

We compare the following approaches:

- Restrict the number of records per patient to $k = 1$ and apply sample level privacy.
- Restrict the number of records per patient to $k > 1$ and train with per-sample gradient clipping. The privacy budget is computed by applying group privacy with group size k.
- User level privacy (ULDP).

In the ECG cardiovascular disease task, we use the PTB-XL dataset [13, 23, 24]. PTB-XL contains 21799 clinical 12-lead ECG records of 10 seconds from 18869 patients. For the multilabel classification task, we use the following six labels: 1st degree atrioventricular block, right bundle branch block, left bundle branch block, sinus bradycardia, atrial fibrillation and sinus tachycardia. In the melanoma classification task, we use two challenge datasets of skin lesion classification. For training, we use the ISIC 2020 dataset of 33126 dermoscopy images from 2056 patients [21]. The test set of the ISIC 2020 challenge is not public, so we use the ISIC 2017 test

[2] https://github.com/meta-pytorch/opacus

Table 1: Disease prevalence in training and test data

Label	Training	Test
1st degree AV block	808	96
right bundle branch block	432	54
left bundle branch block	428	54
sinus bradycardia	509	64
atrial fibrillation	1211	152
sinus tachycardia	661	82
total ECG records	17418	2198
melanoma	584	117
total dermoscopy images	33126	600

set of 600 images [9], which does not overlap with ISIC 2020 [6]. The negative and positive labels in the datasets are not balanced; we give the number of positive labels in Table 1.

There is no established protocol for fair comparison between models trained using different privacy units. Our approach is to fix the privacy related parameters: the (ε, δ) budget and maximum gradient norm C. In addition, we fix the training effort by approximate number of epochs Kq, and choose the sampling rate q such that approximately 1024 samples are used in each optimizer step. Borrowing the concept of epochs makes the spent training effort easy to interpret, and follows the design paradigm of Opacus.

We briefly discuss alternative approaches of determining equal training effort and why we did not choose them. Firstly, it is possible to fix the sampling rate q and use equal number of epochs Kq. However, for the approaches that use a subsample of k records per patients, the amount of training data is reduced. In case of the ISIC 2020 melanoma data and $k = 1$, there are only 2056 training examples, meaning that with a fixed q the number of records per optimizer step becomes small. There is empirical and theoretical evidence that this has detrimental effect with differential privacy [20]. Because we also observed higher training loss compared to variable q experimentally, we abandoned this approach. Secondly, Charles et al. fixed the number of training examples processed in total in order to compare algorithms using equal compute [7]. Again, this approach is problematic in cases with $k = 1$, because with a smaller training sample the number of iterations has to be increased, which also requires increasing the noise variance σ^2. This made it difficult to maintain a stable training regime under our evaluated $\varepsilon = 10$ privacy budget. While theoretically sequential composition of the SGM benefits from extended iterations [1], if the model does not converge this benefit is not realized in practice.

The privacy budget is $\varepsilon = 10$ and $\delta = 10^{-7}$ in all experiments, and $q = \frac{1024}{n}$ where n is the number of training examples available, as discussed previously. We chose the remainder of the hyperparameters through an iterative process with the goal to achieve training stability. Normally the goal of the hyperparameter tuning is to maximize validation performance. In our case, with the amount of noise incurred by the requirement of patient privacy, ensuring that the training process leads to consistently decreasing training loss became the priority. For this reason, we do not maintain a strict protocol of avoiding privacy leak through hyperparameter selection. This is a common practice in empirical papers [1, 7, 8], but should not be adopted in applications. Because we do not compare against non-private models, small inflation of the privacy budget ($\varepsilon = 10$) which we can assume is equal for all approaches, does not interfere with our evaluation.

We set the remaining hyperparameters as follows. The maximum gradient norm $C = 10$ in both classification tasks. For melanoma classification, we used the Adam optimizer with learning rate 0.003 and trained for $Kq \approx 30$ epochs. In ECG classification, we were unable to achieve a stable training regime with the Adam optimizer. The SGD optimizer with learning rate 0.25 behaved more predictably, with sufficiently fast convergence to the training data. With $k = 3$ group privacy, we reduced the learning rate to 0.1 to suppress the divergence in the training process caused by the increased noise scale. We train the ECG models for $Kq \approx 100$ epochs. Recall that Kq is roughly equivalent to the number of epochs, because each of the K steps uses approximately a q sized fraction of the dataset.

