TUM School of Computation, Information and Technology
Technische Universität München

Privacy-preserving Artificial Intelligence in Medicine

Alexander A. Ziller

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:
apl. Prof. Dr. Georg Groh

Prüfende der Dissertation:

1. Prof. Dr. Daniel Rückert
2. Prof. Dr. Dr. Jens Kleesiek
3. Prof. Dr. Sotirios A. Tsaftaris

Die Dissertation wurde am 24.06.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 08.11.2024 angenommen.

Abstract

Artificial Intelligence (AI) has become paramount in many areas over the last decade. It has proven to be a valuable addition to medical workflows, where it can assist doctors in precise evaluations of patient conditions. However, highly performant AI models crucially depend on large and diverse datasets. While these datasets are continuously generated in hospitals and medical institutions, they are inaccessible due to the risks of privacy infringements. The term Privacy-enhancing technologies (PETs) summarises the field of technical approaches and algorithms, which aim to reunite AI training and the protection of its training data from unintended leakage.

In this dissertation, we investigate the use of PETs in the context of medical AI approaches. Specifically, we demonstrate a holistic workflow comprised of various PETs that provides protection from attackers while yielding highly performant AI models, even outperforming expert radiologists. The most important PET in this thesis, Differential Privacy (DP), provides mathematical bounds on the risks of information leakage. We analyse the computational overhead DP implementations impose on the training of AI models and provide an alternative which is competitive in runtime and generically compatible with most AI network architectures. Furthermore, we investigate the impact of using DP for medical AI training on the fairness and non-discrimination of subgroups. Here, in contrast to prior work, we find that not the representation of subgroups in the training data is driving fairness impacts, but rather the difficulty of predicting the respective subgroup. In particular, we see that groups with a lower prediction performance in non-private AI models suffer further performance losses with increasing privacy guarantees. This may impact the way of assembling datasets for the training of privacy-preserving fair AI models. Lastly, we analyse how an appropriate level of protection can be determined and find that, for many scenarios, typical privacy budgets are overly pessimistic. We show that by adapting the privacy budget to a concrete threat model, the negative impact of DP on the performance of AI models can be largely mitigated. With these contributions, we hope to advance the widespread breakthrough of technical and mathematical approaches to protecting patient privacy when training medical AI models.

Zusammenfassung

Künstliche Intelligenz (KI) hat im letzten Jahrzehnt Einzug in viele Bereiche gehalten. Es hat sich als wertvolle Ergänzung für medizinische Arbeitsabläufe erwiesen, in dem es Ärzten hilft medizinische Fragestellungen präzise zu beantworten. Allerdings basieren leistungsfähige KI Modelle entscheidend auf der Verfügbarkeit von umfangreichen und vielfältigen Datensätzen. Diese Datensätze werden zwar fortlaufend in Kliniken und Arztpraxen generiert, allerdings sind sie bedingt durch Datenschutzregeln nicht verfügbar für das Training von KI-Modellen. Privatsphärenwahrende Techniken (Privacy-enhancing Technologies, PETs) beschreiben dabei die Vielfalt an technischen Ansätzen, die darauf abzielen das Training von KI-Modellen und Datenschutz zusammenzubringen.

In dieser Dissertation untersuchen wir den Einsatz von PETs im Kontext von medizinischer KI. Insbesondere zeigen wir einen ganzheitlichen Ansatz bei dem mehrere PETs in Kombination zum Einsatz kommen um dabei Schutz vor datenschutzverletzenden Angriffen zu bieten und gleichzeitig hochperformante KI-Modelle hervorzubringen, die sogar Fachärzten in der Diagnose überlegen sind. Die wichtigste privatsphärenwahrende Technik in dieser Dissertation ist Differential Privacy (DP), welches Grenzen über das Risiko von ungewollten Informationsflüssen mathematisch garantiert. Wir analysieren den zusätzlichen Rechenaufwand den DP für das Training von KI-Modellen impliziert und präsentieren eine Alternative, die vom Rechenaufwand mit vorherigen Ansätzen mithalten kann, gleichzeitig aber nativ kompatibel mit allen zulässigen KI-Netzwerkarchitekturen. Desweiteren untersuchen wir das Zusammenspiel von DP und der Fairness und Nicht-Diskriminierung bei medizinischer KI. Konträr zu vorherigen Arbeiten zeigt sich bei uns, dass nicht die Repräsentation einer Gruppe im Datensatz die Fairness beeinflusst, sondern die Schwierigkeit der jeweiligen Prädiktion. Insbesondere sehen wir, dass Gruppen, die bereits bei nicht-privatsphärenwahrenden KI-Modellen schlechter prädiziert werden, zusätzliche Einbußen erfahren je stärker der Privatsphärenschutz ist. Dies könnte beeinflussen wie in Zukunft Datensätze für das Training von fairen und privatsphärenwahrenden KI-Modellen zusammengestellt werden. Zuletzt analyisieren wir wie angemessen Datenschutzlevel festgelegt werden können und stellen fest, dass in vielen Szenarien typische Privatsphärenbudgets sehr pessimistisch sind. Wir zeigen, dass durch die Anpassung dieser Budgets an eine konkrete Gefährdungslage der negative Einfluss von DP auf die Leistung von KI-Modellen weitgehend abgeschwächt werden kann. Mit diesen Beiträgen hoffen wir, den breiten Durchbruch technischer und mathematischer Ansätze zum Schutz der Privatsphäre von Patienten beim Training medizinischer KI-Modelle voranzutreiben.

Acknowledgements

Bled deafst scho sei, aber z'häifa muast da wissn

Bavarian Proverb

Most of the time, acknowledging the people who have been essential while pursuing a PhD also idealises the time itself. During the last four years, I have experienced some of the best and most challenging moments in my life so far. I met the girl of my dreams and lost both of my grandpas. I have learnt and done so much incredibly cool stuff, which often cost me sleepless nights. I worked at the cutting edge of modern technology, even on weekends, holidays, and nights. I will forever look back fondly on this time, but I am also excited to embrace the next chapter of my life.

George, you taught me almost everything I know about research. To this day, I am still amazed by your genius. You were the best mentor anyone could ask for, and I'm deeply grateful to have you as a friend. Thank you, Daniel, for trusting and guiding me throughout the last few years. I've learnt so much from you, especially that working hard and being laid-back are perfectly compatible. Rickmer, I think this is now the right point to do this: I sincerely apologise that this thesis is not about pancreatic cancer prediction. I can't express how much I appreciate your support in exploring something outside your research field. At the same time, you always reminded me of what really matters: getting these fascinating approaches out of the lab and into the hospital so real patients can benefit.

Any description of my PhD would be incomplete if I didn't mention the people I've shared so much time with. First of all, the real radiology OGs: Leo, Alina, and Tamara. It was so much fun with you! I'm still fascinated by Leo's optimisation and Alina's social skills. Let's do another game night soon. I'm also so glad for the fresh energy and good vibe that Jojo, Sarah and Kiki have brought to our team. I had wonderful collaborators in our lab who did projects and organised the practicals with me. Namely, thank you to Can, Johannes, Fritzi, Philip, Felix, Anne, and Hendrik. I also want to thank Dima, Mo, Flo, Jonas, Reihaneh, Florent, and Reza. It was great discussing research ideas and so much more with all of you. There are so many more I would like to name here, and I'm incredibly proud to have been part of this fantastic group. A special shoutout to you, Deborah. You're our lab mum who cares for everyone (even if you'd never admit it) and rule the office with an iron fist. Working with such talented students was one of the most fulfilling parts during these years. It was an honour to accompany all of you, Nico, Chaima, Jakob, Helena, Can, Timo, Tim, Leon, Alp, Anna, Sarah, Denitsa, Othmane, Simon, Johannes, Thomas, Tobi, Marwa, Pires, and Philipp through your studies.

All of this would not have been possible without my amazing family. My loving parents and grandparents were always on my side and I'm forever grateful. And yes, Markus, I'm doctor now. Thanks for being there for me. Finally, the best part about all of this was meeting you, Tamara. Your unconditional love and support have been my anchor, and I am deeply grateful for every moment we've shared.

Contents

List of Figures

Acronyms

AI	Artificial Intelligence
AIA	Attribute Inference Attack
DP	Differential Privacy
DP-FTRL	DP-Follow-the-regularized-leader
DP-SGD	Differentially Private Stochastic Gradient Descent
DRA	Data Reconstruction Attack
ECG	Electrocardiogram
EU	European Union
FDA	Food and Drug Administration
FL	Federated Learning
FPR	False Positive Rate
FTRL	Follow-the-regularized-leader
GDPR	General Data Protection Regulation
HE	Homomorphic Encryption
LLM	Large Language Model
MIA	Membership Inference Attack
MRI	Magnetic Resonance Imaging
PET	Privacy-Enhancing Technology
PSO	Predicate Singling Out
ROC	Receiver-Operator Characteristic

SMPC	Secure Multi-party Computation
TPR	True Positive Rate

Publication List

This dissertation is based on the following peer-reviewed publications. A * indicates shared first authorship.

[1] G. Kaissis*, **A. Ziller***, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima Jr, J. Mancuso, F. Jungmann, M.-M. Steinborn, A. Saleh, M. Makowski, D. Rueckert and R. Braren. 'End-to-end privacy preserving deep learning on multi-institutional medical imaging'. In: *Nature Machine Intelligence* 3.6 (2021), pp. 473–484.

[2] **A. Ziller***, D. Usynin*, R. Braren, M. Makowski, D. Rueckert and G. Kaissis. 'Medical imaging deep learning with differential privacy'. In: *Scientific Reports* 11.1 (2021), p. 13524.

[3] S. Tayebi Arasteh*, **A. Ziller***, C. Kuhl, M. Makowski, S. Nebelung, R. Braren, D. Rueckert, D. Truhn and G. Kaissis. 'Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging'. In: *Communications Medicine* 4.1 (2024), p. 46.

[4] **A. Ziller**, T. T. Mueller, S. Stieger, L. F. Feiner, J. Brandt, R. Braren, D. Rueckert and G. Kaissis. 'Reconciling privacy and accuracy in AI for medical imaging'. In: *Nature Machine Intelligence* 6.7 (2024), pp. 764–774.

The following additional publications were further written *during the time of the doctoral thesis*. A * indicates shared first or last authorship. In chronological order.

[1] D. Usynin, **A. Ziller**, M. Makowski, R. Braren, D. Rueckert, B. Glocker, G. Kaissis and J. Passerat-Palmbach. 'Adversarial interference and its mitigations in privacy-preserving collaborative machine learning'. In: *Nature Machine Intelligence* 3.9 (2021), pp. 749–758.

[2] D. Usynin, **A. Ziller**, M. Knolle, A. Trask, K. Prakash, D. Rueckert and G. Kaissis. 'An automatic differentiation system for the age of differential privacy'. In: *Privacy in Machine Learning Workshop (PriML)* (2021).

[3] **A. Ziller**, D. Usynin, M. Knolle, K. Prakash, A. Trask, R. Braren, M. Makowski, D. Rueckert and G. Kaissis. 'Sensitivity analysis in differentially private machine learning using hybrid automatic differentiation'. In: *Workshop on Theory and Practice of Differential Privacy, ICML* (2021).

[4] M. Knolle, **A. Ziller**, D. Usynin, R. Braren, M. R. Makowski, D. Rueckert and G. Kaissis. 'Differentially private training of neural networks with Langevin dynamics for calibrated predictive uncertainty'. In: *Workshop on Theory and Practice of Differential Privacy, ICML* (2021).

[5] G. Kaissis, M. Knolle, F. Jungmann, **A. Ziller**, D. Usynin and D. Rueckert. 'A unified interpretation of the gaussian mechanism for differential privacy through the sensitivity index'. In: *Journal of Privacy and Confidentiality* 12.1 (2022).

[6] N. W. Remerscheid*, **A. Ziller***, D. Rueckert and G. Kaissis. 'Smoothnets: Optimizing cnn architecture design for differentially private deep learning'. In: *Workshop on Theory and Practice of Differential Privacy, ICML* (2022).

[7] **A. Ziller**, T. T. Mueller, R. Braren, D. Rueckert and G. Kaissis. 'Privacy: An axiomatic approach'. In: *Entropy* 24.5 (2022), p. 714.

[8] H. Klause, **A. Ziller**, D. Rueckert, K. Hammernik and G. Kaissis. 'Differentially private training of residual networks with scale normalisation'. In: *Theory and Practice of Differential Privacy Workshop, ICML* (2022).

[9] D. Usynin, **A. Ziller**, D. Rueckert, J. Passerat-Palmbach and G. Kaissis. 'Distributed Machine Learning and the Semblance of Trust'. In: *The Third AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI-22): Virtual: February 28, 2022.* AAAI Press (Association for the Advancement of Artificial Intelligence). 2022.

[10] T. T. Mueller, S. Kolek, F. Jungmann, **A. Ziller**, D. Usynin, M. Knolle, D. Rueckert and G. Kaissis. 'How Do Input Attributes Impact the Privacy Loss in Differential Privacy?' In: *The Fourth AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI-22): Virtual: February 13, 2023.* AAAI Press (Association for the Advancement of Artificial Intelligence). 2023.

[11] **A. Ziller***, A. C. Erdur*, F. Jungmann, D. Rueckert, R. Braren and G. Kaissis. 'Exploiting segmentation labels and representation learning to forecast therapy response of PDAC patients'. In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI).* IEEE. 2023, pp. 1–5.

[12] **A. Ziller**, A. Güvenir, A. C. Erdur, T. T. Mueller, P. Müller, F. Jungmann, J. Brandt, J. Peeken, R. Braren, D. Rueckert et al. 'Explainable 2D Vision Models for 3D Medical Data'. In: *arXiv preprint arXiv:2307.06614* (2023).

[13] M. Knolle, R. Dorfman, **A. Ziller**, D. Rueckert and G. Kaissis. 'Bias-Aware Minimisation: Understanding and Mitigating Estimator Bias in Private SGD'. In: *Theory and Practice of Differential Privacy* (2023).

[14] F. Meissen, J. Getzner, **A. Ziller**, G. Kaissis and D. Rueckert. 'How Low Can You Go? Surfacing Prototypical In-Distribution Samples for Unsupervised Anomaly Detection'. In: *arXiv preprint arXiv:2312.03804* (2023).

[15] T. T. Mueller, S. Zhou, S. Starck, F. Jungmann, **A. Ziller**, O. Aksoy, D. Movchan, R. Braren, G. Kaissis and D. Rueckert. 'Body fat estimation from surface meshes using graph neural networks'. In: *International Workshop on Shape in Medical Imaging*. Springer. 2023, pp. 105–117.

[16] G. Kaissis, J. Hayes, **A. Ziller** and D. Rueckert. 'Bounding data reconstruction attacks with the hypothesis testing interpretation of differential privacy'. In: *Theory and Practice of Differential Privacy* (2023).

[17] G. Kaissis, **A. Ziller**, S. Kolek, A. Riess and D. Rueckert. 'Optimal privacy guarantees for a relaxed threat model: Addressing sub-optimal adversaries in differentially private machine learning'. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[18] K. Schwethelm, J. Kaiser, M. Knolle, D. Rueckert, G. Kaissis* and **A. Ziller***. 'Visual Privacy Auditing with Diffusion Models'. In: *arXiv preprint arXiv:2403.07588* (2024).

[19] T. T. Müller, S. Starck, K.-M. Bintsi, **A. Ziller**, R. Braren, G. Kaissis and D. Rueckert. 'Are Population Graphs Really as Powerful as Believed?' In: *Transactions on Machine Learning Research* (2024).

[20] **A. Ziller**, A. Riess, K. Schwethelm, T. T. Mueller, D. Rueckert and G. Kaissis. 'Bounding Reconstruction Attack Success of Adversaries Without Data Priors'. In: *arXiv preprint arXiv:2402.12861* (2024).

1 Introduction

Artificial Intelligence has become paramount in many areas of modern life. At the latest, with the breakthrough of large language models such as ChatGPT, the general population has become aware of the capabilities of learning procedures instead of handcrafted algorithms. It comes with the promise to automate, improve, and generally increase the quality of workflows, which previously could only be executed by humans. Strongly performing Artificial Intelligence (AI) models are fuelled by large and diverse datasets. The breakthrough of AI premised the collection of vast datasets such as ImageNet [1]. Currently, the most popular Large Language Models (LLMs), such as GPT-4 by OpenAI [2] or Gemini by Google [3], are presumably trained on all publicly available data. This includes web documents, books, code, images, audio, and videos [3]. However, even on such publicly accessible data, the training of AI models led to privacy concerns. Most prominently, ChatGPT was banned in Italy as authorities suspected the use of personal data for the training of this algorithm [4]. In order to regulate "what is acceptable" in the context of AI training, authorities have come forward with specialised legislation. Most notably, the European Union (EU) established two legal frameworks with direct implications for training and deploying AI algorithms: (1) The General Data Protection Regulation (GDPR) and (2) the AI Act. The AI Act specifically regulates the requirements for training algorithms based on their risk class. For example, it prohibits "socially unacceptable" algorithms, such as social scoring algorithms. GDPR has a more indirect link to the training of AI models. It regulates the protection of personal data as a central right. Hence, if data is leaked, it must be ensured that it cannot be assigned to one specific person and thus is "anonymous". While this is a general regulation without a specific focus on AI, it has strong implications, as large datasets are a prerequisite to strong AI models.

Improvements in medical workflows through the use of AI can directly lead to enhanced diagnoses and better and more rapid treatment, which in turn will improve patients' life quality and expectancy [5]. The benefit AI can provide was demonstrated for tasks ranging from predicting SARS-Cov-2 variants to all-purpose medical LLMs [6, 7, 8, 9]. While it is still in an early stage, there is a potential to revolutionise medical workflows [10]. Thus, there is a demand for the use of AI models in medicine. Yet again, there are ethical and technical challenges, which are so far unaddressed [10]. The fact that legislative bodies are paying such attention to safe and socially compliant aspects of AI demonstrates the societal desire to enforce ethical standards and trustworthiness in these systems. One of them is that health data is strongly protected, as it can contain information about diseases, genetic variants, and patients' lifestyles, to name only a few attributes that are considered personal and private by many. While GDPR regulates all types of data, there is often also dedicated and stricter healthcare legislation. This leads to a situation where although large amounts of high-quality data exist and are continuously generated, it is inaccessible for training medical AI models. Hence, there is a tension between medical AI algorithms dependent on large and diverse medical databases for training and the release of such databases for AI training, which contain highly sensitive information. This tension is further intensified as it has been shown repeatedly that safeguarding the training data is not sufficient, as it also can be leaked from the AI model [11, 12, 13, 14, 15, 16, 17, 18]. A solution to this field of tension could be the use of Privacy-Enhancing Technologies (PETs). These encompass a collection of techniques which in combination, can allow a holistic workflow for a privacy-preserving way of training AI algorithms. In particular, data governance enhancing techniques such as Federated Learning (FL) ensure that data is processed on-site and does not need to be transferred to a central, potentially untrusted instance. Encryption techniques such as Homomorphic Encryption (HE) or Secure Multi-party Computation (SMPC) allow to perform computations while ascertaining that no unauthorised reading of the data can be performed. Most importantly, Differential Privacy (DP) is the key privacy technique providing formal and mathematically provable guarantees on the protection of sensitive outputs. This guarantee can be translated to an upper bound on the success of all privacy-critical attacks, namely Membership Inference Attacks (MIAs) [19], re-identification [20, 21], and Data Reconstruction Attacks (DRAs) [22, 23]. In this thesis, we outline how the use of PETs can mitigate the tension of training strong AI algorithms while protecting the privacy of data owners.

