At What Price? Exploring the Potential and Challenges of Differentially Private Machine Learning for Healthcare

Aycan Aslan
University of Goettingen
aycan.aslan@uni-goettingen.de

Tizian Matschak
University of Goettingen
tizian.matschak@uni-goettingen.de

Maike Greve
University of Goettingen
maike.greve@uni-goettingen.de

Simon Trang
University of Goettingen
strang@uni-goettingen.de

Lutz M. Kolbe
University of Goettingen
lkolbe@uni-goettingen.de

Abstract

The increased generation of data has become one of the main drivers of technological innovation in healthcare. This applies in particular to the adoption of Machine Learning models that are used to generate value from the growing available healthcare data. However, the increased processing of sensitive healthcare data comes with challenges in terms of data privacy. Differential privacy, the method of adding randomness to the data to increase privacy, has gained popularity in the last few years as a possible solution. However, while the addition of randomness increases privacy, it also reduces overall model performance, generating a privacy-utility trade-off. Examining this trade-off, we contribute to the literature by providing an empirical paper that experimentally evaluates two prominent and innovative methods of differentially private Machine Learning on medical image and text data to deepen the understanding of the existing potential and challenges of such methods for the healthcare domain.

Keywords: Differential privacy, PATE framework, Differentially private stochastic gradient descent.

1. Introduction

The digitization of healthcare data and technological advancements in computer processing and data storage has enabled the development of advanced algorithmic techniques such as Artificial Intelligence, especially in the form of Machine Learning (ML). Beyond the increase in volume, velocity, and variety of available healthcare data, an additional driver of ML applications is financial pressures on the healthcare industry globally, with increasing demands due to a growing and aging population (Stanfill & Marc, 2019). Against this background, the use of ML is gaining popularity not only in research but also in medical practices. For instance, to realize the potential offered by 'precision medicine', a tailored medical treatment of patients based on individual characteristics (Ginsburg & Phillips, 2018), data from a wide range of data sources, such as Electronic Health Records (EHR) or genomics data, must be collected and subsequently analyzed (Ginsburg & Phillips, 2018). Here, the high speeds at which ML models perform make them a suitable tool to efficiently take advantage of a growing, diverse set of healthcare data (Jiang et al., 2017). For example, studies show that ML models can be used to analyze gene expression data and DNA data to predict the treatment response of patients with rheumatoid arthritis (Tao et al., 2021). Other examples of the use of ML show its potential for an automated system of disease classification of medical images (Mehta & Pandit, 2018) and fraud detection within the healthcare system (Matschak et al., 2022).

While the presented examples illustrate the potential of ML for healthcare, one cannot ignore the associated data privacy issues. These issues primarily originate from the high demand ML places on vast amount of data to train on, but there are also privacy issues arising from the inherent nature of ML (Abouelmehdi et al., 2018). These include the privacy of model weights of ML models or the possible data memorization of individual data points during the training of ML models (Kaissis et al., 2020). Studies have shown that, for example, model inversion attacks can be performed to recover recognizable images of people's faces from ML models that used them (Fredrikson et al., 2015). These data privacy issues apply in particular to health data since it represents sensitive and personally identifiable information. To ensure privacy for sensitive health data while capitalizing on the presented potentials by ML, the use of differential privacy (DP) is gaining popularity (Aslan et al., 2022). DP describes the systematic

URI: https://hdl.handle.net/10125/103034
978-0-9981331-6-4
(CC BY-NC-ND 4.0)

modification of data to reduce the potential of retrieving information about individuals (Hu et al., 2019) and has quickly become a well-accepted mathematical quantification of privacy (Wang et al., 2022). While DP itself is a definition of privacy, the implementation of DP into data analysis generates privacy-enhanced data analysis techniques that have been adopted by companies such as Google (Erlingsson et al., 2014) and Microsoft (Ding et al., 2017).

However, despite the socio-technical implications of privacy-enhanced ML, there is little work on DP in Systems Science literature. Prior research has proved that enhanced privacy significantly influences how individuals perceive and interact with information systems (Malhotra et al., 2004). Hence, a sufficient way of implementing privacy protection through DP can influence how individuals and organizations interact with ML models. Nevertheless, despite its socio-technical nature, work on DP in Systems Science literature is still limited. Therefore, the objective of this work is to showcase the potential of differentially private ML with medical text and image data to understand the trade-offs of differentially private ML with respect to added privacy and model performance. This paper aims to contribute to this understanding by answering the following research questions:

RQ1: *What is the nuanced trade-off between privacy and model performance for medical image and text data?*

RQ2: *What future research directions can be derived from such a deepened understanding of this trade-off?*

Our work follows a threefold procedural approach to answer these questions. First, we introduce DP and describe its privacy properties. Second, we analyze the potentials of differentially private ML with an experimental evaluation utilizing the PATE framework and the differentially private stochastic gradient descent. Third, based on the understanding generated from the experimental evaluation, we identify current research gaps and show how System Science scholars can contribute to filling the respective gap to further advance the socio-technical research on DP.