For the ECG classification task we use the ResNet-SE architecture adapted for differentially private training [19], with the number of channels reduced by a factor of 16 compared to the non-private ECG model originally proposed by Zhao et al. [27]. For melanoma classification, we require a pretrained model, because the melanoma datasets lack the scale and diversity to train generalization capable computer vision models [6]. Almost all high performance computer vision architectures use batch normalization layers, which cause information to leak between samples, thus breaking our privacy guarantees [10]. The NFNet architecture is specifically designed to deliver high classification performance without batch normalization [5]. We use the reduced size "l0" model pretrained on the ImageNet 1K dataset from the timm computer vision library [25]. To finetune the NFNet model to the melanoma classification task, we train the final convolution layer and the fully connected classifier.

In the ECG classification task we measure multilabel classification accuracy using macro AUC, the average of the AUC scores for individual cardiovascular disease labels. We report the median of 15 independent tests with different random seeds for model initialization and sample selection. In the melanoma classification task we report the median AUC of 15 independent tests. In both experiments, because we do not assume any distribution of the error of the AUC measurement, we give the confidence interval determined by sign test [4].

4 Results

We summarize the ECG multilabel classification task in Table 2. The number of records with $k = 3$ is only about 14% higher than with $k = 1$, which makes the noise ratio z unfavorable for the group privacy model. Even with the improved privacy bound, the noise ratio computed using group privacy accounting compensates for an up to three times increase of the training examples, overestimating the privacy loss. Noise scale appears to be the main contributor to the performance loss with the group privacy approach. Sample level training has the highest median performance, but the confidence intervals overlap, so the difference to ULDP training is not significant. To illustrate how the $\varepsilon = 10$ privacy budget impacts performance, a model trained using the same neural network architecture without differential privacy achieved macro AUC of 0.98.

Table 2: PTB-XL multilabel ECG classification, $\varepsilon = 10$

Privacy model	Noise scale z	Training records	Macro AUC	Confidence interval
sample $k = 1$	1.77	15023	0.90	+0.01-0.03
group $k = 3$	4.37	17153	0.79	+0.06-0.06
user (ULDP)	1.67	17418	0.89	+0.01-0.02

Table 3: ISIC 2017 melanoma classification, $\varepsilon = 10$

Privacy model	Noise scale z	Training records	AUC	Confidence interval
sample $k = 1$	2.53	2056	0.61	+0.02-0.03
group $k = 3$	4.10	6167	0.67	+0.01-0.03
group $k = 7$	5.57	12314	0.65	+0.01-0.02
user (ULDP)	0.92	33126	0.70	+0.02-0.04

In Table 3, we report the results of the melanoma classification task. Here, the number of records also affects the performance, both because of the sparsity of positive training examples (Table 1) and because there is less privacy amplification from sampling. Although we could reduce the noise by lowering the sampling rate q, this has unfavorable side effects as discussed in Section 3. Reducing the number of training examples to $k = 1$ yields poorly performing models with these datasets and the group privacy approach becomes more competitive. However, ULDP training requires a much lower noise ratio z for almost triple amount of training records. At the same time, the impact of gradient clipping on ULDP training increases with the number of records per patient, but the trade off appears to be favorable in this setting.

Although clinical relevance is outside the scope of this paper, the predictive performance of the differentially private melanoma models is low. In comparison, the best result in the ISIC 2017 challenge was 0.87 AUC in melanoma classification [9]. Despite the low performance of our NFNet model, the finetuning approach we apply is relevant in practice. For example, finetuning a skin lesion foundation model to a classification task on a private dataset would pose similar requirements for preserving patient privacy.