Contributions This dissertation investigates how AI methods in medicine and privacy preservation can be combined into a holistic privacy-preserving workflow as a part of an ethical AI workflow. For this, we investigate several aspects of private deep learning workflows in medicine and healthcare. Furthermore, we detail the current challenges introduced by private AI training with an emphasis on the behaviour in medical problems and under varying threat models. The main contributions can be summarised as follows:

- We showcase how medical workflows can benefit from a holistic privacy-preserving pipeline. We combine several PETs covering different aspects to maintain data governance, ensure confidentiality and guarantee privacy. By this, we can show for the exemplary use case of classifying paediatric pneumonia on chest X-rays that (1) data reconstruction attacks in an honest-but-curious setting can be impeded, (2) distributed training clearly outperforms local AI models, and (3) private AI algorithms can compete or even outperform expert radiologists. These findings lay the foundation for the practical use of privacy-preserving AI workflows, which in turn can unlock access to a larger wealth of health data, in turn leading to stronger models (Section 3.1).

- We evaluate the efficiency of implementations of differentially private trainings and optimise it further to be more generally applicable and time efficient. We benchmark our implementation against open-source frameworks on two medical tasks and find that at the time of our work, we have advantages in the general applicability and memory and/or time efficiency (Section 3.2).

- We investigate the effect of DP on the fairness to subgroups on two relevant medical imaging tasks. We find that, as opposed to the findings of previous works, the loss penalties are not defined by the underrepresentation of a subgroup but rather by their prediction difficulty. In particular, the underperformance of AI models on certain subgroups appears to be exacerbated with increasing privacy protection. Exemplarily, for the task of chest radiograph classification, we observe that older patients, although overrepresented, suffer higher losses on diagnostic accuracy the stronger the guaranteed privacy is. This can be an important consideration for the design of future private AI workflows to compensate utility penalties implied by laying a particular focus on subgroups which are harder to diagnose (Section 3.3).

- We analyse the effect of varying threat models on the trade-off between privacy preservation and AI performance. One of the main obstacles hindering the breakthrough of privacy-preserving machine learning is the imposed trade-off between AI utility and the level of privacy. However, privacy analyses are typically based on worst-case assumptions about the adversary. We demonstrate that relaxations about these assumptions mitigate the trade-off and, in some cases, make it negligible (Section 3.4).

Overview of this Thesis This publication based thesis is structured as follows: Chapter 2 outlines the research field of privacy-preserving AI in medicine. Specifically, in Section 2.1, we introduce the field of trustworthy and ethical AI and, in particular, privacy as a substantial component.We outline that privacy protection could not only be legally mandated but also further improve AI models in sensitive areas such as medicine. Section 2.2 provides detailed descriptions of all relevant Privacy-Enhancing Technologies (PETs), in particular Federated Learning (FL), Encryption, and Differential Privacy (DP). In Section 2.3, we provide an overview of privacy-centred attacks on machine learning systems. Chapter 3 contains all underlying peer-reviewed publications. Lastly, in Chapter 4, we discuss the implications of our publications and give an outlook on potential future research directions.

2 Background

2.1 Trustworthy and Ethical AI

We begin by outlining the characteristics of medical datasets and how these correspond to the components of an ethical and trustworthy AI workflow. In particular, we outline the aspects of privacy preservation and its current challenges.

2.1.1 Medical Dataset Characteristics

AI has pushed the borders of the possibilities in modern medicine. It can boost the early detection rate of breast cancer by $5 - 13\%$ [24]. It can outperform expert cardiologists in diagnosing cardiovascular diseases [25]. It can even automatically identify clusters of leukaemia patients with distinct risks of disease progression without knowing about the outcome beforehand [26]. However, medical datasets and tasks have their unique challenges. Medical data is created in vast amounts in hospitals, medical practices and care centres. These datasets are continuously annotated with high-quality labels as doctors and medical professionals assess the measured data and the patient's condition, i.e., diagnose and treat the diseases. However, at the same time, most AI models used for medical purposes are trained on small datasets. For example, most algorithms approved by the US Food and Drug Administration (FDA) are trained on less than 1 000 data samples [27]. Several factors cause this discrepancy: For one, regulatory obstacles need to be overcome when training with medical data. Ethics committees and data protection officers must agree to trials where medical data is used. Moreover, the data is typically bound to their respective source institutions. At the same time, while large and diverse datasets are often inaccessible, medical tasks are often more complex compared to other use cases. In many cases, examination results are high-dimensional measurements, such as Magnetic Resonance Imaging (MRI) images or Electrocardiogram (ECG). Medical practitioners are trained for years in order to be able to correctly interpret these. Moreover, the relevant information is often subtle and very localised, such as a metastasis in an organ in the MRI or a missing spike on the ECG. Thus, large and diverse datasets are necessary to contain and, by that, allow to learn subtleties and variability over patients. The combination of small datasets, which is harmful to the generalisation of AI models and the complexity of tasks, leads to algorithms which

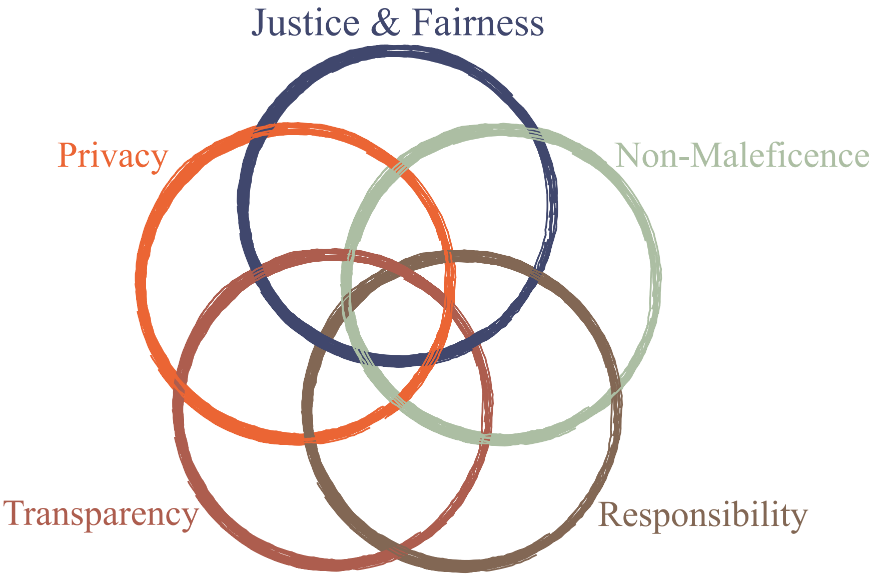


Figure 2.1: Aspects of Ethical AI. Jobin et al. [28] identified five guiding principles for the implementation of ethical AI. This thesis is focused on the aspect of privacy for medical AI and also investigates the interaction with fairness.

are often overly tailored to one specific use case. The immediate drawback is that these models are extremely susceptible to minor changes in the data, such as different measurements (e.g. by using devices from other vendors) or anomalous data, which was not contained in the training set.

In summary, medical tasks are challenging, but the wealth of existing data could enable accurate and robust AI models. Although the data exists, it is inaccessible due to legal and ethical requirements. Hence, AI systems fulfilling these requirements are not only mandated from an ethical perspective but likely also the prerequisite to generalist models.

2.1.2 Components of Trustworthy and Ethical AI

Current AI models are often considered black boxes, where the internal processes are not transparent. However, for the reliable use of AI in sensitive contexts, those models must adhere to basic ethical principles and guidelines [28]. Jobin et al. [28] identify a global agreement on five guiding ethical principles: Transparency, Privacy, Justice & Fairness, Non-Maleficence, and Responsibility (see Figure 2.1). Furthermore, they find additional principles often linked to ethical AI: Beneficence, Freedom & Autonomy, Trust, Sustainability, Dignity, and Solidarity. It is an ongoing research effort to investigate how each ethical principle can be technically implemented for AI systems. In this thesis, we focus on privacy and its potential technical implementations and implications for medical AI models. Privacy has a somewhat exceptional position within these principles, as technical privacy guarantees fulfilling legal definitions by design could unlock access to larger amounts of data. This could, in turn, also boost adherence to other ethical principles. For example, a model without information about a certain subgroup will likely produce larger error rates and thus discriminate. In addition to privacy, we also touch upon the aspect of Justice and Fairness, specifically the interaction with privacy. We note that while other aspects of ethical AI and their technical implementations deserve equal attention, these are out of the scope of this thesis.

Privacy

The main focus of this thesis is the safeguarding of privacy while training medical AI models. In this section, we define privacy and outline the challenge of achieving privacy, particularly outlining why current anonymisation approaches are insufficient and lead to data governance issues.

What is Privacy? Privacy is a social concept, and its exact definition has been refined repeatedly. Perhaps the first characterisation from 1890 defines privacy as "the right to be let alone" [29]. While this fulfils an aspect of privacy many people can likely relate to, follow-up works have identified that this definition is not complete. Jourard [30], or Westin [31] emphasise the control over information. Specifically, Westin defines privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" [31]. Nissenbaum remarked that the control of information flow is not sufficient for a description of privacy, but the appropriate flow of information corresponding to contextual norms is important for maintaining privacy [32]. For example, it is likely considered privacy-conformant for a medical doctor to discuss a patient's condition with a colleague but not with someone unrelated to the medical procedure. Solove moves away from describing privacy as a monolithic definition and instead defines privacy recursively as the solution to a privacy problem [33, 34, 35]. This naturally complements PETs as technical solutions to privacy problems. We have also presented a definition of privacy, which is based on an axiomatic characterisation of the aspects of it [36], incorporating the findings of prior works. Our definition aims for the unification of social science and technical aspects when reasoning about privacy.

There are also various legal frameworks with the goal of imposing privacy. Medical data contains a wealth of sensitive information about patients, ranging from obviously identifying attributes such as names, birth dates, and insurance numbers to more indirect features, such as genetic variants, anomalies, or disease history. Legislation, such as the GDPR, therefore, usually requires the assurance of patient's anonymity. By this formulation, anonymity is achieved if there is a guarantee preventing to re-assign published data, e.g., the weights of an AI model, to a specific person. In technical terms, this corresponds to the notion of predicate singling-out attacks [20]. This type of attack aims to infer a set of attributes which uniquely identify an individual. As this is highly dependent on the available side information, we focus in this thesis on two side-knowledge independent attacks, which are closely related: Membership Inference Attacks (MIAs) and Data Reconstruction Attacks (DRAs). Notably, MIAs are the simplest and DRAs the hardest attack. Hence, the success of re-identification is lower and upper bounded by the success of these attacks. For more details, we refer to Section 2.3.

Current Implementation of Privacy The overwhelming majority of current approaches aim to achieve anonymity of datasets by removing identifiers such as names, birth dates and other highly identifying information. Anonymisation here summarises approaches where these obvious identifiers are removed entirely. In contrast, pseudonymisation describes approaches where identifiers are replaced and a mapping remains, which allows for a later identification of a pseudonymised data sample. There are also more sophisticated variants, such as k-anonymity [37]. Here, the data is discretised in a way such that each value in a column appears at least k-times. Although there is a strong common belief in practice that these procedures are safe, it is not substantiated as non-removed features can still allow for re-identification. This is especially critical for medical data which inherently contains a wealth of identifying information, from rare diseases to genetic information to the shape of the patient's body. Exemplary, the latter has been demonstrated by Schwarz et al. [38], who could match facial photographs to MRI scans of patients. This illustrates the main problem of anonymisation and pseudonymisation, including k-anonymity: All of them are vulnerable to the introduction of side information. In fact, the original publication of k-anonymity already identified the existence of side-information as a vulnerability [37]. Thus, despite their name, current anonymisation procedures fall short of fulfilling the legal requirements of anonymous data.

Data governance Another aspect to consider when training AI models is the governance, i.e. the possession and control, of patient data. Typically, datasets are centralised and copied from their original sources. However, this may induce legal and administrative problems. For example, patients may request the deletion of their data, which is granted in certain legislation, e.g., in the GDPR is incorporated in the right to be forgotten. To avoid this, a strict data tracking system would need to be implemented to have information about each data sample's copies, usage, and storage location. An alternative is to train AI models decentralised, where the data remains at the original owners, and the AI models are sent and trained on-site. Thus, no copies of the data are distributed. We will further explain this concept, which is commonly known as Federated Learning (FL) in Section 2.2.

Fairness

A second key principle of ethical AI besides data privacy is the justice and fairness of AI models [28]. While there are varying definitions, fairness is often expressed as the non-discrimination of subgroups, especially those which are underrepresented. Specifically, whether AI models are robust to introducing biases against underrepresented groups. Several works have drawn substantial attention by showing that biases are often transferred from biased datasets and/or protocols into the models [39, 40, 41, 42]. Especially in the medical field, where AI models are used in delicate tasks, it is crucial –at the very least– to know about such biases. Notably, it has been shown that underrepresented groups are often underdiagnosed by medical AI models and thus may not receive timely attention [43]. It is an active research direction to counteract these biases [42]. However, many studies conclude that the most effective solution is the implementation of an unbiased data acquisition protocol, where protected attributes, such as sex, age or socio-economic status, are accounted for [42, 44]. In this thesis, we are interested in the interaction of PETs and the subgroup fairness of AI models.

2.2 Privacy Enhancing Technologies

This section presents the most important techniques that provide technological solutions to mitigate data governance, confidentiality, and privacy issues. An overview of a holistic workflow comprised of these techniques can be found in Figure 2.2.

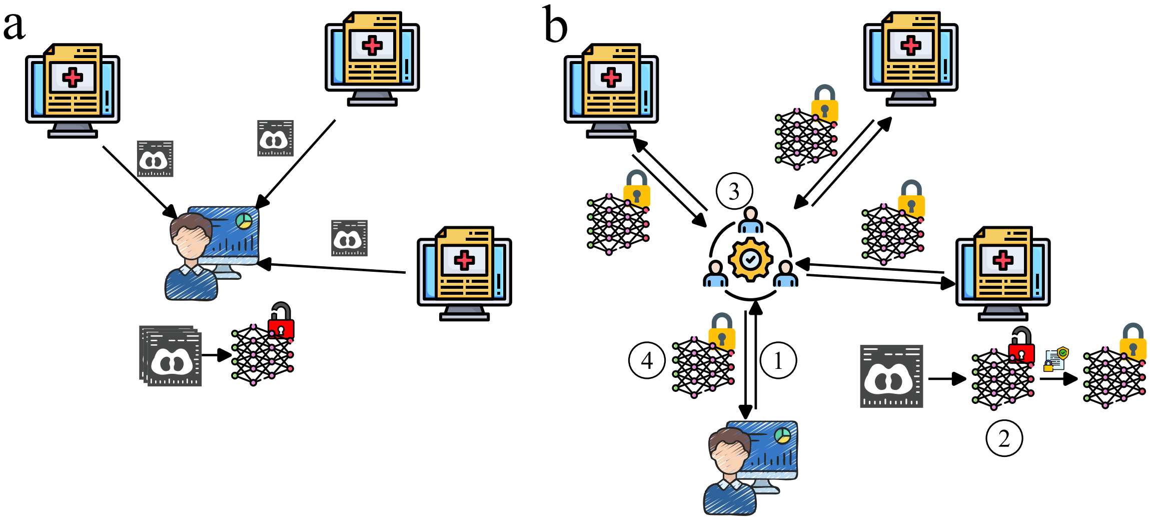


Figure 2.2: Comparison of standard AI training pipeline with a privacy-preserving pipeline. (a) Typical AI training workflow: The data from multiple sites is sent to the AI practitioner, who assembles the data and trains the model. (b) Example for a Privacy-preserving Machine Learning Pipeline: The data remains at the sites. The AI practitioner sends the model to each site (Step 1). A copy of the model is trained on the respective local data at each site. Differential Privacy protects each data point in the training process (Step 2). The trained and privatised AI models of all sites are aggregated using SMPC. The individual models remain secret while the aggregated model is decrypted (Step 4). The aggregated model from all sites is sent back to the AI practitioner (Step 5). This process is iterated either until convergence or until privacy budgets are exhausted.

2.2.1 Federated Learning

In a typical AI workflow, data is collected and centrally aggregated to locally train a model. However, this process requires the copying and distribution of data. Especially for medical data, it can be a legal requirement that its governance remains at the data owner, typically the healthcare provider. Furthermore, legislation such as the GDPR imposes a right to be forgotten, which is practically infeasible if several copies of the data exist. A solution to maintain governance over the data and train AI models could be Federated Learning (FL). FL [45] describes the decentralised training of AI models. Here, the data does not need to be collected to one compute node where the training is happening. Instead, the model is trained on the servers where the data is already stored, and the model updates from all data servers are aggregated. The use of this approach has been demonstrated for CoViD-19 classification [46], cancer detection [47], and industrial drug discovery [48]. However, FL is often mistakenly considered a means to provide data confidentiality or privacy, which is not the case. It has been repeatedly demonstrated that federated learning makes AI training even more vulnerable to data leakage [11, 12, 15, 16]. Hence, for achieving privacy, FL must be combined with other techniques providing these properties presented in the next sections.

2.2.2 Encryption Techniques

When transferring and aggregating AI models of different sites in a FL setting, preventing unauthorised access to the per-site updates is often desirable or even necessary. Encryption techniques are an established and reliable way of providing confidentiality for various scenarios. Standard protocols such as RSA [49] are used for various applications such as browsing, messaging or file transmission. These standard protocols are typically designed to encrypt the message of a sender, which can only be decrypted by the receiver. A special form of encryption is a protocol where computations can be performed on the encrypted data. This would allow the training of a neural network where the processor could not see the network weights or the data but just perform the computations. This family of protocols is commonly referred to as Homomorphic Encryption (HE) [50]. However, so far, the practical use of HE is limited, as these protocols only support very few mathematical operations [51] and are computationally costly [52]. Yet, first works demonstrated the use of HE for the application of AI models [53].

An alternative are so-called Secure Multi-party Computation (SMPC) schemes [54, 55]. These allow several participants to perform computations jointly without any of the participants being able to read the data on their own. These protocols are often more flexible and computationally efficient compared to HE. Thus, they pose a good approach for tasks such as aggregating local models in a federated learning setup, where the participants would like to conceal their individual contributions.

The main difference between HE and SMPC from a methodological point of view is that HE protocols rely on a key-based encryption technique where anyone with the encryption key can read the data, whereas SMPC does not use encryption keys and the data can only be decrypted in collaboration of a certain number of participants.

2.2.3 Differential Privacy

Differential Privacy (DP) is the key technique discussed in this thesis. In the following, we will explain the relevant background knowledge and try to give intuitive access to this topic. The section is mostly based on Dwork et al. [56] and Kaissis et al. [57].

Conceptually, DP is designed as a way to release "answers" about a dataset while limiting the contribution of each individual and, by that, protecting their privacy. For this, DP introduces a "privacy budget" or inversely a "privacy loss". We are using these terms in the following interchangeably. The privacy budget regulates how likely it is that information about specific data samples in the training dataset can be inferred. In other words, it defines the maximum contribution of an individual to the outcome. Dwork et al. [56] frame DP as a promise given to the data owner that DP ascertains that they are not affected by the use of their data. A key factor here is the privacy budget, as larger privacy budgets increase the probability of being affected.

DP is a collection of techniques which provide mathematical guarantees on the maximum influence of individual data samples on the output of a function. More formally, for a randomised Mechanism $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$, all databases $D, D' \in \mathcal{X}$, which differ in exactly one entry, for all $S \in \mathcal{Y} \subseteq \text{Range}(\mathcal{M})$, and a function f, (ε, δ)-DP is satisfied if

$$\Pr(\mathcal{M}(f(D)) \in S) \leq e^{\varepsilon} \Pr(\mathcal{M}(f(D')) \in S) + \delta. \tag{2.1}$$

This guarantee is provided by the randomness in the mechanism $\mathcal{M}$. More concretely, only by the introduction of randomness in the process DP can provide theoretical guarantees. The parameters ε and δ define how "private" the mechanism is. However, as we will outline later in this section, a single pair of (ε, δ) is not sufficient to measure the privacy loss. Intuitively and with a bit of terminological laxity, ε can be thought of as a factor to the "risk" of inferring information about a data sample, while δ is the probability that the actual risk is higher.