2. Background and Related Work

In this section, we will clarify the theoretical background by first giving an overview of traditional methods to ensure data privacy for healthcare and touch on their shortcomings. Building on this, we will present the definition and characteristics of DP and present two frameworks to implement DP.

2.1. Privacy for Healthcare Data

Data privacy is a sub-field of data management whose goal is to provide value from sensitive datasets without compromising the privacy of the individuals' data records within these databases (Kifer & Machanavajjhala, 2011). One example of such sensitive datasets is healthcare data of any kind as soon as it contains personal attributes. In today's healthcare systems, healthcare data is collected along the entire patient journey, for example via EHRs. These records do contain not only sensitive *attributes* such as medical data in the narrow sense (e.g., blood pressure) but also several types of personal patient data. The first type of attribute is an *explicit identifier*, which are labels that can be used to directly identify an individual, such as name or phone number. The second type is a *quasi-identifier*, which does not explicitly reveal identities but may be linked to external data sources to identify an individual. Consequently, using sensitive healthcare data to develop and deploy ML applications has raised substantial legal, ethical, and regulatory challenges (Stanfill & Marc, 2019). Against this background, there are various regulatory frameworks (e.g., the Health Insurance Portability and Accountability Act in the U.S.), all of which address patient privacy as a key concern. In practice, the privacy disclosure risk for patients is twofold: Firstly, identity disclosure, which occurs when it is possible to match a record in an anonymized dataset to an actual individual and, secondly, attribute disclosure, which occurs when adversaries can restore sensitive data of an individual record (Duncan & Lambert, 1989).

Facing these risks and the associated legal regulations on the one hand, and the increased need for the adoption of ML-based applications by healthcare providers at the point of care (Noorbakhsh-sabet et al., 2020) on the other, a research body regarding privacy-preserving ML has emerged. Researchers proposed privacy models such as *k*-anonymity (Sweeney, 2002), *l*-diversity (Machanavajjhala et al., 2007), and *t*-closeness (N. Li et al., 2007) to formalize privacy protection requirements. One of the most widespread methods is *k*-anonymity (Sweeney, 2002), which requires each individual record in a dataset to be indistinguishable from at least $k - 1$ other records in terms of quasi-identifiers. Thus, the *k*-anonymity focuses only on re-identification risk and does not consider attribute disclosure risk. To overcome this drawback, *l*-diversity has been proposed (Machanavajjhala et al., 2007). *L*-diversity addresses attribute disclosure risk by building on *k*-anonymity

and requires that sensitive attributes include at least *l* represented values in the *k*-anonymized data. Besides that, another privacy principle, called *t*-closeness, addresses the attribute disclosure risk by further considering the overall distribution of sensible attributes values (N. Li et al., 2007). *T*-closeness necessitates that the distance between the distributions of sensible attributes values in a group and the overall distribution of these values cannot be larger than a defined threshold *t*.

Since medical text documents typically have many unstructured and no pre-defined sensible attributes, *t*-closeness and *l*-diversity do not fit this use case because these principles typically assume either a single sensible attribute or several pre-defined sensible attributes (X. Li & Qin, 2018). Moreover, *k*-anonymity, *l*-diversity, and *t*-closeness approaches all depend on assumptions about the adversary's additional information about individual targets. If the assumptions do not hold, these approaches may not work well (Dwork, 2011). To overcome these limitations, Dwork & Roth (2013) introduces the notion of DP.

2.2. Differential Privacy

DP was proposed by Dwork & Roth (2013) and describes a promise between the data holder and data subject: The data holder will not be affected adversely by allowing their data to be used for statistical analysis (Dwork & Roth, 2013). This promise between the data holder and the data subject results in a privacy model, which aims at achieving maximum privacy by minimizing the risk of individual record identification by bounding the maximum amount of information that can be learned about any one individual (Dwork & Roth, 2013).

Formally, DP is defined as follows: A randomized algorithm M satisfies (ε, δ) – differential privacy if for all neighboring datasets *D1* and *D2* and all possible outputs *S*:

$$\Pr [M(D1) \in S] \leq \exp(\varepsilon) \Pr [M(D2) \in S] + \delta]$$

where *M* denotes the algorithm which randomizes by adding noise and datasets *D1* and *D2* are considered neighboring if they only differ in the data of a single individual (Dwork & Roth, 2013). Based on this formula, the difference in *M´s* output between the neighboring datasets *D1* and *D2* is identical but only differs in the exponential of ε. Consequently, the maximum change between the probability distributions of *D1* and *D2* is measured by the parameter ε. That means that the presence of the data point of any single individual in the datasets is noticeable from the output up to (the exponential factor of) ε. Intuitively, ε can be seen as a parameter to tune the difference between the datasets *D1* and *D2*: Low levels of ε require the algorithm M to provide very similar outputs with regard to the probability distributions so a potential adversary couldn't determine if a single individuals' data was present in the input dataset, hence leading to higher levels of privacy described. On the contrary, high values of ε allow less similarity between the output of the probability distributions and therefore provide less privacy. Additional to ε, the parameter δ can be seen as a security parameter. It bounds the probability of the privacy guarantee not holding, meaning that the chances of something going unexpectedly wrong is limited.