As noted in Section 3, with ECG models we encountered divergence in the training process. We mitigated this by selecting a learning rate low enough that large fluctuations and long term increases in training loss were no longer observed. For a better understanding of the performance and stability of the training process, we tested the sensitivity to learning rate. The plots in Figure 2 show the median macro AUC of 5 models trained with different random seeds, over SGD learning rates from 0.0625 to 1.0. The drop in performance with higher learning rates is due to most, or all models diverging in the training process. The error bars in the plots indicate increased uncertainty of the outcome close to this divergent regime. Both $k = 1$ sample level and $k = 3$ group privacy approach perform best at learning rates that sit next to the divergent regime. This fragile balance may indicate that the chosen $\varepsilon = 10$ budget is close to the limit where viable models can be trained with this particular dataset.

It is well known that user level privacy requires gradient manipulations that make the training inefficient, but there is little

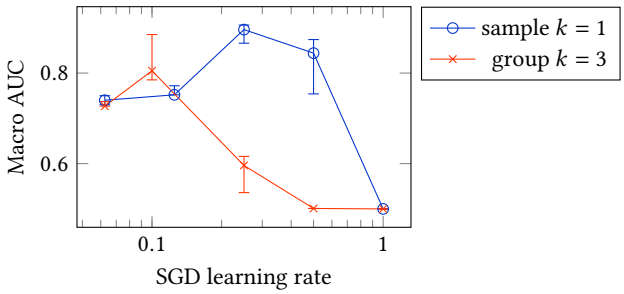


Figure 2: Sensitivity to learning rate of ECG models.

empirical data available to quantify that, so we report our training times here. We used compute nodes with the Nvidia V100 GPU, Xeon Gold 6230 CPU and NVMe storage. The training of ResNet-SE ECG models for estimated 100 epochs took 27-30 minutes with the $k = 1$ and $k = 3$ data samples and 70-80 minutes with user level privacy. The NFNet melanoma models were trained for estimated 30 epochs. The training times in this case were affected by large differences in the size of the training sample. With $k = 1$ sample level privacy the training time was approximately 10 minutes. With user level privacy that uses the full dataset the training took 4 hours. Overall, we observed that our implementation of ULDP is about 3 times slower than the mature implementation of sample level privacy in the Opacus library.

5 Conclusions and Future Work

We summarize our findings w.r.t. the research questions.

RQ1: Our evaluation datasets turned out to be widely apart in the evaluation of sample level privacy. Reducing the training data to one record per patient yielded best median performance with the PTB-XL dataset, and was clearly worst with the ISIC challenge datasets. User level privacy performed consistently better than group privacy, if we do not consider the added computational cost. With our implementation, the training time is increased about 3 times, compared to other approaches.

RQ2: Even with the theoretical advances in determining the group privacy bound, the cost of group privacy is too high in terms of noise, because the group privacy bound estimate is by design imprecise. This is consistent with the findings of earlier papers [7, 8]. The distribution of training examples per patient is the main factor in determining the performance with sample level privacy.

RQ3: User level privacy combined the advantages of a favorable noise ratio and largest training samples, resulting in best or shared best performance in our experiments. Therefore, whenever the distribution of records is such that removing extra records has a non-negligible effect on training data size, substituting sample level privacy with user level privacy is justified.

Our results sketch out a very general picture, because the datasets used in experiments represent two examples of possible distributions, and two modalities – waveform and image. In future work, we intend to address this limitation and fill the gaps with more experiments. We will include more and larger ECG datasets and are evaluating other modalities, like text. The main obstacle is that patient identifiers are rarely included in public medical datasets.

In conclusion, user level differential privacy is a viable approach in preserving patient privacy. Our early results indicate that sample level and group privacy approaches come with a cost to utility, except the edge case where most patients have one data record to begin with. The software implementation we developed will hopefully make further research into ULDP more accessible outside the federated learning setting.

Acknowledgments

This work was supported by Research Council of Finland (grant 358868). Tapio Pahikkala acknowledges the research environment provided by ELLIS Institute Finland. We thank the CSC – IT Center for Science in Finland for providing the computing resources for the experiments.

References

[1] Martín Abadi et al. 2016. Deep Learning with Differential Privacy. In *Proc. SIGSAC*. ACM, New York, NY, USA, 308–318.

[2] Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. 2020. Hypothesis Testing Interpretations and Rényi Differential Privacy. In *Proc. AISTATS*. PMLR, 2496–2506.