Additive Noise Mechanisms

As alluded to in the previous section, the key to obtaining privacy guarantees is the introduction of randomness into a process. Additive noise mechanisms are a family of randomised mechanisms providing DP guarantees, which introduce randomness by the addition of calibrated noise to the release of a query. These are the predominant approaches in privatising numerical queries. In order to choose the "right amount of noise" for a specific query, the randomness is calibrated on the global sensitivity Δ of the query function f. If the space $\mathcal{X}$ is equipped with an ℓ_p-norm ($\|\cdot\|_p$), the sensitivity is defined as

$$\Delta_p(f) = \sup_{D \simeq D'} \left\{ \|f(D) - f(D')\|_p \right\}. \tag{2.2}$$

The sensitivity can be thought of as how much "signal" of one individual is at most in the output of f. Given this global sensitivity, we can define an additive noise mechanism

$$\mathcal{M}(f(D)) = f(D) + \mathcal{Z}(0, \xi) \tag{2.3}$$

$$\xi \propto \Delta_p(f), \tag{2.4}$$

where $\mathcal{Z}$ is a probability measure, and the ℓ_p-norm is an appropriate measure for the sensitivity.

The two prevalent additive noise mechanisms are the Laplacian and the Gaussian mechanism. In the following, we briefly compare both and outline their advantages and drawbacks. As the names suggest, the mechanisms are characterised by the addition of noise sampled from a Laplacian or a Gaussian distribution. The Laplacian mechanism allows for desirable $(\varepsilon, 0)$-DP, where $\delta = 0$, implying that there are no events possible where the actual occurred privacy loss is larger than the guarantee. The main advantage of the Gaussian mechanism is that for repeated queries (such as AI training), it allows a more efficient composition and tighter accounting of the privacy loss [58] (a concept which we further explain in Section 2.2.3). Apart from these key differences, Laplacian noise is also more peaked around its expectation, which often results in small deviations. For this reason, it is optimal for continuous queries and small privacy budgets [59, 60]. However, it also has "heavy tails", i.e., values with large magnitude (outliers) are more likely compared to Gaussian noise. In contrast, strong outliers are substantially less likely for Gaussian noise. In multi-dimensional settings, for the standard Laplacian mechanism satisfying $(\varepsilon, 0)$-DP, the sensitivity is measured by the ℓ_1-norm, while for the Gaussian mechanism satisfying (ε, δ)-DP, both ℓ_1 and ℓ_2-norm can be used. Notably, depending on the situation, the ℓ_1 and ℓ_2-norm can drastically differ in a high-dimensional setting. Specifically, it holds that the ℓ_2-norm is always less than or equal to the ℓ_1-norm. As the added noise is proportional to the sensitivity –which is calculated by the norm– and lower noise is preferable to get more accurate results, the ℓ_2 norm and, thus, the Gaussian mechanism is preferable in high-dimensional settings. In summary, the Laplacian mechanism allows pure ε-DP, whereas the Gaussian mechanism has advantages for repeated and/or high-dimensional queries. As neural network trainings are repeated high-dimensional queries, the Gaussian mechanism is the predominant way of implementing DP guarantees for AI models. Hence, we will focus on the Gaussian mechanism in the following.

The Gaussian Mechanism is described as

$$\mathcal{M}(f(D)) = f(D) + \mathcal{N}(0, \sigma^2 I), \tag{2.5}$$

where σ is the variance based on the sensitivity and privacy budget, and I is the N-dimensional identity matrix, with N being the number of dimensions of the output of the query $f(D)$. Intuitively, we can interpret the relation of sensitivity and noise multiplier as a signal-to-noise ratio $\frac{\Delta}{\sigma}$, where the sensitivity defines the maximum signal, whereas the "strength" of the noise σ distorts the data.

With these preliminaries, the details of the databases can be abstracted away to obtain the most revealing –and by that, least private– output of the query as the least overlapping distributions of two outputs:

$$\mathcal{M}(f(D)) \sim \mathcal{N}(0, \sigma^2 I) \tag{2.6}$$

$$\mathcal{M}(f(D')) \sim \mathcal{N}(\Delta_2(f), \sigma^2 I) \tag{2.7}$$

If this pair of distributions exists, it stands for the outputs on the worst-case pair of databases. These are also referred to as dominating pair [61]. DP can be interpreted as measuring the similarity of the dominating pair to quantify how "private" a mechanism is. There are various ways to measure the similarity between these distributions. In the following, we will focus on the hypothesis testing interpretation of DP, which allows for a very intuitive interpretation of the provided guarantees. However, we note that there are other relevant interpretations, such as Rényi-DP, which measures the similarity as information-theoretic divergence between the distributions [62, 63].

Hypothesis Testing Interpretation

The problem of the adversary to decide whether a given privatised output originates from D or D' can be formulated as a statistical hypothesis test. The null hypothesis H_0 would be that it stems from D, and an alternative hypothesis H_1 that the output comes from D'. Both hypotheses are simple and well-specified, meaning they are fully defined with no unknown parameters. As we know the probability density functions for both hypotheses from the dominating pair, the Neyman-Pearson Lemma can be applied [64]. The lemma states that the optimal test for rejecting the null hypothesis is to compare the likelihood ratio to a cutoff threshold. Specifically, the null hypothesis is rejected if the likelihood ratio is less than or equal to this threshold.

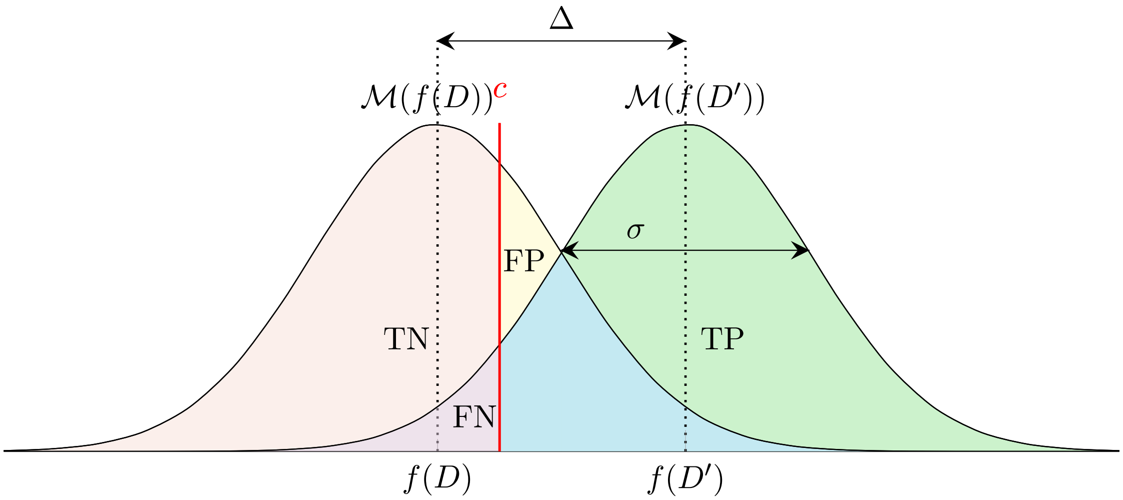


Figure 2.3: Visualisation of the hypothesis testing interpretation of a Gaussian Mechanism. The dominating pair $\mathcal{M}(D)$ and $\mathcal{M}(D')$ define how private a mechanism is. The distance between $f(D)$ and $f(D')$ is, at most, the sensitivity Δ. The "broader" the distributions are, i.e., the higher σ, and by that, the more overlap they have, the more private the output is. Depending on their overlap and the cutoff threshold c chosen by the adversary, the limits on the True Positive Rate (TP, green and blue) and False Positive Rate (FP, yellow and blue) of the adversary are bounded. Figure based on [57].

Based on this cutoff threshold, the adversary can trade off the False Positive Rate (FPR) and True Positive Rate (TPR) (see Figure 2.3). Having simple and well-defined hypotheses implies that the adversary has all knowledge about the process, including the exact function f, the datasets D and D' and the mechanism Therefore, they can not systematically achieve a classification with a better FPR-TPR trade-off (which is also known as Receiver-Operator Characteristic (ROC)-curve) to correctly assign the privatised output to the input. This holds even for the introduction of side information (in contrast to anonymisation) or any post-processing. Thus, the privacy guarantee over the released query cannot be deteriorated by any further computation and thus also holds for any subsequent output. We reiterate that D and D' are datasets which differ in exactly one row and their outputs $f(D)$ and $f(D')$ exhaust the sensitivity of f. Thus, deciding whether the output stems from D or D' is effectively a MIA, where an adversary has to decide whether a given data sample

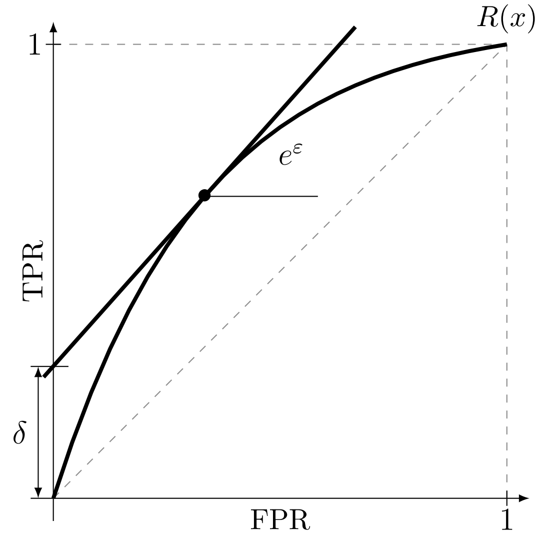


Figure 2.4: Correlation of a privacy-profile and (ε, δ). The optimal ROC-curve of a worst-case MIA adversary describes how private a mechanism is. From this profile, all (ε, δ)-pairs describing this mechanism can be recovered, where e^{ε} is the slope of any tangent and δ is the corresponding intercept. Source: [57].

was included in a query. Hence, by bounding the success of the hypothesis test, the use of DP implicitly also limits the success of MIAs.

The ROC-curve of an adversary describes how private a mechanism $\mathcal{M}$ is. The closer the curve is to the diagonal (i.e., random guessing), the more private the mechanism is. Vice versa, the closer the curve is to the upper left corner (i.e., 100% TPR, 0% FPR), the more revealing the mechanism is. From this privacy profile, all valid (ε, δ) pairs, which the mechanism fulfils, can be recovered. For any point on the ROC-curve where a tangent can be placed, the slope of the tangent is equivalent to e^{ε}, while the intercept of the TPR-axis is δ (see Figure 2.4). For this reason, it is not sufficient to describe a mechanism solely by one (ε, δ)-pair. It also follows that for mechanisms with intersecting privacy profiles, it is not exactly clear which one is “more private”. In contrast, for mechanisms with non-intersecting privacy profiles, i.e., one mechanism dominates the other, the one closer to the diagonal is always more private [65].

Privacy Guarantees for AI Models

Given these preliminaries, we now focus on introducing DP guarantees to the training of AI models. While there are other methods [66], the dominating approach when training private AI models is privatising the intermediate gradients. Neural networks typically have no bound on the output space, i.e., the sensitivity of the function f, where f is calculating a gradient over the weights of a neural network given a specific input is unbounded. A possible solution to limiting the sensitivity of the objective function in training the AI model is implemented by the Differentially Private Stochastic Gradient Descent (DP-SGD) algorithm [67]. Here, the sensitivity is artificially bounded by clipping the norm of the *per-sample* gradients to a pre-defined bound. As per-sample implies, this only holds if the output (i.e. the gradient) depends on the input of only one sample. In this context, a sample describes the data of the instance over which the guarantee is given over. It can be given for each image (per-image guarantee), but in the case of patients contributing multiple data samples, also over the entire data of one patient (per-patient guarantee). The importance of separating the clipping of sample gradients stems from the fact that gradients of multiple samples with higher norms (i.e. sensitivity) can cancel each other out and, by that, avoid being clipped. Therefore, the noise would be calibrated incorrectly, and no valid privacy guarantee would be provided.

Once each per-sample gradient is clipped to a pre-defined norm, the additive noise can be calibrated on the clipping threshold. However, neural network training is an iterative process where each data sample is typically presented many times to the network. Hence, all data samples have a repeated privacy loss. Therefore, an important question is how these repeated privacy losses can be accumulated. This problem is typically referred to as privacy accounting. The simplest way of accounting for the privacy loss of multiple iterations over one sample is the addition of all privacy losses of this data sample. For n iterations, where each is (ε, δ)-DP this yields a final privacy budget of $(n\varepsilon, n\delta)$. However, this accumulates quickly to large values, which are undesirable. Even using the strong composition theorem, which allows for a more efficient accounting of repeated (ε, δ)-DP queries (fulfilled by the Gaussian mechanism), calls for large noise scales for typical privacy budgets and AI trainings. This hindered the breakthrough for the use of DP in AI training, as the injection of large amounts of noise obstructs well-performing networks.

A solution to this problem was proposed by Abadi et al. [68], who derived the so-called moments accountant. Most importantly, they found an effective method of composing privacy losses by tracking the moments of the privacy loss random variable.

Moreover, they incorporated the fact that in AI training data samples only have a low probability of being used in an iteration. Specifically, the probability of a sample used in a training step is defined by the sampling ratio $q = L/N$, where L is the batch size and N is the overall dataset size. Hence, only with probability q information about a sample flows to the model. This effect is known as subsampling amplification. Since then, more sophisticated methods for accounting have been established, such as Rényi-accounting [63] or PRV-accounting [69]. All of them are motivated by having a tight estimate of the true privacy loss the lowest amount of noise is injected into the training process for a given privacy budget and, by that, obtaining better performing models.

2.2.4 Disparity between Private and Non-Private AI Training

DP limits the contribution of individual data samples to the output and, by that, allows the use of data while providing mathematical guarantees for patient privacy. However, it comes at a cost: Generally, PETs and in particular DP-SGD introduce new challenges, which may impede the applicability in practice. One research question in this thesis is how to mitigate or resolve these problems. At the same time, DP also introduces favourable properties. In this section, we outline challenges and open research questions when applying DP in AI training.

Privacy-utility trade-offs

Arguably, the so-called privacy-utility trade-off is the main challenge preventing the widespread breakthrough of DP when training AI models on sensitive datasets. It describes the effect of stronger privacy guarantees typically leading to weaker performance of the resulting private AI models. This imposes a dilemma on practitioners as both of these goals –strong AI models and protection of patient privacy– are important and may be ethically and/or legally mandated [28]. This is because the applications of AI in critical fields such as medicine, lower performing AI models can imply misdiagnosed patients, which may directly or indirectly affect the treatment outcome and life quality. The technical reasons for this are twofold. Most obviously, the introduction of noise calibrated to the sensitivity of the algorithm limits the information about the data. While this is necessary to guarantee the privacy of the patient data, it implies that the signal from which the model learns is overlaid by noise and, therefore, may mislead the search for an optimal set of weights. In

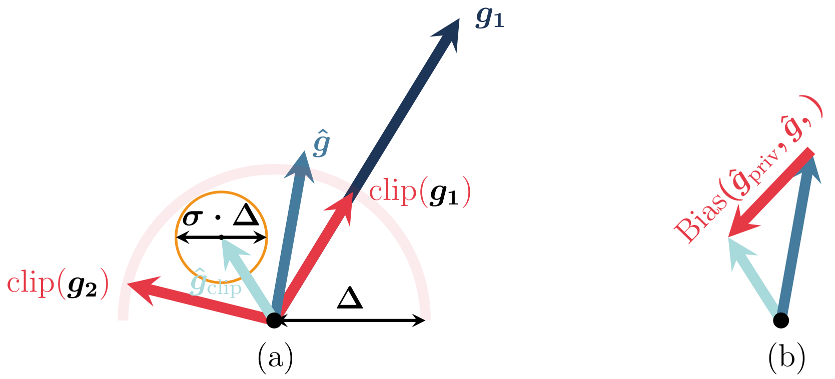

$\boldsymbol{g_1}$
$\hat{g}$
$\mathrm{clip}(\boldsymbol{g_1})$
$\boldsymbol{\sigma \cdot \Delta}$
$\mathrm{clip}(\boldsymbol{g_2})$
$\hat{g}_{\mathrm{clip}}$
$\boldsymbol{\Delta}$
(a)
$\mathrm{Bias}(\hat{g}_{\mathrm{priv}}, \hat{g},)$
(b)

Figure 2.5: Visualisation of the effects on the gradient during DP-SGD. Per-sample gradients g_1 and g_2 are clipped to a maximum sensitivity Δ. The average of clipped gradients $\hat{g}_{\text{clip}}$ is further distorted by the addition of noise with scale σ calibrated to the sensitivity Δ. The direction of the resulting privatised gradient can substantially differ from the non-private gradient $\hat{g}$ and introduce a bias in the learning process. Figure based on [70].

addition to that, clipping the output to a maximum sensitivity as done in DP-SGD can introduce a bias in the learning process [71, 72, 73, 70] (see Figure 2.5). Hence, it is important to find appropriate values for the clipping norm and the strength of the noise, which yield the optimal trade-off between privacy protection and model utility. These performance losses are further exacerbated as other methods which often improve the performance in standard AI training –such as large models or contrastive pretraining– are not straightforward compatible with DP, which we further explain in the following sections.

Increased computational requirements

A practical challenge of providing DP guarantees in the training of AI models is the additional computational overhead. Frameworks for the training of deep neural networks, such as PyTorch [74], are optimised to efficiently compute the average gradient of a batch of input data samples. However, DP-SGD requires the calculation of per-sample gradients. While, in theory, the operations are closely related, in practice, the algorithms are less optimised and thus impose additional time and memory requirements for the training of private AI models. This can be a particular concern for the training of larger models (such as LLMs), large datasets, or restricted compute resources. Mitigating this increased computational overhead is an active research and engineering field [75, 76].

Impact on Subgroup Fairness

An important aspect of critical areas such as medicine is that AI tools work equally well for all patient subgroups. In other words, it is ethically and legally mandated not to systematically discriminate against patients with specific ethnicity, age, sex, or other characteristics [28]. However, several studies have found that DP can negatively affect fairness characteristics of AI models [77, 78, 79, 80, 81, 82, 83]. So far, the effect is not entirely understood and is part of an ongoing research direction.

Other remarks

While the above aspects are investigated in detail in the following sections of this thesis, there are other impacts of DP on AI training, which we would like to briefly remark on here.

Privacy and Generalisation Although DP introduces new challenges, there are also findings which indicate positive effects on the training of AI models. A key problem of machine learning is how to get models to learn from a limited set of data samples to not only perform well on the training data but also on unseen data. In other words, how to find a trade-off between learning enough to “understand” the problem and *generalise* without learning the specific training data “by heart” and *memorise*. Following this observation, it has been shown that DP can lead to a better translation of training performance to test time performance [84]. Furthermore, for the same reason, DP leads to provable robustness guarantees, as small perturbations in the input do not largely impact the output [85]. At the same time, it has been shown that learning, to a certain extent, requires memorisation of out-of-distribution samples [86]. DP guarantees that each sample’s influence is bounded, which implies that the memorisation of individual data samples is bounded [86]. Thus, this can, in turn, contribute to the privacy-utility trade-off.