This presented notion of DP has some critical implications in the context of ML. Under perfect privacy, it would mean that training a model on a dataset should return the same model even if we remove any (one) person from the training dataset. While perfect privacy is not desired in most cases, this notion of privacy aims to create the most accurate model with the highest amount of privacy possible. Hence, the goal is to maximize both the utility and privacy of the model simultaneously.

The described formal definition of DP does not create privacy by itself but instead presents the constraints that a researcher can analyze to understand whether the query is leaking private information and, more importantly, to what extent. We will use this notion of data privacy since it leads to a quantifiable privacy budget and therefore allows us to understand trade-offs between privacy and utility.

2.2.1. PATE framework. One of the frameworks proposed to implement the stated boundaries of DP is the Private Aggregation of Teacher Ensembles (PATE), which will be presented and explained in this sub-section.

PATE is a framework based on knowledge aggregation of an ensemble model and knowledge transfer (Papernot et al., 2017). The intuition behind PATE is that if two different classifiers trained on disjoint datasets agree on how to classify a new input example, that classification decision does not reveal information about any single training example. Since the model with and without the given example reaches the same conclusion, the classification decision would have been made with or without the given training example. When the classifiers disagree, publishing any of the two decisions may leak private information contained in the respective training data. Therefore, noise is added to ensure the privacy guarantees stated by DP.

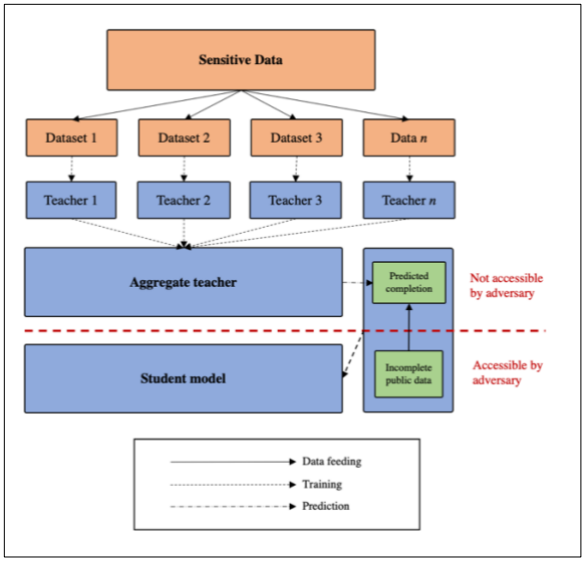

Figure 1. Architecture of the PATE framework.

The architecture of PATE can be seen in Figure 1 and consists of three key parts: 1) an ensemble of *n* teacher models which are trained on labeled private data, 2) an aggregation mechanism that adds noise, and 3) a student model which consists of unlabeled public data (Papernot et al., 2017). The goal is to utilize the private dataset in a differentially-private manner, to generate the missing labels for the public dataset. First, the private dataset which contains sensitive data is portioned into n subsets of data. Next, each of these models, also called teachers, is trained on the subsets of data using an ML model. The models are then used to make predictions on each input data of the public dataset. Hereby, while aggregating the predictions of each teacher, noise is added to realize the privacy guarantees defined by DP. The number of teachers who voted for a class is counted, and the count is then perturbed by adding random noise. Then only the noised up highest votes are outputted. When two output classes receive a close number of votes, which might pose a privacy risk, the added noise will ensure that the output class with the most counts will be chosen randomly. If most teachers agree on the same class, the added randomness will not change the fact that the respective class received the most votes. This aggregation mechanism is essential since it enables knowledge transfer from the private dataset (teachers) to the public dataset (students) in a differentially-private manner. Finally, the noisy labels with the highest votes are used to train the unlabeled data of the student to create a differentially-private student model. At this point, teacher models must be discarded since their outputted labels may leak information about the private datasets. Now, the student model is the only classifier used for inference since it does not pose privacy risks.

2.2.2. Differentially-private stochastic gradient descent. Another form to implement the stated privacy boundaries of DP is differentially private stochastic gradient descent (DPSGD) proposed by (Abadi et al., 2016).

The privacy guarantees of DPSGD build on the characteristics of standard stochastic gradient descent (SGD). SGD is a widely employed method to train ML models by optimizing a given objective function (Wang et al., 2022). It is an iterative algorithm that replaces the true gradient with a randomized gradient estimated from a random subset of available data (Wang et al., 2022). The learning phase of the model is depicted in the top part of Figure 2.