[3] Daniel J Beutel et al. 2020. *Flower: A Friendly Federated Learning Research Framework.* arXiv:2007.14390

[4] James Vandiver Bradley. 1960. *Distribution-free Statistical Tests*. WADD technical report, Vol. 60. U. S. Air Force, Washington, DC, USA.

[5] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. 2021. High-performance Large-scale Image Recognition Without Normalization. In *Proc. ICML*. PMLR, 1059–1071.

[6] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. 2022. Analysis of the ISIC Image Datasets: Usage, Benchmarks and Recommendations. *Medical Image Analysis* 75 (2022), 102305.

[7] Zachary Charles et al. 2024. *Fine-Tuning Large Language Models with User-Level Differential Privacy.* arXiv:2407.07737

[8] Lynn Chua et al. 2024. *Mind the Privacy Unit! User-level Differential Privacy for Language Model Fine-tuning.* arXiv:2406.14322

[9] Noel CF Codella et al. 2018. Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). In *Proc. ISBI*. IEEE, New York, NY, USA, 168–172.

[10] Ali Davody, David Ifeoluwa Adelani, Thomas Kleinbauer, and Dietrich Klakow. 2020. *On the Effect of Normalization Layers on Differentially Private Training of Deep Neural Networks.* arXiv:2006.10919

[11] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 3-4 (2014), 211–407.

[12] Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Raghu Meka, and Chiyuan Zhang. 2023. User-level Differential Privacy with Few Examples per User. In *Proc. NeurIPS*. Curran Associates, New York, NY, USA, 19263–19290.

[13] Ary L Goldberger et al. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 101, 23 (2000), e215–e220.

[14] Alexey Kurakin, Shuang Song, Steve Chien, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. 2022. *Toward Training at Imagenet Scale with Differential Privacy.* arXiv:2201.12328

[15] Daniel Levy et al. 2021. Learning with User-level Privacy. In *Proc. NeurIPS*. Curran Associates, New York, NY, USA, 12466–12479.

[16] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning Differentially Private Recurrent Language Models. In *Proc. ICLR*. PMLR, 14 pages.

[17] Ilya Mironov. 2017. Rényi Differential Privacy. In *Proc. CSF*. IEEE, New York, NY, USA, 263–275.

[18] Ilya Mironov, Kunal Talwar, and Li Zhang. 2019. *Rényi Differential Privacy of the Sampled Gaussian Mechanism.* arXiv:1908.10530

[19] Zoher Orabe, Antti Vasankari, Tapio Pahikkala, Matti Kaisti, and Antti Airola. 2026. Securing Deep Learning Models with Differential Privacy for Cardiovascular Disease Prediction. *Biomed. Signal Process. Control.* 112 (2026), 108502.

[20] Ossi Räisä, Joonas Jälkö, and Antti Honkela. 2024. Subsampling is not Magic: Why Large Batch Sizes Work for Differentially Private Stochastic Optimisation. In *Proc. ICML*. PMLR, 41959–41981.

[21] Veronica Rotemberg et al. 2021. A Patient-centric Dataset of Images and Metadata for Identifying Melanomas Using Clinical Context. *Scientific Data* 8, 1 (2021), 34. https://doi.org/10.1038/s41597-021-00815-z

[22] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *Proc. SP*. IEEE, New York, NY, USA, 3–18.

[23] Patrick Wagner et al. 2020. PTB-XL, a Large Publicly Available Electrocardiography Dataset. *Scientific Data* 7, 1 (2020), 154.

[24] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. 2022. *PTB-XL, a Large Publicly Available Electrocardiography Dataset.* https://doi.org/10.13026/kfzx-aw45 Version 1.0.3.

[25] Ross Wightman. 2019. PyTorch Image Models. https://github.com/huggingface/pytorch-image-models.

[26] Zheng Xu et al. 2023. Learning to Generate Image Embeddings with User-Level Differential Privacy. In *Proc. CVPR*. IEEE, New York, NY, USA, 7969–7980.

[27] Zhibin Zhao et al. 2020. Adaptive Lead Weighted ResNet Trained With Different Duration Signals for Classifying 12-lead ECGs. In *Proc. CinC*. IEEE, New York, NY, USA, 1–4.

[28] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep Leakage from Gradients. In *Proc. NeurIPS*. Curran Associates, New York, NY, USA, 14747–14756.