Further challenges DP training has further intricacies, complicating the straightforward adoption of non-private training processes. One key difference to standard AI training is that DP-SGD requires that gradients are calculated per sample. However, batch normalisation [87], which is one of the most common neural network layers, intermixes the information of the samples in a batch by calculating a statistic over all samples and, based on that, rescales them. As previously outlined, this violates the assumptions of DP-SGD. Therefore, batch normalisation must be replaced by other normalisation layers, which normalise each sample individually. Typical choices for this are group normalisation [88] or kernel normalisation [89]. The latter, which was recently published, does not impose performance losses compared to standard batch normalisation anymore [90]. Still, many pre-trained models and architectures cannot be used out of the box and must be adapted to be privacy-conformant. Another difference to standard AI training is that using a weighted sampling scheme to draw data from the dataset is not straightforward when training on imbalanced training data. This is because current accounting methods assume that all samples are equally likely. While simple approximations to obtain valid guarantees exist [84], there is a lack of tight accounting methods for weighted data sampling. On this note, we remark that the weights for such a sampling scheme are often based on count queries, which also have to be differentially private. The privacy budgets must be offset against the total privacy budget to maintain a complete picture of the privacy loss. Related to that, a crucial part of standard AI training is a tuning of hyperparameters in order to retrieve the optimal model for a specific dataset. Yet, this can lead to information leakage in differentially private machine learning, which is not captured in the privacy budget [91]. Although there are procedures to mitigate this problem [91], it is often neglected in practice. Lastly, it is also important to note that the privacy loss of repeated private queries accumulates. This implies that if a pre-defined privacy budget of a specific dataset is used, this data cannot be used for other procedures. In other words, data cannot be used for an arbitrary amount of AI trainings but is consumed once the privacy budget is expended. These considerations are also relevant to the idea of "data economics" [92].

2.3 Privacy Attacks on AI Models

Implementing defence mechanisms to provide privacy guarantees for the training of AI models premises an understanding of attack vectors. In this section, we outline the most important types of privacy-related attacks and provide an overview of the relation to formal privacy guarantees provided by DP.

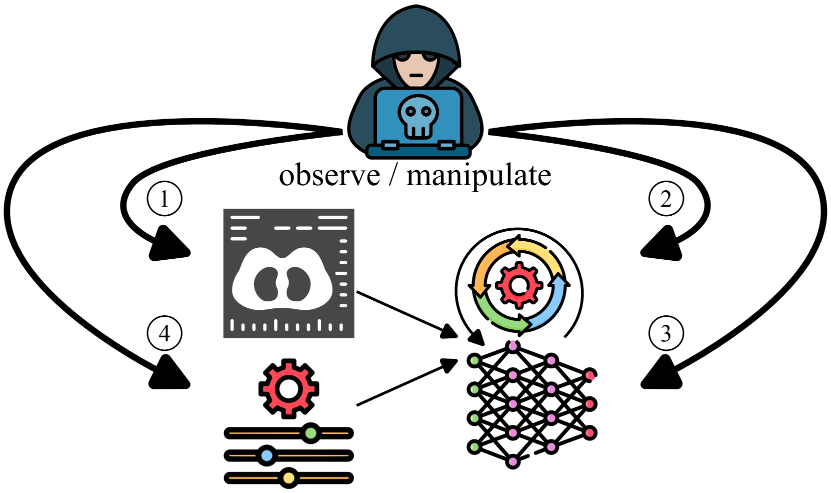


Figure 2.6: Options for various threat models. The threat model describes the capabilities of an adversary in a certain scenario. In a worst-case threat model, the adversary (1) knows the input dataset, (2) can observe all intermediate steps (gradients), (3) can manipulate model architecture and weights, and (4) can modify hyperparameters used for training. The assumptions for real-world adversaries are often not as strong. In particular, the knowledge about the input data is limited for malicious adversaries. Honest-but-curious adversaries do not manipulate or modify any parts of the training pipeline but only observe the outputs.

2.3.1 Threat models

Threat models are a concept from security research used to describe an attacker's capabilities on the system (see Figure 2.6). It is crucial to clearly describe these when analysing the attack's success –i.e. the risk– as it can vary drastically for stronger or weaker adversaries. For example, an adversary that can infiltrate a modified model architecture designed to memorise sensitive data has a much higher chance of succeeding in reconstructing this data than an adversary that only observes [16]. In literature, there are three typical threat models, which we refer to in the following as *worst-case*, *malicious*, and *honest-but-curious*. While the worst-case threat model stems from DP literature, malicious and honest-but-curious come from general AI and especially FL attack research. Hence, the latter two focus on practical scenarios, while the worst case is more used for theoretical risk assessment. In more detail, the threat models can be described as follows: (1) The *worst-case* threat model considers an adversary with unlimited capabilities and full knowledge of the scenario.

Notably, this includes unbounded computational power and knowledge about the data in question. This attacker only lacks knowledge of (a) the answer to the attack's question, e.g., in the case of a MIA, whether a given data sample was used in training, and (b) in the case of DP, the exact random noise, which was used in the mechanism. This threat model is the predominant model in DP literature. While studying these adversaries is primarily of a theoretical nature, the analysis has one crucial advantage: The resulting analyses of such adversaries are absolute worst-case bounds, implying that irrespective of the power or side knowledge that can be added, no result can produce a higher attack success than these adversaries. This includes that the results are robust to any form of post-processing. However, the risk estimates are often very pessimistic for any practical case. Examples of such attacks and analysis can be found in Nasr et al. (Dataset) [93] for MIA, and Hayes et al. [23] for DRA. (2) The *malicious* threat model considers adversaries who actively interfere with the setup in order to gain their optimal advantage in performing the attack. This can include modifications to the model architecture, hyperparameters or other training details. However, the notable difference to worst-case adversaries is that they do not have unlimited power and knowledge, rendering them as powerful and also a realistic threat to AI systems. Examples are described by Nasr et al. (Poison) [93], Fowl et al. [15], Boenisch et al. [16] or Feng & Tramèr [94]. (3) In the *honest-but-curious* threat model, an adversary attempts to infer information about the training data without interfering with the setup. Specifically, these adversaries only distil information from a given scenario, including fixed pre-defined model architectures and hyperparameters. Examples are demonstrated by Nasr et al. (API) [93] and Geiping et al. [11]. Notably, in literature, there is also a common distinction between white-box (i.e. the adversary can observe and potentially manipulate intermediate steps) and black-box access (i.e. the adversary only observes the final output, depending on the definition, even just the predictions without weights). From a theoretical point of view, it has been shown that for the success of a MIA, these scenarios are equally susceptible [95].

2.3.2 Attack Types

A wide variety of attacks has been developed aiming to undermine various aspects of AI systems, most importantly, privacy of input data and utility of the model. We refer to Usynin et al. [96] for an overview. In this thesis, we focus on privacy-centred attacks and within those on Membership Inference Attacks (MIAs) and Data Reconstruction Attacks (DRAs). An overview can be found in Figure 2.7.

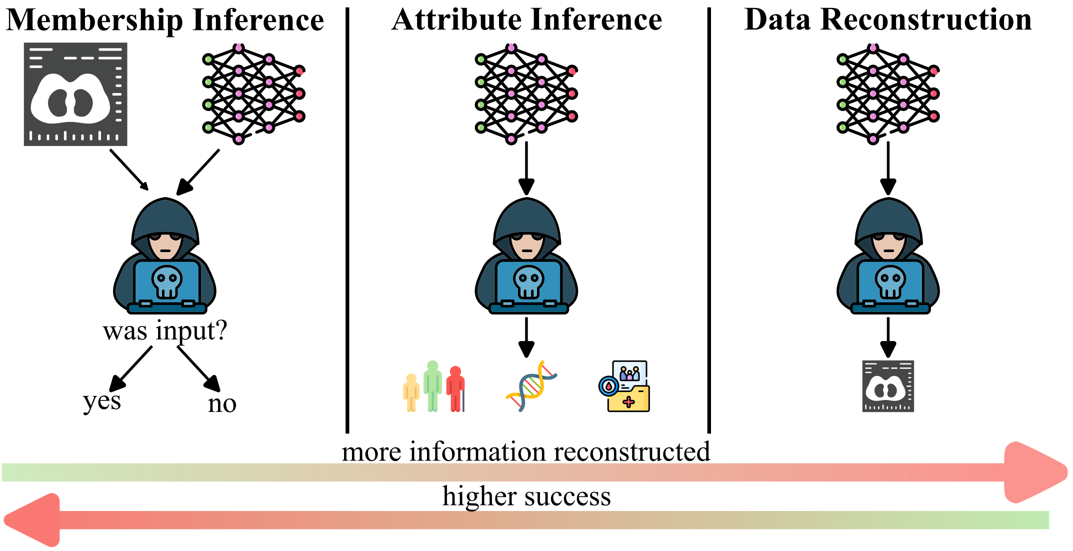


Figure 2.7: Types of privacy attacks on machine learning models. Membership Inference Attacks are the simplest attacks, revealing just one bit of information, namely if a data sample was used for training an AI model. Data Reconstruction Attacks, at the other extreme, aim to recover the full input but are the hardest to perform. Attribute Inference Attacks are a trade-off between information recovery and complexity as they recover only some attributes of the input, e.g., age, genetics or medical history of the patient's input data.

Membership Inference Attacks

In a MIA, an attacker attempts to infer whether a given data sample was part of the input to the neural network during training [97]. Revealing this information can –depending on the scenario– be privacy critical. For example, it can reveal a patient's diagnosis if their data was used in a specific application [97]. However, MIA is the *simplest* and *weakest* attack as it recovers the least amount of information, namely a binary state: Member or Non-member. For this reason, MIA is –just like a worst-case threat model– attractive for theoretical analysis. If the success of a MIA is bounded, it is also bounded for any other type of attack. Notably, the hypothesis testing interpretation of DP immediately provides a bound for the success rate of a MIA by a worst-case adversary. The bounds have also been formalised for weaker adversaries, which do not have exact knowledge of the training data, allowing to also provide theoretical risk bounds for more realistic adversaries [98]. Empirically, it has been shown that the bounds provided by DP are tight and relaxing the threat model decreases the attack success [93].

Data Reconstruction Attacks

While MIA recovers the least amount of information, DRAs, also known as model inversion attacks, are on the other end of the spectrum by attempting to recover the full information, i.e. the exact input. Hence, these are the most challenging but strongest attacks for an attacker and thus, any weaker attack will achieve *at least* the same success rate.

Empirical DRAs Several empirical works have shown that reconstructing the input data works well to a certain extent: In an honest-but-curious scenario, the predominant approaches are based on gradient matching, i.e. optimising a random noise image, which yields the same gradient as the one observed by the adversary [11, 12]. Nearly perfect reconstructions are achievable if the adversary can also manipulate the setting in their favour in a malicious threat model [15, 16]. Moreover, it has been shown that specifically Diffusion Models [99, 100] as generative models are vulnerable to leaking their input data [17]. Several works [22, 14, 18] demonstrated that input data can also be reconstructed from trained model weights in an honest-but-curious threat model. However, these have certain limitations, either having strong assumptions on the model architectures or limited reconstruction quality. Feng & Tramér showed that by manipulating the architecture, which they term privacy backdoors, these limitations vanish [94]. So far, there has been no attack that successfully reconstructed input data of an arbitrary model without observing intermediate steps or manipulating the architecture (black-box attacks).

Theoretical Bounds on DRAs To formalise theoretical bounds on the success of reconstruction attacks and specifically under the use of DP as a privacy-preserving mechanism, Balle et al. [22] proposed the notion of (η, γ)-Reconstruction Robustness (ReRo). Based on an arbitrary reconstruction error function and a prior over the input data, they measure what the probability γ is that the actual reconstruction error is lower or equal to a fixed threshold η. If any defence mechanism can guarantee a limit on γ for a specific value of η, it is considered (η, γ)-ReRo.

Balle et al. [22] proved that DP fulfils (η, γ)-ReRo. In other words, any mechanism that fulfils the requirements imposed by DP, including DP-SGD, also provides bounds on the success of reconstructing input data. Hayes et al. [23] derived tight bounds for a worst-case adversary attempting to achieve $(0, \gamma)$-ReRo under DP. Notably, this analysis is based on an adversary with access to a prior set containing the target data sample. The attack is then successful if the adversary can match the output (i.e., the model or gradient) to the correct input sample. Thus, while this analysis is valuable for the theoretical understanding of risks, it may have limited real-world significance. Currently, it remains an open challenge to provide tight and real-world applicable bounds on the success of reconstruction attacks under DP conditions.

Attribute Inference Attacks

As outlined in the previous sections, MIA represents the minimal information recovery, while DRA recovers the full information. Hence, the question arises if there is a middle ground which recovers *sufficient* information. This would have stronger implications on breaches of privacy compared to MIA, but at the same time might achieve higher success than DRAs. A family of attacks fulfilling this definition are Attribute Inference Attacks (AIAs). As the name suggests, these attempt to infer specific attributes from the model about its input data, which are privacy critical [101]. This is especially critical in combination with Predicate Singling Out (PSO) [20, 21], which attempt to infer enough attributes to uniquely identify an individual. Allowing such an attack is explicitly prohibited by the GDPR [20]. Again, DP can serve as a provable defence against this type of attack [20].

3 Publications

3.1 End-to-end privacy preserving deep learning on multi-institutional medical imaging

Synopsis: Using large, multi-national datasets for high-performance medical imaging AI systems requires innovation in privacy-preserving machine learning so models can train on sensitive data without requiring data transfer. Here we present PriMIA (Privacy-preserving Medical Image Analysis), a free, open-source software framework for differentially private, securely aggregated federated learning and encrypted inference on medical imaging data. We test PriMIA using a real-life case study in which an expert-level deep convolutional neural network classifies paediatric chest X-rays; the resulting model's classification performance is on par with locally, non-securely trained models. We theoretically and empirically evaluate our framework's performance and privacy guarantees, and demonstrate that the protections provided prevent the reconstruction of usable data by a gradient-based model inversion attack. Finally, we successfully employ the trained model in an end-to-end encrypted remote inference scenario using secure multi-party computation to prevent the disclosure of the data and the model.

Contributions of thesis author: code development, experiment design and evaluation, paper writing.

nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00337-8

End-to-end privacy preserving deep learning on multi-institutional medical imaging

Georgios Kaissis[1,2,3,4,13], Alexander Ziller[1,2,4,13], Jonathan Passerat-Palmbach[3,4,5], Théo Ryffel[4,6,7], Dmitrii Usynin[1,2,3,4], Andrew Trask[4,8], Ionésio Lima Jr[4,9], Jason Mancuso[4,10], Friederike Jungmann[1], Marc-Matthias Steinborn[11], Andreas Saleh[11], Marcus Makowski[1], Daniel Rueckert[2,3] and Rickmer Braren[1,12] ✉

Using large, multi-national datasets for high-performance medical imaging AI systems requires innovation in privacy-preserving machine learning so models can train on sensitive data without requiring data transfer. Here we present PriMIA (Privacy-preserving Medical Image Analysis), a free, open-source software framework for differentially private, securely aggregated federated learning and encrypted inference on medical imaging data. We test PriMIA using a real-life case study in which an expert-level deep convolutional neural network classifies paediatric chest X-rays; the resulting model's classification performance is on par with locally, non-securely trained models. We theoretically and empirically evaluate our framework's performance and privacy guarantees, and demonstrate that the protections provided prevent the reconstruction of usable data by a gradient-based model inversion attack. Finally, we successfully employ the trained model in an end-to-end encrypted remote inference scenario using secure multi-party computation to prevent the disclosure of the data and the model.

The rapid evolution of artificial intelligence (AI) and machine learning (ML) in biomedical data analysis has recently yielded encouraging results, showcasing AI systems able to assist clinicians in a variety of scenarios, such as the early detection of cancers in medical imaging[1,2]. Such systems are maturing past the proof-of-concept stage and are expected to reach widespread application in the coming years as witnessed by rising numbers of patent applications[3] and regulatory approvals[4]. The common denominator of high-performance AI systems is the requirement for large and diverse datasets for training the ML models, often achieved by voluntary data sharing on behalf of the data owners and multi-institutional or multi-national dataset accumulation. It's common for patient data to be anonymized or pseudonymized at the originating institution, then transmitted to and stored at the site of analysis and model training (known as centralized data sharing)[5]. However, anonymization has proven to provide insufficient protection against re-identification attacks[6,7]. Therefore, large-scale collection, aggregation and transmission of patient data is critical from a legal and an ethical viewpoint[8]. Furthermore, it is a fundamental patient right to be in control of the storage, transmission and usage of personal health data. Centralized data sharing practically eliminates this control, leading to a loss of sovereignty. Moreover, anonymized data, once transmitted, cannot easily be retrospectively corrected or augmented, for example by introducing additional clinical information that becomes available.

Despite these concerns, the increasing demand for data-driven solutions is likely to increase health-related data collection, not only from medical imaging datasets, clinical records and hospital patient data, but also for example via wearable health sensors and mobile devices[9]. Hence, innovative solutions are required reconcile data and protect privacy. Secure and privacy-preserving machine learning (PPML) aims to protect data security, privacy and confidentiality, while still permitting useful conclusions from the data or its use for model development. In practice, PPML enables state-of-the-art model development in low-trust environments despite limited local data availability. Such environments are common in medicine, where data owners cannot rely on other parties' privacy and confidentiality compliance. PPML can also provide guarantees to model owners that their model will not be modified, stolen or misused, for example by its encryption during use. This lays the groundwork for sustainable collaborative model development and commercial deployment by alleviating concerns of asset protection.

Evidence from prior work

Recent work has shown the utility of PPML in biomedical science and medical imaging in particular. For instance, federated learning (FL) is a decentralized computation technique based on distributing machine learning models to the data owners (also referred to as computation nodes) for decentralized training instead of centrally aggregating datasets. It has been proposed as a method to facilitate multi-national collaboration while obviating data transfer. In the setting of the COVID-19 pandemic[10,11] FL was used to allow the retention of data sovereignty and the enforcement of local governance policies over data repositories. In medical imaging, recent studies[5,12] demonstrated that federated training of deep learning models on brain tumour segmentation or breast density classification performs on-par with local training and that it fosters the inclusion of data from more diverse sources, leading to improved generalization.

[1]Institute of Diagnostic and Interventional Radiology, Technical University of Munich, Munich, Germany. [2]Artificial Intelligence in Medicine and Healthcare, Technical University of Munich, Munich, Germany. [3]Department of Computing, Imperial College London, London, UK. [4]OpenMined. [5]ConsenSys Health, New York, NY, USA. [6]INRIA, ENS, PSL University, Paris, France. [7]Arkhn, Paris, France. [8]Centre for the Governance of AI, University of Oxford, Oxford, UK. [9]Universidade Federal de Campina Grande, Campina Grande, Paraíba, Brazil. [10]Cape Privacy, New York, NY, USA. [11]München-Klinik Schwabing, Munich, Germany. [12]German Cancer Consortium (DKTK), Partner Site Munich, Munich, Germany. [13]These authors contributed equally: Georgios Kaissis and Alexander Ziller. ✉e-mail: rbraren@tum.de

However, FL in itself is not a fully privacy-preserving technology. Previous studies[13,14] demonstrate that inversion attacks can reconstruct images from model weights or gradient updates with impressive visual detail. Moreover, in the setting of inference-as-a-service[15], exposure of the model to a non-trusted third party can enable model misuse or outright theft. Therefore, FL must be augmented by additional privacy-enhancing techniques to truly preserve privacy. For example, FL with secure aggregation (SecAgg) of weights or gradient updates or differential privacy (DP) can prevent dataset reconstruction attacks, and the utilization of secure multi-party computation (SMPC) protocols during model inference can protect the models in use. We provide an overview of these techniques in our previous work[16].