In the presented context of SGD, DPSGD modifies the minibatch stochastic optimization process to limit the privacy loss per gradient update during the stochastic optimization (Abadi et al., 2016). The intuition is that if the model's training itself is differentially-private, so are the resulting model outcomes.

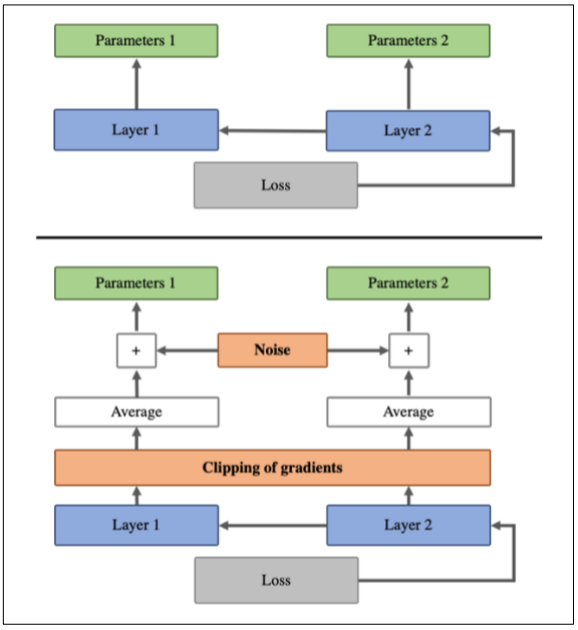

Figure 2. Comparison of standard SGD (top) and DPSGD (bottom).

The practical implementation of DPSGD can be seen at the bottom of Figure 2. During the depicted backward pass, privacy risks might be embedded in the gradients since the contribution of individual data points might be too large. Hence, rather than updating with raw gradients as per standard SGD, the gradients

are 'clipped,' meaning that they have a maximum defined gradient norm. This clipping of gradients limits the amount of information that is learned from any given example. Next, the gradients are aggregated, and noise is added. These noisy aggregated gradients are used to update the model parameters.

3. Experiments

This section will clarify the context of the experiments conducted. We introduce the datasets used, describe the specific experimental set-up in terms of steps conducted and metrics used to evaluate the experiment, and finally, present the results of the experiments.

3.1. Datasets

For the evaluation of the PATE framework, we utilize lung X-Ray images provided by (Kermany et al., 2018). The dataset consists of a total of 5856 images in two categories: 'Normal' and 'Pneumonia.' All images were screened for quality control, with the subsequent grading of two expert physicians afterward (Kermany et al., 2018).

For the evaluation of DPSGD, we utilize a medical text dataset provided by the University of California Irvine via their Machine Learning Repository with 918 observations in total (Blake et al., 1998). The dataset consists of 11 medical attributes (e.g., 'Chest pain type' or 'Resting blood pressure') that describe the output class 'heart disease'. The output class 'HeartDisease' is binary: 1 (heart disease) and 0 (normal).

3.2. Experimental Set-up and Metrics

In this sub-section we will elaborate on the data processing, the conducted steps in the experiment, and the used metrics for the PATE and DPSGD analysis set-up respectively. The goal for both experimental set-ups is to test a non-private baseline model against private models with varying privacy security. Here, based on the work by (Sun et al., 2019) we divide the level of privacy guaranteed into three scales: *Small* level of privacy (ε = 8), *medium* level of privacy (ε = 2), and *high* level of privacy (ε = 0.5). Here, ε represents the privacy parameter of DP and reflects the noise added to the data (Recall that lower levels of ε guarantee higher levels of privacy). The metrics to compare the baseline models to the private models are adopted from Guo et al. (2020), Sun et al. (2019), and Zhang et al. (2021) and consist of: Accuracy, recall, precision, and F1-score. All values were obtained based on applying a 20/80 random train-test split to the datasets. The resulting sub-sets were then class-balanced by using a class-balanced loss function (for the PATE analysis, as the X-ray images were unbalanced, between the two classes 'Normal' and 'Pneumonia'). In addition, for each analysis data transformation steps (such as resizing, cropping, random rotation, and channel-wise data normalization (for the PATE analysis) and encoding and imputing (for the DPSGD analysis)) were performed. Exemplary images after the transformation of the X-Rays for the PATE analysis can be seen in Figure 3.

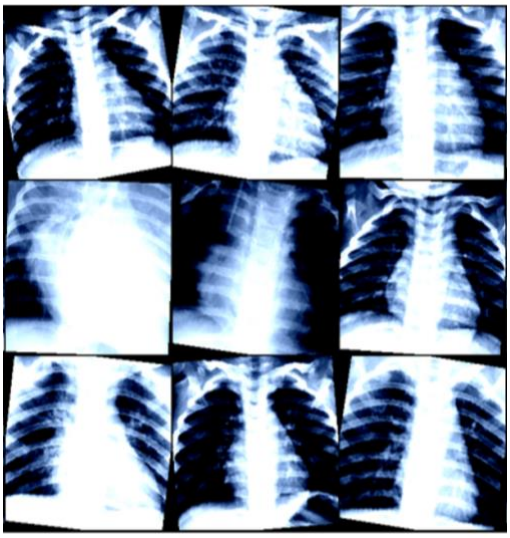

Figure 3. Exemplary images of the X-Rays for the PATE analysis (after transformation).