Aim and contributions

The clinical application of PPML in medical imaging requires the development of frameworks for security and privacy, and their validation on non-trivial clinical tasks. Here we present PriMIA, a free, open-source framework for end-to-end privacy-preserving decentralized deep learning on medical images. Our framework incorporates differentially private federated model training with encrypted aggregation of model updates as well as encrypted remote inference. Our contribution provides the following innovations:

- We demonstrate the training of a deep convolutional neural network (CNN) on the clinically challenging task of paediatric chest radiography classification using FL augmented with PriMIA's privacy-enhancing techniques over the public Internet.
- Our framework is compatible with a wide range of medical imaging data formats, easily user-configurable and introduces functional improvements to FL training (weighted gradient descent/federated averaging, diverse data augmentation, local early stopping, federation-wide hyperparameter optimization, DP dataset statistics exchange), increasing flexibility, usability, security and performance.
- We examine the computational and classification performance of models trained with and without privacy-enhancing techniques against models trained centrally on the accumulated dataset, personalized models trained on subsets of the data and against expert radiologists on unseen real-life datasets to evaluate various scenarios typical in medical imaging research.
- We assess the theoretical and empirical privacy and security guarantees of our framework and provide examples of applying a state-of-the-art gradient-based model inversion attack against the models under a number of training scenarios.
- Finally, we showcase the utilization of the trained model in a secure inference-as-a-service scenario without the disclosure of either the data or the model in plain text and demonstrate the improvements in inference latency of our SMPC protocol.

Library functionality

PriMIA was developed as an extension to the PySyft/PyGrid ecosystem of open-source PPML tools. PySyft (https://github.com/OpenMined/PySyft) is a Python framework allowing the remote execution of machine learning tasks (for example, tensor manipulation) and for encrypted deep learning by interfacing with common machine frameworks such as PyTorch. PyGrid provides server/client functionality for the deployment of such workflows on servers and edge computing devices. A detailed description of the generic functionality provided by these frameworks can be found in our previous work[17]. PriMIA builds upon this functionality towards medical-imaging-specific applications by being natively compatible with medical imaging data formats such as DICOM and able to operate on medical datasets of arbitrary modality and dimensionality (for example, computed tomography, radiography, ultrasound and magnetic resonance imaging). Outside of the above-mentioned PPML techniques, it offers solutions to common challenges in medical imaging analysis workflows, such as dataset imbalance, advanced image augmentation, federation-wide hyperparameter tuning functionality. Furthermore, it provides an accessible user interface for applications ranging from local experimentation on the user's machine to distributed training on remote compute nodes to facilitate the application of PPML best practices in medical consortia. The source code and documentation for the library and the publicly available data are provided at https://doi.org/10.5281/zenodo.4545599[18].

Case study, system design and threat model

We present a case study for the application of PriMIA on clinical data by training an 11.1 million parameter ResNet18 CNN[19] on the paediatric pneumonia dataset originally proposed by Kermany et al.[20] on cloud compute nodes over the public Internet with the aim of classifying paediatric chest radiographs into one of three categories: normal (no signs of infection), viral pneumonia or bacterial pneumonia. Pneumonia is a leading cause of paediatric mortality[21]. Chest radiography is routinely performed for differential diagnosis and therapy selection, but classifying paediatric chest radiographs is challenging. The case study is set up according to the following real-life scenario:

FL training phase. A confederation of three hospitals wishes to train a deep learning model for chest radiography classification. As they neither possess enough data on their own nor the expertise to train the model on this data, they enlist the support of a model developer to orchestrate the training on a central server. In the training phase, we refer to the hospitals holding patient data as the data owners. We utilize the term 'model' throughout the manuscript to refer to the structure and parameters of a deep neural network. We assumed an honest-but-curious threat model as defined previously[22] for the training phase. Here, participants trust each other to not actively undermine the learning protocol with utility degradation in mind, for example by actively supplying adversarial inputs or low-quality data (honest). However, individual participants and colluding groups of participants are assumed to actively attempt to extract private information from other participants' data (curious). Our framework's privacy-enhancing techniques are designed to protect from this behaviour, which we describe in detail in later sections. In brief, DP gradient descent[23] extends the guaranteed properties of DP to deep neural network training. Specifically, it bounds the worst-case privacy loss of individual patients in the datasets and provides privacy guarantees against model inversion/reconstruction attacks carried out against federation participants or against model owners at inference time. PriMIA implements DP for each FL node (local DP) to provide patient-level guarantees. Per-node privacy budgeting is performed using the Rényi Differential Privacy Accountant[24]. SMPC allows parties to jointly compute a function over a set of inputs without disclosing their individual contributions. During training, it is utilized to securely average the network weight updates (SecAgg). Additive secret sharing based on the SPDZ protocol[25] is used for SecAgg. The training phase is shown in Fig. 1. It concludes with all participants holding a copy of the fully trained final model.

Remote inference phase. Once fully trained, the model can be used for remote inference. In our case study, we assume that a different data owner, in this case a physician at a remote location holds some patient data and wants to receive an inference result for diagnostic assistance from the model. The inference service is provided over the internet by the model owner. The data and model owners do not trust each other and wish their data and model to remain private. PriMIA's SMPC protocol guarantees the cryptographic security of

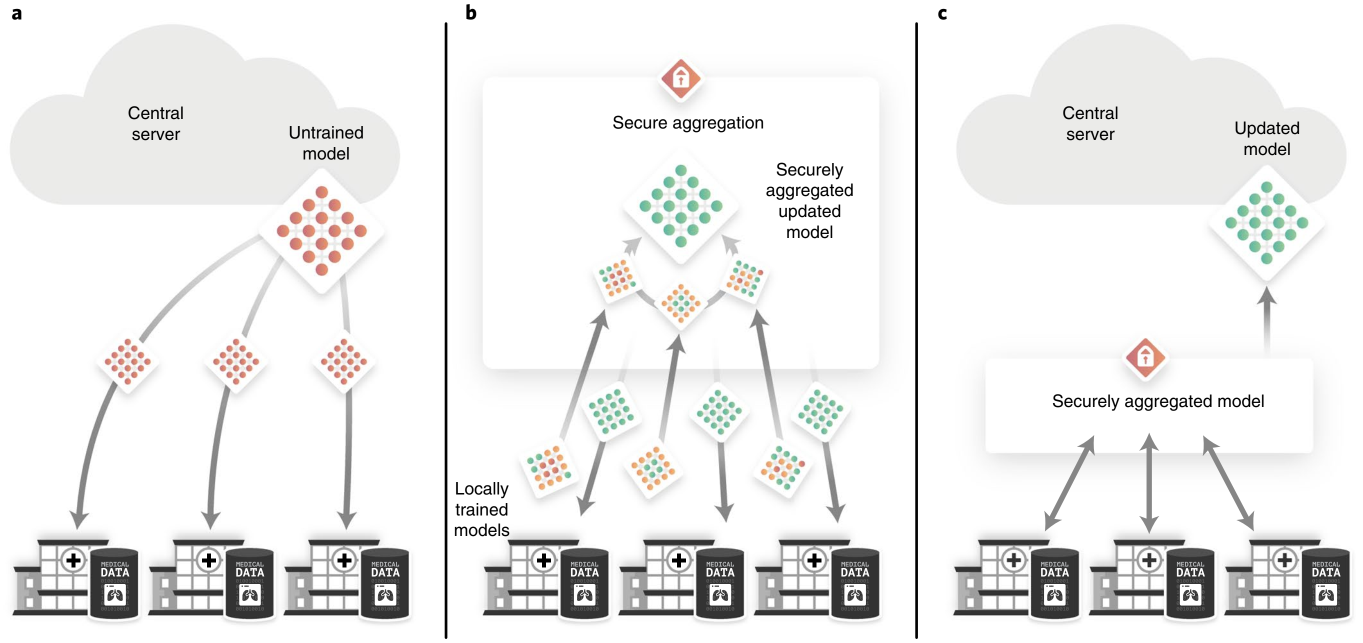


Fig. 1 | Overview of the FL training phase in the PriMIA case study. Three data owners (hospitals) wish to cooperate to train a model; a central server orchestrates the training. **a**, At the beginning of training, the central server sends the untrained model (red) to the computation nodes (hospitals/data owners) for training. **b**, Until convergence is achieved, the models are trained locally at each hospital. Intermittently, the models (coloured) are securely averaged (SecAgg). The SecAgg procedure occurs only between the three data owners. The SMPC protocol guarantees that the individual models cannot be exposed by other participants. After SecAgg, the updated model (green) is redistributed for another round of training. **c**, After the final iteration, the central model is updated with the (now fully trained) securely aggregated model (green) and can be used for inference.

both the model and the data in the inference phase. The AriaNN framework described in our previous work[26] is used, which we have adapted to end-to-end encrypted inference.

A common SMPC technique[25] is the utilization of cryptographically secure random numbers (cryptographic primitives) generated ahead of time (so-called offline phase) to accelerate certain computations. The trusted system (for example, a hardware device) providing these primitives is referred to as a cryptographic provider and is not involved in the actual inference procedure (online phase), nor does it ever come in contact with any party's data. In fact, a 'stockpile' of cryptographic primitives can be provided to the protocol participants ahead of time to be used up over multiple inference procedures. The encrypted inference process is summarized in Fig. 2.

Classification performance

We trained FL models without SecAgg or DP (DP-/SecAgg-), with SecAgg only (DP-/SecAgg+) and with both techniques (DP+/SecAgg+). Furthermore, we trained a model on the entire dataset pooled on a single machine (centrally trained) and separate models on the individual data owners' subsets of the dataset (personalized). The centrally trained model represents the centralized data sharing scenario described in the introduction. The personalized models each represent a single institution training exclusively on their own data, a typical case in current medical imaging research workflows. FL aims to enable the training of models that are better than personalized training and—ideally—as good as the centrally trained model.

We tested the classification performance of the models on the validation set and against the classification performance of two expert radiologists on test set 1 (145 images) and against clinical ground truth data on test set 2 (345 images). We used accuracy, sensitivity/specificity (recall), receiver-operator-characteristic-area-under-the-curve (ROC-AUC) and the Matthews correlation coefficient (MCC)[27] for assessment. Details can be found in the Methods section. Model and expert classification performance on the datasets can be found in Table 1.

The FL model trained with neither SecAgg nor DP performed best with no statistically significant difference to the centrally trained model. The addition of SecAgg to the model slightly, but non-significantly reduced performance. Both FL models and the centrally trained model significantly outperformed the human observers. The DP training procedure ($\epsilon = 6.0$, $\delta = 1.9 \times 10^{-4}$ at an α-value (divergence order) of 4.4) significantly reduced model performance, however the model still performed statistically on par with human observers and retained stable performance on the out-of-sample data of test sets 1 and 2. We note that the ϵ-value represents the total privacy budget spent at the end of training. The personalized models trained only on the data owners' individual data subsets performed approximately on par only on the validation data, but significantly worse on the out-of-sample data of test sets 1 and 2, indicating poor generalization. The statistical evaluation of these results alongside inter-rater/model agreement metrics can be found in Supplementary Section 2 and Supplementary Tables 1 and 2.

Training and inference performance benchmarking

To assess the performance ramifications of PriMIA's privacy-enhancing techniques, we benchmarked the training and inference performance in a variety of scenarios, shown in Fig. 3. Training timings were measured as average time per batch at a constant batch size to decouple them from dataset size. Compared to training locally, FL incurs a performance penalty due to network communications, which is further increased by the addition of SecAgg and DP, yielding a threefold increase in training time when both SecAgg and DP are used. Large neural network architectures require proportionally longer to train due to network transfer

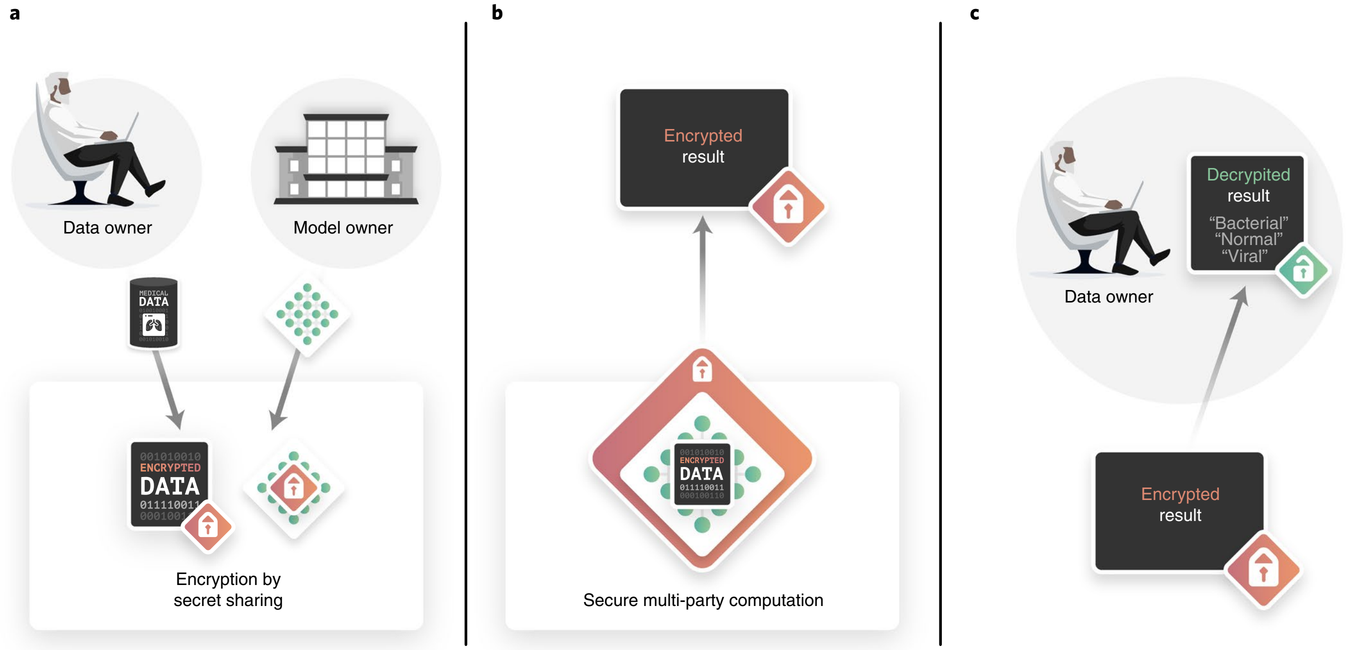


Fig. 2 | Overview of the encrypted inference process. The data owner (in this case, a physician located at a remote location) requests an inference result from the model over the Internet but wants the confidential patient data they hold to remain secret. Similarly, the model owner provides inference as a service but wants to keep their model confidential. The use of SMPC enables the following scenario. **a**, Initially the data owner and model owner respectively encrypt the data and model using secret sharing. This process relies on splitting the data/model into shares, which in themselves do not contain any usable information and can therefore be exchanged (shared) with the other party. **b**, Inference is then carried out by jointly computing a function (in this case the neural network inference procedure) using SMPC. **c**, The data owner receives an encrypted result, which only they can decrypt.

requirements, providing justification for the use of the ResNet18 architecture in our study compared with larger ResNets. The addition of more worker nodes led to a linear increase in times when utilizing SecAgg due to the communication overhead of the protocol. However, due to the small number of operations per round, the protocol scales well to multiple parties: linear regression analysis of the scaling yielded $t(w) = 0.57w + 2.61$ with t expressing time in seconds and w the number of workers ($R^2 = 0.98$, $p < 0.001$, $N = 100$ samples per number of workers tested). Training time was nearly constant without SecAgg. Training times per batch were constant for larger dataset sizes, signifying that training duration is dependent only on dataset size all other things being equal. Lastly, we benchmarked our encrypted inference implementation[26] based on the function secret sharing (FSS) protocol[28], which offers increased efficiency in the evaluation of comparison operations, max-pooling and batch normalization layers compared to the widely used SecureNN[29]. The utilization of FSS for encrypted inference significantly reduced inference times. In particular, in the high-latency setting, FSS yielded a proportionally better performance in comparison to SecureNN. Implementation details can be found in the Methods section and the statistical evaluation can be found in Supplementary Section 3.

Model inversion attack

Prior work[13,30] has demonstrated that model inversion attacks are able to reconstruct features or entire dataset records (in our case, chest radiographs), rendering them a threat to patient privacy in FL settings. To exemplify the susceptibility of models trained with and without the privacy-enhancing techniques offered by PriMIA, we utilized the improved deep leakage from gradients attack[31,32] with small modifications detailed in the Methods section. We chose this method because it was the first technique shown to be highly effective against the ResNet18 architecture used in our case study. Figure 4 shows exemplary results from the chest radiography case study. We utilized the pixelwise mean squared error (MSE), signal-to-noise ratio (SNR) and Fréchet inception distance (FID) metrics for quantifying attack success. Empirical evaluation yielded that the attack's success depends highly on the L2-norm of the gradient updates and the batch size used. To thus generate a best-case baseline of a highly successful attack, we attacked the centrally trained model with a batch size of one at the start of training, when the loss magnitude (and thus gradient norm) is highest. The attacks on the FL model with SecAgg used for our case study were not successful, most likely due to the high effective batch size of 600. Consistent with DP's privacy guarantees, the attacks were ineffective when DP training was used. Results showing that DP negates the attack even when the model is attacked locally or when SecAgg is not used are shown in Supplementary Section 5 and Supplementary Fig. 2.

To further underline the high risk of privacy-centred attacks in the healthcare imaging setting and thus the importance of privacy-enhancing techniques for collaborative model training, we performed additional experiments on the publicly available MedNIST dataset and were able to recover images disclosing sensitive patient attributes when DP was not utilized. No images could be recovered with DP in place (Fig. 5). Further details on the attack and the statistical evaluation can be found in the Methods and Supplementary Sections 4 and 6.

Discussion

We've presented PriMIA, an open-source framework for privacy-preserving FL and encrypted inference on medical images. We've demonstrated the decentralized collaborative training of an expert-level deep convolutional neural network in the challenging clinical task of paediatric chest radiography classification. Further, we've showcased end-to-end encrypted inference, which can be leveraged for secure diagnostic services without the disclosure of confidential data or exposure of the model. Our work serves

Table 1 | Classification performance comparison of models on the validation set and test sets 1 and 2

	Accuracy			Sensitivity/specificity			ROC-AUC			MCC		
	Val	Test 1	Test 2	Val	Test 1	Test 2	Val	Test 1	Test 2	Val	Test 1	Test 2
Federated DP−/SecAgg−	0.89	0.89	0.90	0.95	0.88	0.90	0.92	0.92	0.93	0.84	0.84	0.85
				0.86	0.88	0.88						
				0.86	0.94	0.93						
Federated DP−/SecAgg+	0.88	0.88	0.89	0.98	0.88	0.89	0.90	0.92	0.92	0.83	0.83	0.83
				0.86	0.88	0.88						
				0.78	0.91	0.91						
Federated DP+/SecAgg+	0.85	0.85	0.84	0.97	0.87	0.86	0.89	0.88	0.87	0.78	0.76	0.77
				0.76	0.81	0.83						
				0.82	0.85	0.86						
Centrally trained	0.92	0.90	0.91	0.96	0.90	0.93	0.93	0.93	0.94	0.87	0.85	0.87
				0.90	0.88	0.89						
				0.87	0.94	0.92						
Personalized 1	0.89	0.67	0.63	0.90	0.96	1.00	0.92	0.72	0.71	0.83	0.48	0.47
				0.88	0.19	0.25						
				0.88	0.71	0.65						
Personalized 2	0.87	0.69	0.58	0.88	0.85	0.91	0.90	0.74	0.67	0.80	0.51	0.37
				0.85	0.65	0.29						
				0.87	0.41	0.50						
Personalized 3	0.87	0.68	0.66	0.86	0.68	1.00	0.90	0.75	0.79	0.80	0.50	0.48
				0.90	0.79	0.72						
				0.84	0.53	0.00						
Expert 1	-	0.79	-	-	0.96	-	-	-	-	-	0.70	-
					0.47							
					0.88							
Expert 2	-	0.79	-	-	0.96	-	-	-	-	-	0.68	-
					0.84							
					0.41							

Federated, model trained with federated learning; DP+/−, model trained with (+) or without (−) DP gradient descent; SecAgg+/−, model trained with (+) or without (−) SecAgg; Centrally trained, model trained on the entire dataset on a single machine. Personalized 1–3, models trained only on the data owner's local data set. Expert 1/2, human experts. Sensitivity/specificity metrics refer to normal/bacterial/viral, respectively.

as the first step towards the implementation of next-generation privacy-preserving methods in medical imaging workflows. It applies to both multi-institutional research and to enterprise model development settings, allowing the preservation of data governance and sovereignty over confidential patient health data. Our framework can be used in inference-as-a-service scenarios in which diagnosrsquo support can be provided remotely with theoretical and empirical guarantees of privacy, confidentiality and asset protection. PriMIA represents a targeted evolution of our previous work[17] towards healthcare-sector-focused deployment. Although we focused on a classification task for the presented case study, PriMIA is highly adaptable to a variety of medical imaging analysis workflows employing different network architectures, datasets and more. We present an additional case study focused on semantic segmentation in computed tomography scans of the abdomen in Supplementary Section 7 and Supplementary Fig. 3, to demonstrate this flexibility.