PATE analysis. For the implementation of the PATE framework, first the images are loaded. Next, the dataset is split equally among the three teachers into training and validation sets. Each trainer is trained using a pre-trained deep learning model (VGG-16 model) (Simonyan & Zisserman, 2015). After the teacher models have been trained, they are able to generate labels. Each of the three teacher models will generate one label for each image in the dataset. To aggregate the labels in a differentially private manner, the predicted labels are noised up using random variables that are Laplacian distributed. Here, as stated, the ε values tested are 8, 2, and 0.5. Building the non-private baseline model is more straightforward. First, the images are loaded and the model is trained based on the learning algorithm defined, without any measures in terms of privacy. The non-private baseline model is then tested.

DPSGD analysis. In comparison to the standard classifier, for the DPSGD models, we make some changes using the *PyTorch Opacus* framework

(Opacus AI, 2022). The Opacus framework is an Open-Source framework that supports PyTorch machine learning models and allows the implementation of DP easily, with only minimal modifications to the original model. The Opacus framework enables us to create a 'privacy engine' that can be attached to our SGD optimizer, making it differentially private. The privacy engine allows us to define a maximum norm of the per-sample gradients. Any gradient with the norm higher than this will be clipped to this value. Finally, both the model with standard SGD and the DPSGD with the privacy engine are evaluated using the test data.

3.3. Results of the Experiments

The results of the PATE analysis can be seen in Table 1. The first column lists the non-private baseline model and the varying values of ε. The top row depicts the metrics used to assess the performance of the given model, namely model accuracy, model recall, model precision, and F1-score (in percentage terms). Firstly, our experiments demonstrate that the non-private baseline model is a performant classifier for detecting pneumonia on X-Ray images. The baseline model has an accuracy of 89.9%, recall of 88.2%, precision of 93.1%, and F1-score of 90.5%. As stated, we tested three privacy security levels against this baseline model: Small level of privacy ($\varepsilon = 8$), medium level of privacy ($\varepsilon = 2$), and high level of privacy ($\varepsilon = 0.5$). We note that small privacy protection with an ε level of 8 still leads to a model that reliably predicts our target labels with an accuracy of 87.5%, recall of 99.0%, precision of 80.5%, and F1-score of 88.8%. However, for medium-level privacy protection, we note a strong drop in model performance. Medium privacy protection with level an ε level of 2 leads to a model with an accuracy of 75.5%, recall of 99.4%, precision of 67.1%, and F1-score of 80.3%. Consequently, the model with the highest privacy protection and ε of 0.5 leads to the weakest-performing model. This model has an accuracy of 61.1%, recall of 96.2%, precision of 56.5%, and F1-score of 71.2%.

Table 1. Results for the PATE analysis.

Privacy level	Accuracy (%)	Recall (%)	Precision (%)	F-1 score (%)
Baseline model	89.9	88.2	93.1	90.5
Small ($\varepsilon = 8$)	87.5	99.0	80.5	88.8
Medium ($\varepsilon = 2$)	75.5	99.4	67.1	80.3
High ($\varepsilon = 0.5$)	61.1	96.2	56.5	71.2

The results of the DPSGD analysis can be seen in Table 2. Here, the first column lists the non-private baseline model with the standard SGD optimizer and the private models with the differentially private SGD.

Table 2. Results for the DPSGD analysis.

Privacy level	Accuracy (%)	Recall (%)	Precision (%)	F-1 score (%)
Baseline model	89.9	88.2	93.1	90.5
Small ($\varepsilon = 8$)	85.5	84.2	88.9	86.5
Medium ($\varepsilon = 2$)	73.9	77.6	75.6	76.7
High ($\varepsilon = 0.5$)	68.8	68.4	73.2	70.7

As depicted in Table 2 the baseline shows an overall good performance in predicting heart disease. The accuracy of the baseline model is 89.9%, the recall is 88.2%, the precision is 93.1%, and F1-score is 90.5%. Hence, the baseline model yields a potent model to be compared to. We note that small privacy protection with an ε level of 8 still leads to a model that performs well with an accuracy of 85.5%, recall of 84.2%, precision of 88.9%, and F1-score of 86.5%. However, for medium-level privacy protection we note a strong drop in model performance. Medium privacy protection with level an ε level of 2 leads to a model with an accuracy of 73.9%, recall of 77.6%, precision of 75.6%, and F1-score of 76.7%. In line with these results, the model with the highest privacy protection and ε of 0.5 leads to the worst performing model along all metrics. This model has an accuracy of 68.8%, recall of 68.4%, precision of 73.2%, and F1-score of 70.7%.