Model classification performance. Recent work has evaluated the ramifications of data quality (overly homogeneous/independent and identically distributed data versus overly heterogeneous data) and distributed system topology on federated model performance, for example generalization to out-of-sample data. In our case study, models trained with FL performed on par with the centrally trained model similar to ref. [5] and outperformed human observers. Models trained only on subsets of the data (personalized models) showed drastically diminished performance on out-of-sample data. Since personalized model training is the standard in most mono-centric medical imaging studies, this finding serves as a reminder that the inclusion of larger quantities of more diverse data from multiple sources enabled through FL can allow the training of models with better generalization performance, as is demanded by current best practices[33]. DP model training is able to offer objective privacy guarantees and resilience against model inversion attacks[30,32]. The utilization of DP diminished model performance, which was, however, still on par with human observers. At the same time, the DP guarantees achieved ($\epsilon=6$) by the selected model are only moderate. This phenomenon (privacy–utility trade-off) is a well-known observation in the still nascent area of deep learning with DP. For instance, previous work[23] reached an ϵ-value of approximately 8 on the CIFAR-10 dataset and another study reported[34] ϵ-values between 6.9 and 8.48. Both studies also report a diminished performance by the final model. We regard methods to improve the training of DP models as a promising direction for future research.

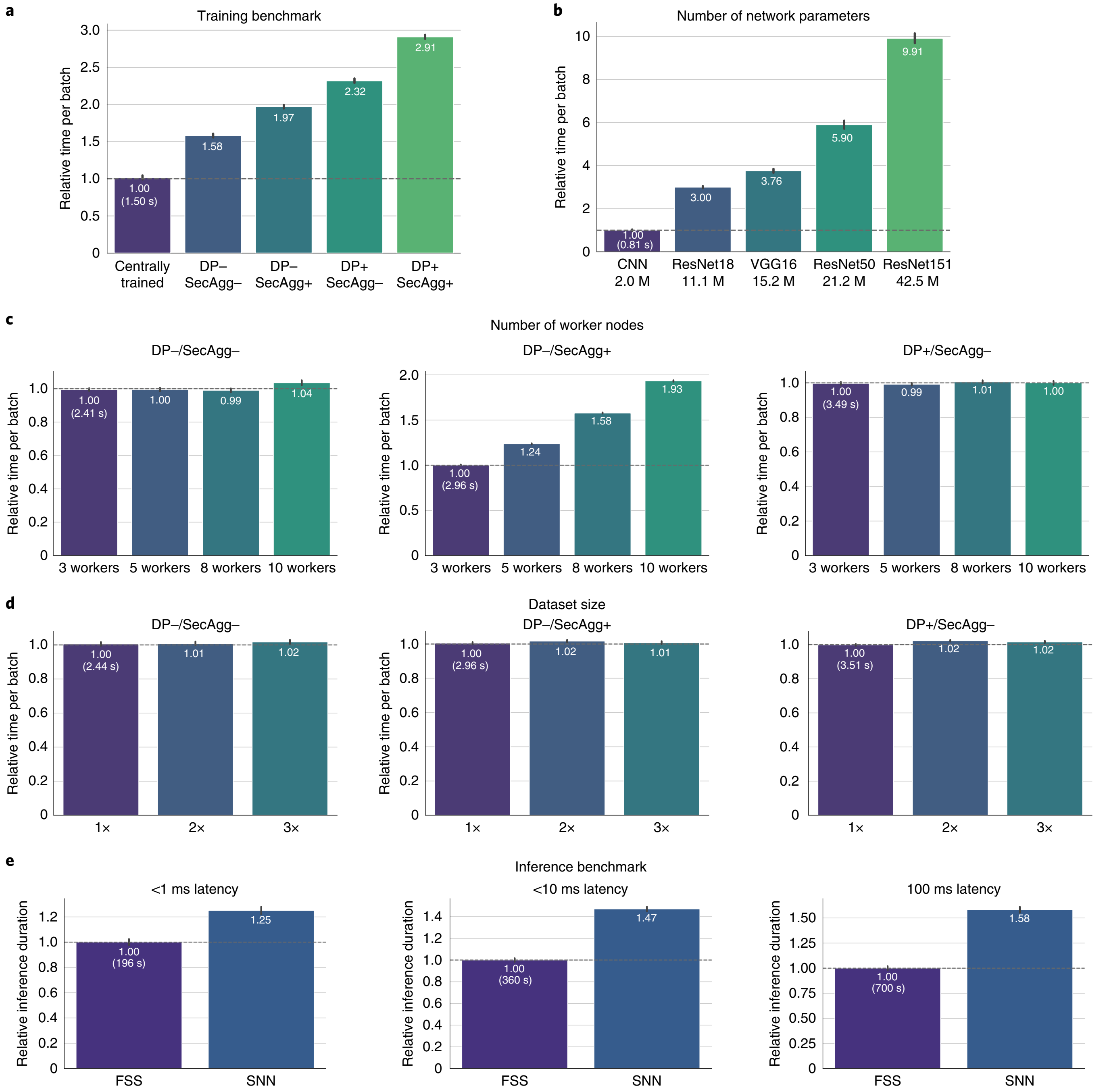


Fig. 3 | Results of training and inference benchmarks. a–d, Timing benchmarks in the training phase. All times shown in white are relative to the baseline for a batch size of 8 at a constant synchronization rate of 1 averaged over 100 runs. For DP, a microbatch size of 1 was used. The baseline is provided in parentheses. Bars denote standard deviation. Centrally trained: local training. DP+/− and SecAgg+/−: with/without DP gradient descent/SecAgg. **a**, Training latency for local training in various scenarios. **b**, The influence of neural network model parameters. Models shown: CNN architecture included with PriMIA (2.0 million parameters), ResNet18 (11.1 million parameters), VGG16 (15.2 million parameters), ResNet50 (21.2 million parameters) and ResNet151 (42.5 million parameters). **c**, The influence of the number of workers (data owners) in the federation. **d**, The influence of the dataset size per worker between one (1×) and three (3×) times the amount of data. As times shown are per batch, timings are independent of dataset size. **e**, Timing benchmark in the inference phase. FSS, function secret sharing-based inference (ours). SNN, SecureNN protocol[29]. 100 repetitions each. Latency, average 10-round-trip ping latency.

Functional improvements to FL. To increase framework usability and flexibility as well as FL model performance, our framework includes the following functional improvements. (1) Besides incorporating adaptive client optimization in the form of the Adam optimizer recently shown to yield improved convergence results[35], we include a wide range of advanced image augmentation techniques including MixUp, which has been shown to encompass privacy-enhancing attributes[36]. (2) We implement techniques to address imbalances in data volume between nodes (local early stopping), as well as between dataset classes (class-weighted gradient descent and federated averaging[37]). (3) We include facilities to carry out centrally coordinated hyperparameter optimization

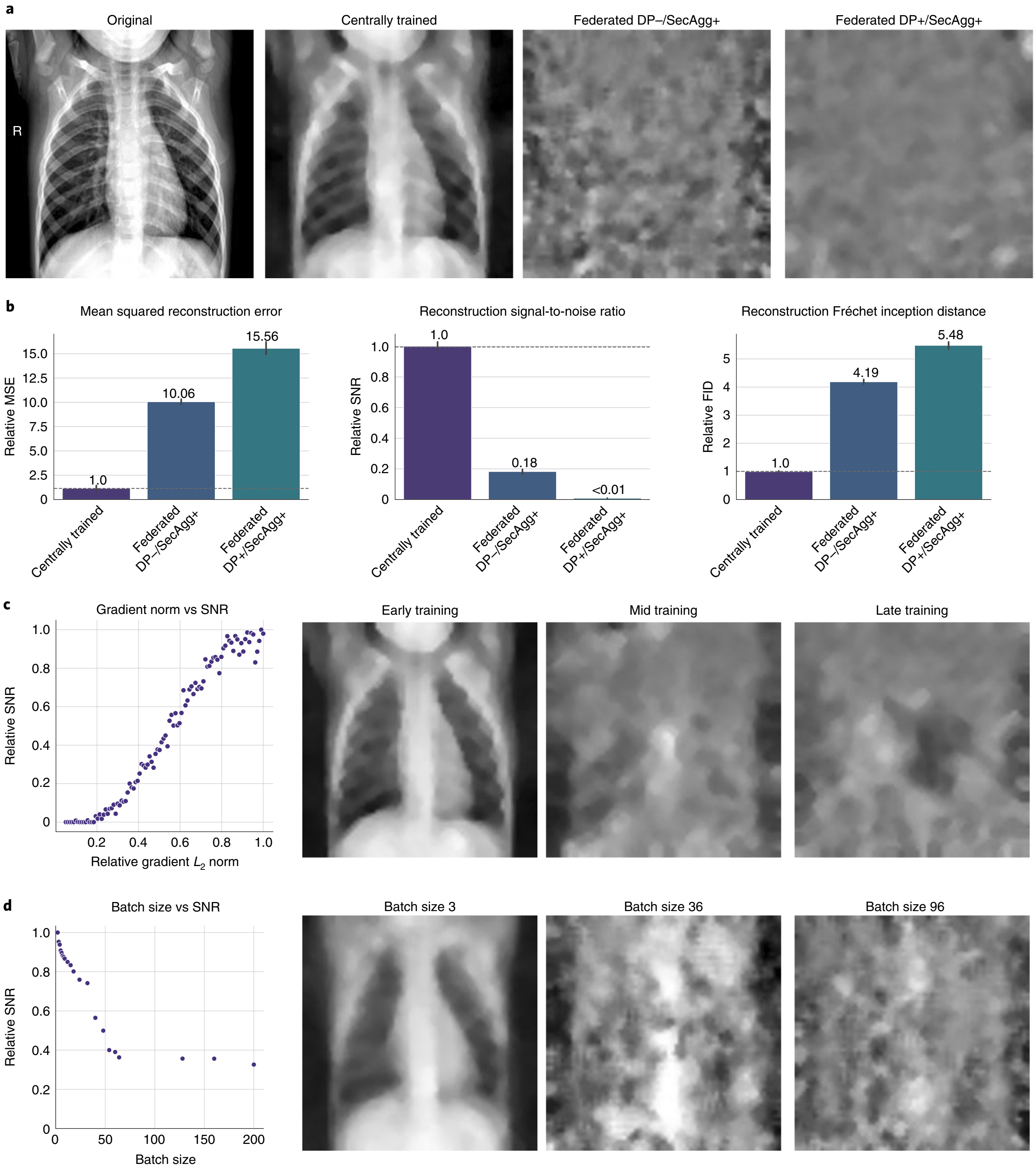


Fig. 4 | Overview of the gradient-based privacy attacks against PriMIA using the paediatric pneumonia dataset. a, Left to right: the target image (original); best-case reconstruction derived from attacking the centrally trained model early during training with a batch size of 1; typical case of an attack against the FL model trained with SecAgg (effective batch size 600, epoch 5 of 20); worst-case attack performed against a model trained with DP. **b**, Normalized metrics of attack success. Lower values for pixel-wise MSE and FID (mirroring human perception of similarity) and higher values for signal-to-noise ratio indicate increased success, respectively. **c**, Attack success, measured as relative signal-to-noise ratio dependent on the model's global L_2-norm. As training progresses, loss decreases and thus the gradient norm diminishes, reducing attack success. **d**, The influence of effective batch size on attack success measured as relative signal-to-noise ratio. High batch sizes substantially impede attack success. Chest radiographs from Mendeley Data[67].

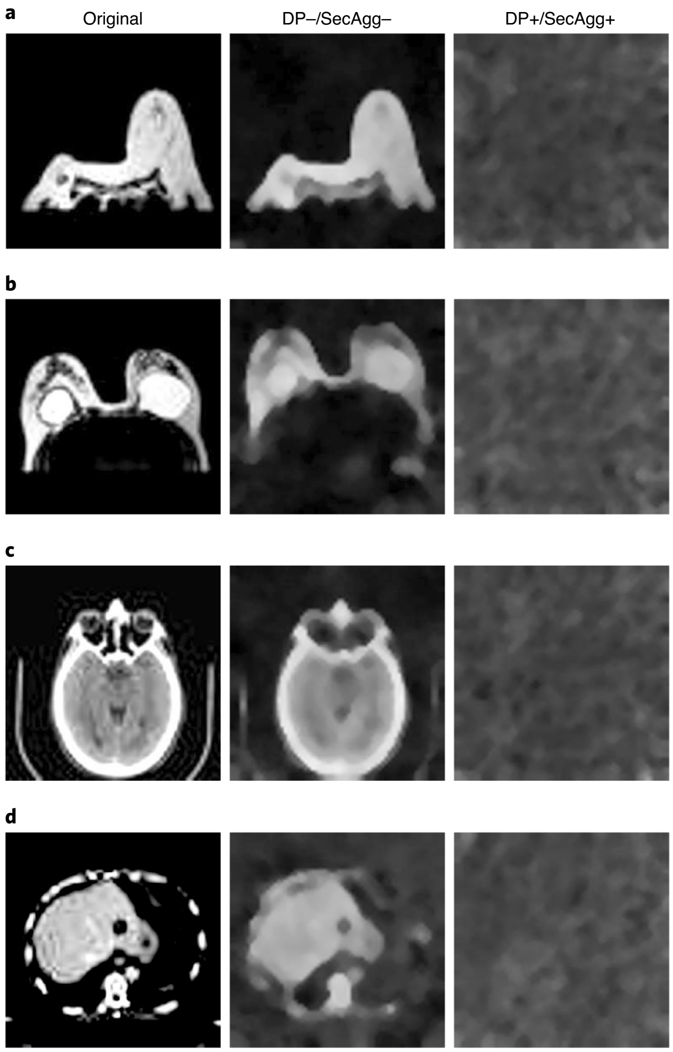


Fig. 5 | Overview of the gradient-based privacy attacks against PriMIA using the MedNIST dataset in a variety of scenarios. The original image is shown (original) alongside the reconstruction results from a model trained without secure aggregation or DP (DP–/SecAgg–) as well as a model trained with DP and SecAgg (DP+/SecAgg+). In every case, the attack reveals confidential information about the patient when the model is trained without privacy-enhancing techniques. **a**, Breast MRI revealing absence of the right breast, likely due to operative removal due to breast cancer. **b**, Breast MRI revealing breast implants. Both **a** and **b** also allow assumptions about the patient's sex. **c**, Cranial computed tomography image at the level of the nose. Facial contours reconstructed from such images can lead to personal identification[39]. **d**, Abdominal CT at the level of the liver, allowing visualization of a hypodense lesion in the left liver lobe in the reconstructed image. In every case, using DP thwarts the attack, disallowing any usable image features from being visualized. CT images licensed under the Creative Commons CC BY-SA 4.0.

over the entire confederation using the Tree-Structured Parzen Estimator algorithm[38]. Experimental data showcasing the utilization of our hyperparameter selection framework to search for the optimal FL model can be found in Supplementary Section 1 and Supplementary Fig. 1. All above-mentioned training optimizations are implemented locally on the nodes and do not negatively impact privacy guarantees. Hyperparameter tuning, however, must be considered when DP is utilized, as it relies on multiple training repetitions.

Discussion on privacy-enhancing techniques. The inclusion of methods offering provable privacy and security guarantees in the FL process is a crucial step towards the widespread implementation of privacy-preserving AI technologies[8]. The successful reconstruction of images from unprotected models in our attack experiments underline the risks of such attacks to patient privacy, which has also been discussed in previous work[6,39]. DP training provides objective privacy guarantees in case of attacks against the model both by confederation members and during inference and is not limited to the gradient-based inversion attack we use in our example. SecAgg utilizing SMPC only discloses the aggregate model update to the parties, even in case up to $n-1$ out of n parties collude to reveal data. The DP secure aggregation of dataset statistics (means and standard deviations) we propose can protect FL participants from data leakage, especially when non-imaging data is included in model building (for example clinical records, in which the means of features such as age represent sensitive information). Finally, encrypted inference reveals no information about the data or the model to either party.

Compared with fully homomorphic encryption protocols[40] relying on key-based cryptography, whose implementation for neural network training and inference is impeded by the computational complexity of the encryption process and the performance decrease due to function approximation for for example activation functions, communication overhead has traditionally been the limiting factor for SMPC. In our recent work, we introduced AriaNN[26], an SMPC protocol leveraging function secret sharing (FSS)[28] and building upon SPDZ[25]. It represents an alternative to protocols like SecureNN[29] or Falcon[41], and computes private comparisons with a single round of communication. This renders FSS substantially more communication-efficient than other SMPC protocols, especially when parties are geographically distant and communicate with high latency, for example when performing inference over the public web as showcased in our study. Through the present use-case, we confirm the results obtained in our previous work on other datasets: secure inference gains proportionally greater benefits from the FSS protocol in the high-latency setting. Thus, we propose its utilization over SecureNN in cases a reduction in latency is desired in an honest-but-curious setting.

Comparison to prior work. Several current works aim to introduce PPML techniques to biomedical imaging: Silva et al.[42] present a front-end FL framework for biomedicine, but do not consider DP, SecAgg or encrypted inference. Xu and colleagues (https://bit.ly/3pl5dD1) provide a framework for FL using homomorphic encryption for SecAgg, but do not utilize DP or provide encrypted inference capabilities. Sheller et al.[43] showcase an FL use-case based on segmentation. They do not assess either DP, SecAgg or the option for encrypted inference. Li et al.[44] also demonstrate an FL segmentation task. Their DP implementation relies on an alternative technique (sparse vector) and the framework does not provide secure aggregation or encrypted inference. The work by Lu and colleagues[45] demonstrates FL with DP, however their use-case is focused around pathology slides and does not employ SecAgg or provide encrypted inference capabilities. Li et al.[46] utilize DP, however assume a fixed sensitivity and do not conduct privacy analysis. Their framework does not offer SecAgg or encrypted inference.

Limitations. We consider the following limitations of our work. The computational requirements for deploying our system are substantial, and the latency resulting from encrypted inference is still very high compared to unencrypted inference, despite the proposed protocol improvements. The underlying remote execution environment currently offers experimental graphics processing unit (GPU) support, with full support planned for an upcoming version. The success of FL models is largely dependent on high data quality on the nodes. The auditing and curation of the data and its quality, methods to quantify the contribution of individual datasets to the model or to detect local overfitting are still under investigation[47]. Our library is designed to be used in an honest-but-curious regime, which we believe to represent the standard in healthcare consortia. Thus, although we provide comprehensive privacy protection measures, we included no specific countermeasures against malicious contributions of low-quality or adversarial data to the FL process or to verify/guarantee to the data owner that the model used in the inference setting is the one promised. Furthermore, we point out that discussions of the theoretical threat model are a level of abstraction that cannot fully represent the complexity of real-life situations. For instance, threat modelling is typically undertaken on the level of FL participants representing entire hospitals, however this cannot take every individual person working for these hospitals and their specific motivations into account. Similarly, questions about participant reimbursement or model ownership in FL were outside the scope of our current investigation. Further studies in this developing field are required to fully illuminate such details. Lastly, as mentioned above, the utilization of DP causes a direct trade-off between model privacy and utility. Future work will need to address this trade-off through improved privacy analysis and training techniques, as the privacy guarantees of current studies, including the ϵ-value of around 6.0 seen in our study, are not yet sufficiently rigorous to be considered generally applicable.

Conclusion

We present a free, open-source software framework for privacy-preserving FL and end-to-end encrypted inference on medical imaging data, which we showcase in a clinically relevant real-life case study. Further research and development will enable the larger-scale deployment of our framework, the validation of our findings on diverse cross-institutional data, and further the widespread utilization of PPML techniques in healthcare and beyond.

Methods

Dataset collection. For model training, we used the previously proposed paediatric pneumonia dataset[20]. The dataset was reviewed by a specialist radiologist for image quality and representativeness and included 5,163 training images in the above-mentioned three categories, as well as a validation set of 624 images. For FL model development, the training set was randomly subsampled into three equally sized non-overlapping partitions. Class balance between nodes was not enforced.

For model testing on unseen data, we retrospectively collected 497 chest radiographs of the same classes of an age-matched cohort from two university hospitals (test set 1: 145 images (43 bacterial, 68 normal, 34 viral), test set 2: 352 images (120 bacterial, 126 normal, 106 viral)). Ethics committee and data protection votes for data collection and exchange were granted by all institutions waiving the requirement for informed consent in this retrospective study (protocol number 111/20 S-KH). All procedures were carried out in accordance with clinical best practices, applicable laws and regulations as well as the Declaration of Helsinki. Ground-truth labels for the dataset were generated from clinical records based on validated laboratory results and clinical parameters (c-reactive protein (CRP), body temperature, antibiotic response for bacterial, sputum or sweat polymerase chain reaction (PCR) and/or absence of bacterial infection signs for viral) as well as clinical assessment of specialist paediatricians/neonatologists not involved in image evaluation.

Model training. *Privacy-preserving processing of dataset statistics.* For the training of neural networks, data is typically pre-processed by mean subtraction and division by the standard deviation. In federated learning, dataset statistics from the local nodes or aggregated statistics from all nodes can be used. Additionally, the provision of the final model in an inference setting requires these statistics for rescaling incoming images. However, dataset statistics can contain private information that should not be shared, especially in case non-imaging data is included (for example, age in the case of clinical record data). Hence, we propose and implement differentially private secure aggregation of dataset statistics. Here, sensitivity-calibrated Laplacian noise is added to the statistics to satisfy a user-defined ϵ DP value before SMPC is used to average them, and they are then stored on the central server for later use. Before inference starts, the data is rescaled with the (differentially private) securely aggregated mean and standard deviation of the training set. For training, the nodes use their local dataset statistics. Thus, data leakage is prevented, especially in the case individual nodes contain few, or just one, dataset(s).

Model architecture, hyperparameters and augmentation. We used the ResNet18 architecture[19], pretrained on ImageNet[48], with the final average pooling layer replaced by a single linear layer with 512 units and randomly initialized with the Kaiming Uniform initializer[49]. Images were cropped to squares such that the entire chest section of the radiograph is preserved and resized to 224×224 pixels.

The following standard augmentation techniques were employed: random horizontal flips, random affine transformations, Gaussian noise injection. In extension, we used the Albumentations library[50] to apply the following transformations: random changes in the gamma value and brightness, blurring, optical distortions, grid shuffles/dropouts/distortions, elastic transforms, changes in hue-saturation-value (HSV) colour space, inverting images, cutouts of the image, artificial shadows, fog, solarizations and sun flares. We also provide the option for histogram equalization or contrast-limited adaptive histogram equalization (CLAHE), both as an augmentation and a standardization technique. The individual augmentations were introduced with a probability p_1 and augmentation was activated overall with a probability p_2. Furthermore, we applied a modified variant of MixUp augmentation[51] by which the mixing parameter (λ) is randomly sampled from a uniform distribution similar to that in ref. [36].

Training was performed for 40 epochs using the Adam optimizer[52] with a log-linearly decreasing learning rate initially set at 10^{-4}. PriMIA caches models automatically after each round, and selects the model with the highest validation set Matthews correlation coefficient (MCC). The centralized model was trained by pooling all data on a single machine and training the model on the accumulated dataset. Personalized models were trained on the respective nodes using only the local dataset. PriMIA implements the ability to carry out centrally coordinated automated hyperparameter tuning on the entire federation or locally, which was used to determine the best model in every case according to highest validation set MCC. An example is provided in Supplementary Section 1 and Supplementary Fig. 1. Model hyperparameters are centrally set for all nodes, but image augmentation, local early stopping and weighted gradient descent are performed locally and independently on the nodes. Federated training and inference experiments were conducted over the public Internet on cloud instances with 32 CPU cores at 3.1 GHz and 64 GB of random access memory (RAM). Centralized model training was performed on a server with 36 CPU cores at 2.4 GHz and 512 GB of RAM.

Differentially private model training. DP model training entails several additional considerations. We describe these alongside PriMIAs DP implementation and the process of training the final DP model at length in Supplementary Section 8. In brief, PriMIA implements DP gradient descent[23] based on clipping the gradient L_2-norm of each individual sample, then adding calibrated Gaussian noise. This process occurs on each node independently with independent noise sources (local DP). We considered the paediatric pneumonia dataset private, therefore did not perform hyperparameter optimization based on multiple training runs. Furthermore, due to the relatively small size of the dataset, we determined it would not be possible to train the model with sufficient utility while maintaining acceptable privacy guarantees. Hence, we used the pre-training technique described previously[23] and employed a publicly available dataset trained on a related task to determine the optimal parameters for the DP mechanism and pre-train the model. Details can be found in Supplementary Section 8.2.2 and Supplementary Fig. 4.

Training topology, gradient descent and secure aggregation. We selected the hub-and-spoke system topology due to its reported improved final model performance over techniques such as incremental or cyclical training[5,43] and its higher flexibility with respect to node availability and asynchronous training[53]. In PriMIA training is carried out asynchronously in rounds. Initially, the model is sent from the central server to all computation nodes. During each round, nodes locally perform a variant of gradient descent in which gradient updates are weighted inversely by the frequency of the individual dataset classes present on the node (class weighted gradient descent). After a number of batches (denoted by σ) have been processed on every node, the updated models are securely averaged (SecAgg[54]) using the FSS SMPC protocol (see below), before being distributed back to the nodes. For model averaging, we utilize class-weighted federated averaging[37] whereby the central model updates are weighted by the class frequency on the nodes before a new training round begins.

Model synchronization and the σ parameter. Previous work has investigated the federated synchronization rate parameter (σ) as central in controlling network input/output and training duration[55]. We found the choice of this parameter to also affect model performance and training time, and it has recently been described as an important open research target in FL with respect to the optimal trade-off between model accuracy and training time[47]. We provide further details on these findings in Supplementary Section 10 and Supplementary Fig. 6.

Measures against FL training deterioration. Literature findings and our own evidence indicate that, in case one of the federation's nodes contains less data than others, continuing training beyond convergence until other nodes have completed training can lead to overfitting or training collapse. Alternatively, not including the updates from this node can lead to catastrophic forgetting[56] of the node's data and reduced generalization performance. We empirically determined that local early stopping, that is, terminating training on the local node once the node's local dataset is exhausted, then using the state of the node's local model for all future update steps until a full round of training is completed, led to improved training performance.

Secure multi-party computation protocols. *Function secret sharing.* FSS belongs to the family of SMPC protocols, in which several parties share a secret (for example, data or a model) to ensure privacy. A party alone holds a random share of the private value and cannot reconstruct the value on their own. A quorum of parties (sometimes all parties) need to collaborate to reconstruct the private data. The terms encrypted and obfuscated are used interchangeably in this scenario to denote secret-shared data.

Unlike classical data secret sharing schemes like SecureNN[29], where a shared input $\llbracket x \rrbracket$ is applied on a public function *f*, FSS applies a public input *x* on a private shared function $\llbracket f \rrbracket$. Shares or *keys* ($\llbracket f \rrbracket_0$, $\llbracket f \rrbracket_1$) of a function *f* satisfy $f(x) = \llbracket f \rrbracket_0(x) + \llbracket f \rrbracket_1(x)$. Both approaches output a secret shared result. In our case, assume two parties respectively own shares $\llbracket y \rrbracket_0$ and $\llbracket y \rrbracket_1$ of a private input *y*, and they want to compute $\llbracket y \geq 0 \rrbracket$. They receive some cryptographic primitives (see below), namely each get a share of a random value (or *mask*) $\llbracket \alpha \rrbracket$ and a share of the shared function $\llbracket f_\alpha \rrbracket$ of $f_\alpha: x \rightarrow (x \geq \alpha)$. They first mask their shares of $\llbracket y \rrbracket$ using $\llbracket \alpha \rrbracket$, by computing $\llbracket y \rrbracket_0 + \llbracket \alpha \rrbracket_0$ and $\llbracket y \rrbracket_1 + \llbracket \alpha \rrbracket_1$ and then revealing these values to reconstruct $x = y + \alpha$. Next, they apply this public *x* on their function shares $\llbracket f_\alpha \rrbracket_{j=0,1}$, to obtain a shared output $(\llbracket f_\alpha \rrbracket_0(x), \llbracket f_\alpha \rrbracket_1(x)) = \llbracket f_\alpha(y + \alpha) \rrbracket = \llbracket (y + \alpha) \geq \alpha \rrbracket = \llbracket y \geq 0 \rrbracket$. Previous studies on FSS[57,58] have shown the existence of such function shares for comparison which perfectly hide *y* and the result. For more details about the concrete implementation of FSS we refer to our previous work[26]. SMPC and the FSS protocol provide theoretical security guarantees in the honest-but-curious regime. FSS offers high communication efficiency and can be thus employed to reduce transaction latency. FSS is based in part on the SPDZ protocol[25]. To increase efficiency for specific mathematical operations (for example multiplication) by reducing the rounds of communication required to perform the operation, protocols such as SPDZ partition encrypted operations into an offline phase, during which no communications between parties take place, and an online phase, where parties communicate. During the offline phase, a trusted third party, referred to in PriMIA as a cryptographic provider (and in ref. [25] as a trusted dealer), provides cryptographic primitives. In practice, it is not a requirement for parties to use the PriMIA cryptographic provider, as the framework can be modified to use a trusted third party of their own choosing. These primitives can be computed in advance as they require no knowledge of the exact functions evaluated during the online phase, and the cryptographic provider does not participate in the online phase in which these computations take place. A schematic representation of the two phases and further terminology are provided in Supplementary Section 9 and Supplementary Fig. 5.

Secure aggregation. The SecAgg operation, consisting of a private addition and a public multiplication is performed using the additive secret sharing scheme of the underlying SPDZ[25] protocol. The protocol is designed such that random shares are distributed between participants, which individually contain no usable information and only the sum of their contributions (that is, the aggregated model updates) are revealed. Collusion between up to $n-1$ out of n participants (in the case study, two out of three) is insufficient to disclose the other participant's private information. SecAgg is performed without a need for cryptographic primitives or the cryptographic provider.

Secure inference. Secure inference represents a transaction between two parties, by which the data owner wishes to receive the model's prediction without disclosing their data, and the model owner wishes to keep their model hidden. We adapt our previous work on AriaNN[26], based on FSS, for encrypted inference to leverage its high communication efficiency, which allows the evaluation of private comparisons with minimal communication overhead. Such comparison operations are important for example for the evaluation of maximum pooling layers or rectified linear units. The cryptographic primitives provider is again not required for the actual inference process (online phase), which occurs exclusively between the two parties. In our framework, the data owner initiates a request to the system, the data and model are obfuscated by secret sharing and inference takes place using SMPC. Secure inference scenario is thus—in the sense described above—an end-to-end encrypted transaction, whereby both the data and the model is obfuscated. This guarantees both parties single-use accountability, that is, the guarantee that the data and model can be used for no other purpose than the one explicitly designated by the involved parties.

We note that while the data enjoys information-theoretic secrecy guarantees, the party requesting inference has access to the model's predictions and can perform black-box membership inference[59] or model inversion attacks[60]. PriMIA's DP training procedure provides effective protection against such attacks[30,32,59] to the individuals whose data was used to train the model used for inference.

Classification performance assessment. Classification performance was evaluated as follows. For expert readers, accuracy, sensitivity/specificity (recall) and MCC[27] were calculated on test set 1. The model's performance was evaluated in terms of accuracy, sensitivity/specificity (recall), ROC-AUC MCC on the validation set and on both test sets. MCC was employed due to its invariance to class imbalance and its indication of prediction concordance alongside quality of classification, leading to recent recommendations for its use over the usually employed accuracy or F1-Score metrics[61]. McNemar's test was used to test for statistical significance in classification performance. Cohen's κ (kappa) was used to test inter-rater/-model agreement. Statistical significance is defined as $p < 0.05$.

Inference and training latency assessment. We compared the average ± standard deviation duration in seconds of 1 epoch of training over 100 epochs as well as the average ± standard deviation duration of one inference transaction over 100 transactions in three settings: utilizing inter-process communication locally (using the PySyft VirtualWorker abstraction (no latency), utilizing the websocket/HTTP protocol on the local network (LAN) (low latency) and utilizing the public Internet (WAN) (high latency) with a 10-round-trip ping latency of 100 ms. Student's *t*-test was used to assess statistical significance.

Model inversion utilizing gradient updates. To exemplify the susceptibility of models trained without privacy-enhancing techniques against adversarial agents that attempt to expose sensitive data, we employ the Improved Deep Leakage from Gradients, iDLG, method with modifications as proposed previously[32], itself a variant of previously shownn techniques[31,62]. iDLG was found highly successful against the ResNet18 architecture used in our case study. We additionally modified the attack following newer evidence from[63] by utilizing the AdamW optimizer and initializing images with uniform sampling to further improve its success. The overview of the attack is as follows:

1. Adversary generates a randomized pair of a dummy model update and a corresponding label
2. Adversary captures the gradient update submitted by an honest client
3. Using a suitable cost function, the adversary attempts to minimize the difference between the honest update and the dummy update
4. The algorithm is repeated until either the loss starts diverging or the final iteration is reached

In the original implementation of the protocol, the difference between gradients is calculated using

$$||\Delta W' - \Delta W||^2 = ||\frac{\delta l(F(x', W), y')}{\delta W} - \Delta W||^2$$

where x' and y' are the data point and its label respectively, while W and W' are the victim's and attacker's gradient respectively. Following Geiping et al.'s implementation, we used the cosine similarity metric and utilized images of size 224 × 224, as authors show that this is the upper bound for acceptable reconstruction quality[32]. The empirical evaluation of various batch sizes showed that larger batch sizes drastically reduce the success of the reconstruction. We indicate an averaged model update from *n* parties each trained with a batch size of *k* to have been trained with an effective batch size of $n \times k$. Our observation matches ref. [32] which shows batch sizes above eight to substantially deteriorate the attack. We furthermore found the L_2-norm of the gradient update to strongly influence attack success. Thus, attacks at the beginning of training, when the loss (and thus the gradient with respect to it) is largest, were most successful. A low MSE value did not always signify a successful attack, since a specific model update can be generated by more than one image, resulting in noise that is able to mimic the update, but not the corresponding data. To improve attack evaluation, we also supply signal-to-noise ratio and perceptual metrics which more robustly assess the reconstruction quality and human perception of image similarity as performed in[32,64–66]. As an active attack, iDLG can be executed by an adversarial client or central server. We note that in the case of an adversarial central server, the usage of SMPC prevents the disclosure of individual model updates, therefore only allowing the adversary to utilize averaged model updates instead. For the attacks on the FL system we assumed that one out of three data owners is an adversary. For the 'baseline' attack on the centralized model, we used a batch size of 1. Attacks were performed against 100 randomly selected images from the training set. For the gradient norm experiments, 100 gradient samples were taken at equispaced intervals during model training. Batch size experiments were carried out under identical circumstances only varying batch size. Model and dummy image initialization was deterministically set for all experiments. Each attack was performed in triplicate with at most 24,000 iterations per run and the instance with the highest cosine similarity was selected. One way analysis of variance (ANOVA) followed by the Student's *t*-test were used to assess statistical significance between the MSE, SNR and FID scores. Details of the attack against the MedNIST dataset can be found in Supplementary Section 6.

Data availability

The paediatric pneumonia dataset is publicly available from Mendeley Data at https://doi.org/10.17632/rscbjbr9sj.3. The MedNIST dataset was assembled by B. J. Erickson (Department of Radiology, Mayo Clinic) and is available at https://github.com/Project-MONAI/MONAI/. The MSD Liver Segmentation Dataset is available at http://medicaldecathlon.com. test sets 1 and 2 contain confidential patient information and cannot be shared publicly. Source data are provided with this paper.

Code availability

The current version of the PriMIA source code is publicly available at https://github.com/gkaissis/PriMIA and permanently archived at https://doi.org/10.5281/zenodo.4545599[18]. PriMIA includes source code from PySyft (https://github.com/OpenMined/PySyft), PyGrid (https://github.com/OpenMined/PyGrid) and Opacus (https://github.com/pytorch/opacus) re-used under open-source licence terms.