4. Discussion and Contributions

To sum up the results of the experiments conducted, we see that the added privacy through DP comes at the cost of model performance. For both the PATE and DPSGD analysis, our experiments have shown that with higher levels of privacy (lower ε levels), the model performance of the classifier is decreasing. While this is not too surprising considering the mathematical dependencies for the noise-adding mechanisms, we were able to show a nuanced understanding of how different levels of ε influence the classifier's overall performance for the given image and text datasets. Nevertheless, the experiments also prove that medium levels of privacy (e.g., $\varepsilon = 2$) still allow for building potent ML models. Since the noise addition can be fine-tuned through the level of ε, for the given dataset, we are able to manage how performant the model remains. Hence, we can conclude that the privacy addition through DP negatively influences the model performance but still leaves space for performant classifiers that can be utilized for privacy-enhanced data analysis.

4.1. Contributions to Literature

Our study contributes to the literature in several ways. First, this study challenges the known problem of personal data protection in healthcare-based ML models and builds on the existing knowledge base by contextualizing privacy-preserving ML methods using two concrete real-world examples from the healthcare domain. Our study also demonstrated the potential of privacy-preserving ML by utilizing both medical image and text data, showing the potential of such methods for a variety of existing data types in the healthcare domain.

Second, this study contributes to the existing literature by demonstrating the trade-off between data privacy and ML model performance. Building on the empirical evidence provided by the given image and text datasets, we add to the understanding of how different levels of ε influence the ML classifier's overall performance. We, therefore, quantify the difference in the model performance for three chosen levels of privacy (high, medium, and small) (Sun et al., 2019) and granularly show the impact of multiple relevant metrics (model accuracy, model recall, model precision, and F1-score) for the implementation of such methods in the healthcare context.

Third, this study can serve as a foundation that paves the way forward toward flexible privacy strategies for healthcare data. As prior research has shown, the perceived level of privacy risk is highly dependent on the given context and circumstances (Nissenbaum, 2004). In the context of this discussion in privacy research, our study demonstrates that instantiating privacy through DP offers an easy way to adapt the level of privacy offered. DP allows for bounding the maximum level of information leakage through the parameter ε which can be easily changed for different circumstances. In the healthcare context, this might be changes based on the preferences of the patient with regard to the given medical condition or the existing level of trust between the patient and the healthcare institution.

Based on that, we prove not only the applicability of privacy-preserving methods within the healthcare domain but also provide a blueprint for further research (see sub-section 4.3).

4.2. Contributions to Practice

Additional to the contributions to literature, our study has important implications for practice. First, our study provides stakeholders of the healthcare domain with two applicable methods to implement privacy-enhanced ML. This is especially important considering the growing need for sufficient privacy when handling sensitive health data. The shown methods, PATE and DPSGD, can be implemented to ensure a level of privacy that is mathematically bounded due to the instantiation of DP. The approaches also show that despite the general need for ML approaches for bigger datasets, it is possible to implement ML that is private-by-design. Hence, practitioners in the healthcare domain can benefit from the offered potential of ML, while respecting and protecting the data privacy of the patients at the same time. Moreover, our study clearly indicates the privacy-utility trade-off of privacy-enhanced ML, showing practitioners at which 'cost' the added privacy comes. As both the PATE and DPSGD analysis show, higher levels of privacy result in less potent ML models. Quantifying this trade-off offers healthcare practitioners a tangible view of privacy-enhanced ML and sets realistic expectations with regard to current challenges in implementing additional levels of privacy. Based on this understanding, healthcare practitioners can adapt their overall ML strategy, for example, with regard to the size of the available datasets. As they are sensitized to privacy-utility trade-off, they might attempt at increasing the dataset to train on, as a potential countermeasure for the loss of model performance. Accordingly, such an understanding of the privacy-utility trade-off can be seen as a building block for a widespread implementation of privacy-enhanced ML in practice.

4.3. Limitations and Future Research

Besides our stated contribution to literature and practice, we must note important limitations to our work. First, the size of the available dataset limits the use of DP. If the size of the available training dataset is too small, the addition of noise will too strongly influence the statistical analysis and make the data useless to generate any knowledge. Hence, the potentials of DP presented in this work require settings in which large datasets are available. Moreover, as shown in our experiments, the addition of DP to ML increases computational complexity. Thereby, privacy-preserving ML also requires additional computation time. The addition of privacy through DP does not only come at the cost of model performance but also added computation time which must be taken into consideration.

As DP is still an emerging field, based on the findings of our study, we identify three key areas in which future research is needed, to deepen the knowledge of the socio-technical perspectives on DP. These are summarized in Table 3. First, as shown in sub-section 2.2, DP provides a complex definition of

privacy. It is a mathematically rigorous definition, computationally efficient, and handy to work with for researchers since it creates a mathematically bounded privacy budget. However, there is little work on the user perception of DP. These stated advantages, namely the mathematical nature of DP, can make it challenging to communicate its characteristics to users. Here, adapting the behavioral science lens of Systems science research can help answer questions regarding the perception and ultimately adaption of DP. These include that DP provides nuanced privacy protection that is not binary (private/not private). Hence, the question arises of how organizations, such as healthcare facilities can communicate to their patients that by design, DP will lead to some privacy leakage. In our experiment, we noted the impact of changes in the level of ε. Consequently, further Systems science research should explore how these technical nuances can be addressed to users in an acceptable way. Second, our findings of both the PATE and DPSGD analysis show that the higher level of privacy provided by DP comes at the price of model performance, leading to a privacy-utility trade-off. However, while computer science scholars work on the technical characteristics of these trade-offs, there is little work on how organizations assess these trade-offs. Future Systems science research can contribute to understanding the organizational perspective on these trade-offs. These include how organizations such as healthcare facilities assess and value different interests (e.g., privacy vs. model performance) against each other depending on the context at hand. Moreover, Systems science research could analyze organizational processes to understand how organizations distribute the accountability of decisions about these trade-offs. Third, our findings demonstrate that DP generates a flexible privacy budget that can be changed, with changes to the technical nuances of the protection model (e.g., level of ε). Hence, DP opens the possibility to adapt the privacy budget based on the context and circumstantial factors. For instance, if, for a given context, two parties do not trust each other, a high level of privacy (lower levels of ε) can be enforced, which will be at the cost of model performance. However, if the trust between the parties is high, for example, due to past positive interactions, it might be sufficient to enforce higher levels of ε, compared to the low trust setting. Thereby, the model remains more performant, and the existing trust between the interacting parties is not 'wasted.' With improved designs, it is possible to create tailored and flexible privacy strategies, which both enforce the needed level of privacy and do not sacrifice model performance when a certain trust level is given. Future Systems Science research can advance the understanding of these flexible privacy strategies by investigating user perception. They can also investigate whether patients appreciate such flexible solutions, for example in the context of varying medical conditions which differ in their perceived sensitivity. Additionally, studies could analyze which medical context is perceived by patients as a low and high trust context, with the respective implemented privacy levels.

Table 3. Future Research Opportunities.

Research gap	Related Research	Potential Contribution
User-understanding of differential privacy	Cummings et al. (2021), Xiong et al. (2020)	- How can healthcare organizations efficiently communicate that DP protection is not binary? - How can technical nuances of differential privacy (such as ε) be communicated to patients?
Assessment of privacy-utility trade-offs by organizations	Alvim et al. (2012), Pannekoek & Spigler (2021)	- How do healthcare organizations weigh model performance against privacy? - How do healthcare organizations distribute the accountability for making context-based decisions on these trade-offs?
Patient perception of flexible privacy budgets	Ebadi et al. (2015), Lee & Clifton (2011)	- Do patients appreciate context-based privacy budgets? - Do patients utilize flexible privacy-budgets based on the trust setting?

5. Conclusion

Due to the increased generation and analysis of sensitive data, we consider research on sufficient data privacy measures as very important and necessary. As such, this work aims to show the potential, but also the challenges present when utilizing DP in the form of the PATE and DPSGD frameworks. With the increased pressure of regulatory bodies on data privacy, we can expect that the use of DP will increase in the upcoming years. The findings of our study can be used to carefully consider the use of DP for protecting sensitive healthcare data.

In conclusion, we are confident that this paper provides a good understanding of the potential and challenges of DP. We aspire to this understanding to stimulate future research and motivate scholars to engage in this emerging field to facilitate a more privacy-oriented use of healthcare data.

6. References

Abadi, M., McMahan, H. B., Chu, A., Mironov, I., Zhang, L., Goodfellow, I., & Talwar, K. (2016). Deep learning with differential privacy. *Proceedings of the ACM Conference on Computer and Communications Security*, 308–318.

Abouelmehdi, K., Beni-Hessane, A., & Khaloufi, H. (2018). Big healthcare data: preserving security and privacy. *Journal of Big Data*, *5*(1), 1–18.

Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., Degano, P., & Palamidessi, C. (2012). Differential privacy: On the trade-off between utility and information leakage. *Lecture Notes in Computer Science*, *7140*, 39–54.

Aslan, A., Greve, M., Diesterhöft, T. O., & Kolbe, L. M. (2022). Can Our Health Data Stay Private? A Review and Future Directions for IS Research on Privacy-Preserving AI in Healthcare. *International Conference on Wirtschaftsinformatik (WI)*.

Blake, C., Keogh, E., & Merz, C. J. (1998). UCI repository of machine learning databases. *Department of Information and Computer Science, University of California, Irvine, CA*.

Cummings, R., Kaptchuk, G., & Redmiles, E. M. (2021). *"I need a better description''': An Investigation Into User Expectations For Differential Privacy."* 1–26.

Ding, B., Kulkarni, J., & Yekhanin, S. (2017). Collecting telemetry data privately. *Advances in Neural Information Processing Systems* (Nips), 3572–3581.

Duncan, G., & Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, *7*(2), 207–217.

Dwork, C. (2011). A firm foundation for private data analysis. *Communications of the ACM*, *54*(1), 86–95.

Dwork, C., & Roth, A. (2013). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3–4), 211–487.

Ebadi, H., Sands, D., & Schneider, G. (2015). Differential Privacy: Now it s Getting Personal. *ACM SIGPLAN Notices*, *50*(1), 69–81.

Erlingsson, Ú., Pihur, V., & Korolova, A. (2014). RAPPOR: Randomized aggregatable privacy-preserving ordinal response. *Proceedings of the ACM Conference on Computer and Communications Security*, 1054–1067.

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. *Proceedings of the ACM Conference on Computer and Communications Security*, 1322–1333.

Ginsburg, G. S., & Phillips, K. A. (2018). Precision medicine: From science to value. *Health Affairs*, *37*(5), 694–701.

Guo, Y., Liu, F., Cai, Z., Chen, L., & Xiao, N. (2020). FEEL: A Federated Edge Learning System for Efficient and Privacy-Preserving Mobile Healthcare. *ACM International Conference Proceeding Series*, *19*.

Hu, Y., Ge, L., Zhang, G., & Qin, D. (2019). Research on differential privacy for medical health big data processing. *20th International Conference on Parallel and Distributed Computing*, 140–145. https://doi.org/10.1109/PDCAT46702.2019.00036

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243.

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

Kermany, D. S., Goldbaum, M., & Cai, W. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, *172*(5), 1122–1131.

Kifer, D., & Machanavajjhala, A. (2011). No free lunch in data privacy. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 193–204.

Lee, J., & Clifton, C. (2011). *How Much Is Enough? Choosing Epsilon for Differential Privacy*. 325–340.

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. *IEEE 23rd International Conference on Data Engineering*, *2*, 106–115.

Li, X., & Qin, J. (2018). Protecting Privacy When Releasing Search Results from Medical Document Data. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 3770–3778.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). ℓ-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1).

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, *15*(4), 336–355.

Matschak, T., Prinz, C., Rampold, F., & Trang, S. (2022). Show Me Your Claims and I'll Tell You Your Offenses: Machine Learning-Based Decision Support for Fraud Detection on Medical Claim Data. *Proceedings of the 55th Hawaii International Conference on System Sciences*, *January*.

Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics*, *114*(March), 57–65.

Nissenbaum, H. (2004). Washington law review: Privacy as contextual integrity. *Washington Law Review*, *79*(1), 119–157.

Noorbakhsh-sabet, N., Zand, R., Zhang, Y., States, U., States, U., & Tech, V. (2020). Artificial Intelligence Transforms the Future of Healthcare Nariman. *Am J Med*, *132*(7), 795–801.

Opacus AI. (2022). *PyTorch Opacus*. https://opacus.ai/

Pannekoek, M., & Spigler, G. (2021). *Investigating Trade-offs in Utility, Fairness and Differential Privacy in Neural Networks*.

Papernot, N., Goodfellow, I., Abadi, M., Talwar, K., & Erlingsson, Ú. (2017). Semi-supervised knowledge transfer for deep learning from private training data. *5th International Conference on Learning Representations*, 1–16.

Simonyan, K., & Zisserman, A. (2015). Very deep

convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–14.

Stanfill, M. H., & Marc, D. T. (2019). Health Information Management: Implications of Artificial Intelligence on Healthcare Data and Information Management. *Yearbook of Medical Informatics*, *28*(1), 56–64.

Sun, Z., Wang, Y., Shu, M., Liu, R., & Zhao, H. (2019). Differential Privacy for Data and Model Publishing of Medical Data. *IEEE Access*, *7*, 152103–152114.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *IEEE Security And Privacy*, *10*(5), 1–14.

Tao, W., Concepcion, A. N., Vianen, M., Marijnissen, A. C. A., Lafeber, F. P. G. J., Radstake, T. R. D. J., & Pandit, A. (2021). Multiomics and Machine Learning Accurately Predict Clinical Response to Adalimumab and Etanercept Therapy in Patients With Rheumatoid Arthritis. *Arthritis and Rheumatology*, *73*(2), 212–222.

Wang, P., Lei, Y., Ying, Y., & Zhang, H. (2022). Differentially private SGD with non-smooth losses. *Applied and Computational Harmonic Analysis*, *56*, 306–336.

Xiong, A., Wang, T., Li, N., & Jha, S. (2020). Towards effective differential privacy communication for users' data sharing decision and comprehension. *Proceedings - IEEE Symposium on Security and Privacy*, 392–410.

Zhang, X., Ding, J., Wu, M., Wong, S. T. C., Van Nguyen, H., & Pan, M. (2021). *Adaptive Privacy Preserving Deep Learning Algorithms for Medical Data*. 1168–1177.