Received: 4 October 2020; Accepted: 26 March 2021;
Published online: 24 May 2021

References

1. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
2. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
3. Patent Index 2019: Spotlight on digital technologies. *European Patent Office* https://www.epo.org/about-us/annual-reports-statistics/statistics/2019.html (accessed 10 March 2021).
4. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
5. Sheller, M. J. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 12598 (2020).
6. Schwarz, C. G. et al. Identification of anonymous MRI research participants with face-recognition software. *N. Engl. J. Med.* **381**, 1684–1686 (2019).
7. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* 111–125 (IEEE, 2008).
8. Price, W. N. & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
9. Banerjee, S., Hemphill, T. & Longstreet, P. Wearable devices and healthcare: data sharing and privacy. *Inf. Soc.* **34**, 49–57 (2018).
10. Raisaro, J. L. et al. SCOR: a secure international informatics infrastructure to investigate COVID-19. *J. Am. Med. Inform. Assoc.* **27**, 1721–1726 (2020).
11. Vaid, A. et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. *JMIR Med. Inform.* **9**, e24207 (2021).
12. Roth, H. R. et al. Federated learning for breast density classification: a real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning* 181–191 (Springer, 2020).
13. Fredrikson, M., Jha, S. & Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS 2015)* (ACM Press, 2015).
14. Wang, Z. et al. Beyond inferring class representatives: user-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications* 2512–2520 (IEEE, 2019).
15. La, H. J., Kim, M. K. & Kim, S. D. A personal healthcare system with inference-as-a-service. In *2015 IEEE International Conference on Services Computing* 249–255 (IEEE, 2015).
16. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
17. Ryffel, T. et al. A generic framework for privacy preserving deep learning. Preprint at https://arxiv.org/abs/1811.04017 (2018).
18. Kaissis, G. & Ziller, A. PriMIA version 2021.02 https://doi.org/10.5281/zenodo.4545599 (2021).
19. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
20. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).
21. Gupta, G. R. Tackling pneumonia and diarrhoea: the deadliest diseases for the world's poorest children. *Lancet* **379**, 2123–2124 (2012).
22. Evans, D., Kolesnikov, V. & Rosulek, M. A pragmatic introduction to secure multi-party computation. *Found. Trends Privacy Secur.* **2**, 70–246 (2018).
23. Abadi, M. et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2016).
24. Mironov, I., Talwar, K. & Zhang, L. Rényi differential privacy of the sampled gaussian mechanism. Preprint at https://arxiv.org/abs/1908.10530 (2019).
25. Damgård I., Pastro V., Smart N. & Zakarias S. Multiparty computation from somewhat homomorphic encryption. In *Advances in Cryptology – CRYPTO 2012* (eds. Safavi-Naini, R. & Canetti R.) (Springer, 2012).
26. Ryffel, T., Pointcheval, D. & Bach, F. ARIANN: low-interaction privacy-preserving deep learning via function secret sharing. Preprint at https://arxiv.org/abs/2006.04593 (2020).
27. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct.* **405**, 442–451 (1975).
28. Boyle, E., Gilboa, N. & Ishai, Y. Function secret sharing. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques* 337–367 (Springer, 2015).
29. Wagh, S., Gupta, D., & Chandran, N. Securenn: 3-party secure computation for neural network training. In *Proc. Privacy Enhancing Technologies* 26–49 (Sciendo, 2019).
30. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J. & Song, D. The secret sharer: evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)* 267–284 (2019).
31. Zhao, B., Mopuri, K. R. & Bilen, H. iDLG: improved deep leakage from gradients. Preprint at https://arxiv.org/abs/2001.02610 (2020).
32. Geiping, J., Bauermeister, H., Dröge, H. & Moeller, M. Inverting gradients. How easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems* 16937–16947 (NeurIPS, 2020).
33. Bluemke, D. A. et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the radiology editorial board. *Radiology* **294**, 487–489 (2020).
34. Wu, B. et al. P3SGD: patient privacy preserving SGD for regularizing deep CNNs in pathological image classification P3SGD. In *Proc. Conference on Computer Vision and Pattern Recognition* 2099–2108 (CVPR, 2019).
35. Reddi, S. et al. Adaptive federated optimization. Preprint at https://arxiv.org/abs/2003.00295 (2020).
36. Fu, Y., Wang, H., Xu, K., Mi, H. & Wang, Y. Mixup based privacy preserving mixed collaboration learning. In *2019 IEEE International Conference on Service-Oriented System Engineering (SOSE)* 275–2755 (IEEE, 2019).
37. McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* 1273–1282 (PMLR, 2017).
38. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems* 2546–2554 (Curran Associates, 2011).
39. Parks, C. L. & Monson, K. L. Automated facial recognition of computed tomography-derived facial images: patient privacy implications. *J. Digital Imaging* **30**, 204–214 (2016).
40. Qaisar Ahmad Al Badawi, A. et al. Towards the AlexNet moment for homomorphic encryption: HCNN, the first homomorphic CNN on encrypted data with GPUs. In *IEEE Transactions on Emerging Topics in Computing* (IEEE, 2020).
41. Wagh, S. et al. Falcon: honest-majority maliciously secure framework for private deep learning. In *Proc. Privacy Enhancing Technologies* 188–208 (Sciendo, 2021).
42. Silva, S., Altmann, A., Gutman, B. & Lorenzi, M. Fed-BioMed: a general open-source frontend framework for federated learning in healthcare. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning* (eds Albarqouni, S. et al.) 201–210 (Springer, 2020).
43. Sheller, M. J., Reina, G. A., Edwards, B., Martin, J. & Bakas, S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* 92–104 (Springer, 2019).
44. Li, W. et al. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging* 133–141 (Springer, 2019).
45. Lu, M. Y. et al. Federated learning for computational pathology on gigapixel whole slide images. Preprint at https://arxiv.org/abs/2009.10190 (2020).
46. Li, X. et al. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Med. Image Anal.* **65**, 101765 (2020).
47. Kairouz, P. & McMahan, H. B. *Advances and Open Problems in Federated Learning* (Now, 2021).
48. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR09)* (2009).
49. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision* 1026–1034 (2015).
50. Buslaev, A. et al. Albumentations: fast and flexible image augmentations. *Information* **11**, 125 (2020).
51. Huang, L., Zhang, C. & Zhang, H. Self-adaptive training: beyond empirical risk minimization. In *Advances in Neural Information Processing Systems* Vol. 33 (NeurIPS, 2020).

52. Kingma, P. & Ba, J. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representations* (ICLR, 2015).
53. Rieke, N. et al. The future of digital health with federated learning. *npj Digital Med.* **3**, 119 (2020).
54. Bonawitz, K. et al. Practical secure aggregation for federated learning on user-held data. In *NIPS Workshop on Private Multi-Party Machine Learning* (NIPS, 2016).
55. Wang, S. et al. Adaptive federated learning in resource constrained edge computing systems. *IEEE J. Sel. Areas Commun.* **37**, 1205–1221 (2019).
56. Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl Acad. Sci. USA* **114**, 3521–3526 (2017).
57. Boyle, E., Gilboa, N. & Ishai, Y. Function secret sharing: improvements and extensions. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* 1292–1303 (2016).
58. Boyle, E., Gilboa, N. & Ishai, Y. Secure computation with preprocessing via function secret sharing. In *Theory of Cryptography Conference* 341–371 (Springer, 2019).
59. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* 3–18 (IEEE, 2017).
60. He, Z., Zhang, T. & Lee, R. B. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference* (ACM, 2019).
61. Chicco, D. & Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
62. Zhu, L., Liu, Z. & Han, S. Deep leakage from gradients. In *Advances in Neural Information Processing Systems* 14774–14784 (2019).
63. Wang, Y. et al. SAPAG: a self-adaptive privacy attack from gradients. Preprint at https://arxiv.org/abs/2009.06228 (2020).
64. Oh, H. & Lee, Y. Exploring image reconstruction attack in deep learning computation offloading. In *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications: EMDL '19* (ACM, 2019).
65. Gao, W. et al. Privacy-preserving collaborative learning with automatic transformation search. In *Proc. Conference on Computer Vision and Pattern Recognition* (CVPR, 2021).
66. Yanchun, L. & Nanfeng, X. Generative adversarial networks based on denoising and reconstruction regularization. In *2019 IEEE 21st International Conference on High Performance Computing and Communications IEEE 17th International Conference on Smart City IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)* (IEEE, 2019).
67. Kermany, D., Zhang, K. & Goldbaum, M. Large dataset of labeled optical coherence tomography (OCT) and chest X-ray images. *Mendeley Data* https://doi.org/10.17632/rscbjbr9sj.3 (2018).

Acknowledgements

We acknowledge funding from the following sources, funders played no role in the design of the study, the preparation of the manuscript or the decision to publish. The Technical University of Munich, School of Medicine Clinician Scientist Programme (KKF), project reference H14 (G.K.). German Research Foundation, SPP2177/1, German Cancer Consortium (DKTK) and TUM Foundation, Technical University of Munich (R.B. and G.K.). European Community Seventh Framework Programme (FP7/2007-2013 grant no. 339563 – CryptoCloud) and FUI ANBLIC Project (T.R.). UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare (D.R. and G.K.). Technical University Munich/ Imperial College London Joint Academy of Doctoral Studies (D.U.). We thank B. Farkas for creating Figures 1 and 2 as well as the PriMIA logo; P. Cason and H. Emanuel for assisting with PyGrid debugging, M. Lau for his input, D. Testuggine for his input on differentially private gradient descent, M. Jay for helping with PySyft debugging, N. Remerscheid for his work on the liver segmentation case study, the PySyft and PyGrid development teams for their foundational work and the OpenMined community for their scientific input, contributions and discussion.

Author contributions

G.K. conceived and coordinated the project, evaluated the chest radiography data, helped with PriMIA programming and wrote the paper. A.Z. conceived and developed PriMIA, oversaw PriMIA development, trained the models, performed data analysis and wrote the paper. J.P.-P. helped with project management, oversaw the security and cryptography aspects of PriMIA, supervised model inversion attacks and helped write the paper. T.R. designed and developed PySyft and PyGrid, designed and implemented the AriaNN FSS protocol, helped with PriMIA programming and performed inference latency assessment. D.U. performed the model inversion attacks and helped write the paper. A.T. conceived the OpenMined project, designed and developed PySyft and PyGrid, provided project guidance and assistance and helped with PriMIA development. I.D.L.C.J. developed PyGrid and helped with PriMIA programming. J.M. provided project guidance and prototype code. F.J. performed data curation and helped with data analysis on the chest radiographs. M.-M.S. performed data curation and evaluated the chest radiography data. A.S. and M.M. provided project management, support and guidance. D.R. provided oversight, project management, support, guidance and scientific input, and helped write the paper. R.B. provided oversight, project management, support and guidance, helped with data procurement, provided scientific input and helped write the paper. All authors proof-read and accepted the final version of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00337-8.

Correspondence and requests for materials should be addressed to R.B.

Peer review information *Nature Machine Intelligence* thanks Haixu Tang, Holger Roth and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.2 Medical imaging deep learning with differential privacy

Synopsis: The successful training of deep learning models for diagnostic deployment in medical imaging applications requires large volumes of data. Such data cannot be procured without consideration for patient privacy, mandated both by legal regulations and ethical requirements of the medical profession. Differential privacy (DP) enables the provision of information-theoretic privacy guarantees to patients and can be implemented in the setting of deep neural network training through the differentially private stochastic gradient descent (DP-SGD) algorithm. We here present deepee, a free-and-open-source framework for differentially private deep learning for use with the PyTorch deep learning framework. Our framework is based on parallelised execution of neural network operations to obtain and modify the per-sample gradients. The process is efficiently abstracted via a data structure maintaining shared memory references to neural network weights to maintain memory efficiency. We furthermore offer specialised data loading procedures and privacy budget accounting based on the Gaussian Differential Privacy framework, as well as automated modification of the user-supplied neural network architectures to ensure DP-conformity of its layers. We benchmark our framework's computational performance against other open-source DP frameworks and evaluate its application on the paediatric pneumonia dataset, an image classification task and on the Medical Segmentation Decathlon Liver dataset in the task of medical image segmentation. We find that neural network training with rigorous privacy guarantees is possible while maintaining acceptable classification performance and excellent segmentation performance. Our framework compares favourably to related work with respect to memory consumption and computational performance. Our work presents an open-source software framework for differentially private deep learning, which we demonstrate in medical imaging analysis tasks. It serves to further the utilisation of privacy-enhancing techniques in medicine and beyond in order to assist researchers and practitioners in addressing the numerous outstanding challenges towards their widespread implementation

Contributions of thesis author: code development, experiment design and evaluation, paper writing.

scientific reports

Medical imaging deep learning with differential privacy

Alexander Ziller[1,2,4,5], Dmitrii Usynin[1,2,3,4,5], Rickmer Braren[1], Marcus Makowski[1], Daniel Rueckert[2,3] & Georgios Kaissis[1,2,3,4]✉

The successful training of deep learning models for diagnostic deployment in medical imaging applications requires large volumes of data. Such data cannot be procured without consideration for patient privacy, mandated both by legal regulations and ethical requirements of the medical profession. *Differential privacy* (DP) enables the provision of information-theoretic privacy guarantees to patients and can be implemented in the setting of deep neural network training through the *differentially private stochastic gradient descent* (DP-SGD) algorithm. We here present *deepee*, a free-and-open-source framework for differentially private deep learning for use with the *PyTorch* deep learning framework. Our framework is based on parallelised execution of neural network operations to obtain and modify the per-sample gradients. The process is efficiently abstracted via a data structure maintaining shared memory references to neural network weights to maintain memory efficiency. We furthermore offer specialised data loading procedures and privacy budget accounting based on the *Gaussian Differential Privacy* framework, as well as automated modification of the user-supplied neural network architectures to ensure DP-conformity of its layers. We benchmark our framework's computational performance against other open-source DP frameworks and evaluate its application on the *paediatric pneumonia dataset*, an image classification task and on the *Medical Segmentation Decathlon Liver* dataset in the task of medical image segmentation. We find that neural network training with rigorous privacy guarantees is possible while maintaining acceptable classification performance and excellent segmentation performance. Our framework compares favourably to related work with respect to memory consumption and computational performance. Our work presents an open-source software framework for differentially private deep learning, which we demonstrate in medical imaging analysis tasks. It serves to further the utilisation of privacy-enhancing techniques in medicine and beyond in order to assist researchers and practitioners in addressing the numerous outstanding challenges towards their widespread implementation.

Artificial Intelligence (AI) is a heavily data-centric domain: the success of machine learning (ML) models depends on the quality and quantity of data that is available during training. This is especially problematic in applications such as medical image analysis, in which high quality data is sparse and data utilisation is restricted. Medical data is highly sensitive, and regulatory, ethical and moral requirements restrict its sharing. These restrictions, although crucial, hinder the development of algorithms that generalise well and therefore prevent widespread deployment. Recent work[1] finds that even algorithms approved for diagnostic use are often trained on small (i.e. less than 1000 cases), single centre datasets. Considering that state-of-the-art generic computer vision models are customarily trained on datasets such as ImageNet[2] containing orders of magnitude more images, it becomes readily apparent that the access to more data will be strictly necessary for the development of the majority of deep learning applications in medical imaging to achieve the same success. Privacy-preserving machine learning is a nascent area of AI which proposes to bridge the gap between data utilisation and data protection through the application of privacy-enhancing techniques[3]. Among these, collaborative learning protocols such as federated learning have arguably witnessed the widest publicity[4]. They allow a confederation of clients to train ML models in a decentralised fashion and without sharing the raw data. However, a number of works suggest[5–7] that on its own, federated learning is an insufficient measure of privacy preservation. In the setting of medical imaging, this can result in catastrophic privacy loss for affected patients. Prior work demonstrates that federated learning without additional privacy-enhancing techniques can be reverse-engineered to reconstruct high-fidelity images which encode diagnostic information about patients, such as the absence of a breast indicative of a prior history

[1]Institute for Diagnostic and Interventional Radiology, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany. [2]Artificial Intelligence in Medicine and Healthcare, Technical University of Munich, Munich, Germany. [3]Department of Computing, Imperial College, London, UK. [4]OpenMined, Oxford, UK. [5]These authors contributed equally: Alexander Ziller and Dmitrii Usynin. ✉email: g.kaissis@tum.de

of breast cancer[8]. Moreover, three-dimensional medical imaging can be volumetrically rendered to reconstruct facial contours which enable patient re-identification[9]. Lastly, even when identifying attributes are not directly present in the image, the exploitation of side information by adversaries in the setting of linkage attacks, proven to represent a highly effective method for membership inference[10], is also applicable to medical imaging databases given that large-scale public datasets of medical images are being assembled and—increasingly—publicly released. Thus, solutions based on information-theoretic privacy measures are required to provide comprehensive and quantifiable guarantees to the involved parties. Differential privacy (DP)[11] has arisen as the gold standard in this regard. In brief, DP is the attribute of an algorithm to be approximately invariant to the inclusion or exclusion of individual patients, providing them with formal and quantifiable privacy guarantees. Although formally an information-theoretic privacy guarantee, in practice DP is typically achieved through *computationally secure* means, that is, an addition of carefully calibrated noise to the training process, making individual contributions indistinguishable from each other. In their seminal paper, Abadi et al.[12] demonstrated the successful application of DP in the training of deep neural networks, termed *differentially private stochastic gradient descent* (DP-SGD). However, the authors of this and subsequent works noted that the utilisation of DP-SGD unavoidably negatively affects the utility of the resulting models, a well-known effect termed the *privacy-utility trade-off*[13]. Addressing this trade-off[14] and ultimately enabling the widespread real-world utilisation of privacy-preserving ML in medical imaging and beyond requires the introduction of robust software tools, suitable for implementation within widely-used deep learning libraries and implementing current best practices.

We here present *deepee*, a software framework for differentially private deep learning based on the PyTorch[15] machine learning library. Our main contributions can be summarised as follows:

- We present a technical implementation of the DP-SGD algorithm based on parallelised execution, which makes our framework universally compatible with any neural network layer while enabling substantial performance improvements.
- We implement state-of-the-art tools for production-level DP-SGD application including cryptographically secure random noise generation, automatic architecture modifications and privacy budgeting based on the *Gaussian Differential Privacy* (GDP) framework which offers a tight analysis of privacy consumed.
- We benchmark our toolkit against comparable DP-SGD implementations and analyse the behaviour of DP-SGD in the setting of two medical imaging deep learning tasks: classification and semantic segmentation
- Our framework is aimed at facilitating the application of DP-SGD to arbitrary data by non-experts. For this purpose, it exposes standardised application programming interfaces, is highly compatible with the *PyTorch* deep learning framework and automatically enforces the relevant details to ensure the formal correctness of the DP-SGD algorithm application.
- The source code of our framework is documented in detail, fully tested and available publicly and freely under a permissive, open-source license to enable easy maintenance, rapid detection and correction of potential security vulnerabilities and to encourage open-source contributions.

Two notable works have presented DP frameworks for the *PyTorch* machine learning library based on different technical implementations. The *Opacus* framework[16] provides an implementation of the DP-SGD algorithm based on temporarily caching intermediate backpropagation results. This enables very high performance for specific deep neural network layer types. However, it does not ensure generic compatibility with any given neural network operation unless the procedure for obtaining said backpropagation results is explicitly defined on the user's side. At the time of writing, the framework's privacy analysis is still based on *Rényi DP* (RDP)[17], whose guarantees are not as tight as Gaussian DP (GDP). The *Pyvacy*[18] framework implements a generic version of DP-SGD based on serial execution. Despite its broad compatibility, this implementation is highly computationally inefficient, rendering it impractical for production-level use. The framework also lacks cryptographically secure random number generation and utility functions for automatic neural network architecture modification.

The *TensorFlow Privacy* framework[19] and previous work based on the *JAX* machine learning framework[20] share some characteristics of our library, such as utilisation of the GDP accounting technique or parallelisation, but they are based around different base libraries and thus are not directly comparable to our work.

Results

Technical overview.

We begin by providing a brief technical overview of our framework. Implementation details can be found in the "Methods" section. In brief, *deepee* implements the DP-SGD algorithm in a memory-efficient and parallelised manner by increasing the efficiency of the *per-sample-gradient* calculation step drastically compared to serial processing. This occurs by creating one zero-memory-cost reference to the network's weights for each sample in the minibatch, then performing a simultaneous (parallelised) forward and backward pass. This process introduces no additional assumptions about the network's architecture and thus allows the application of the DP-SGD algorithm to any neural network architecture. This represents an improvement compared to prior work, which requires substantial user effort to manually specify the *per-sample* gradient calculations for unsupported layer types (e.g. *pixel shuffle or transposed convolutions, transformers, etc.*) or relies on performing forward and backward passes serially, thus magnifying time complexity. The framework furthermore is designed to guarantee the formal correctness of the DP-SGD procedure by e.g. removing Batch Normalisation layers from the architecture, employing cryptographically secure random noise and automatic privacy budgeting.

In the following, we demonstrate the utilisation of our framework in the settings of medical image classification and semantic segmentation. We present model performance in private and non-private settings to evaluate

Model	ROC-AUC	GDP ε	RDP ε
Non-private	0.960 [0.946 to 0.971]	∞	∞
Private	0.848 [0.814 to 0.881]	0.52	0.64
Private (relaxed)	0.882 [0.868 to 0.899]	2.69	2.81

Table 1. Classification performance (measured as mean receiver-operator characteristic area-under-the-curve (ROC-AUC)) on the paediatric chest radiography binary classification dataset. Ranges in angled brackets. The non-private model significantly outperformed the private model in both the high-privacy setting and the relaxed privacy setting, while the private model trained with relaxed privacy guarantees significantly outperformed the private model with strict guarantees.

Model	Dice coefficient	GDP ε	RDP ε
Non-private	0.950 [0.948 to 0.951]	∞	∞
Private	0.943 [0.941 to 0.945]	0.12	0.35

Table 2. Segmentation performance (measured by the mean Dice coefficient) on the liver semantic segmentation dataset. Ranges in angled brackets. The privately trained and the non-privately trained models performed on par despite the provision of stringent privacy guarantees in the privately trained setting.

Task	*deepee* (ours)	*Opacus*	*Pyvacy*
Classification	38.82 s [38.67 to 39.08]	16.39 s [16.29 to 16.69]	73.11 s [72.41 to 75.40]
	6366 MiB [6201 to 6448]	7014 MiB [6816 to 7213]	2044 MiB [1992 to 2102]
Segmentation	70.89 s [70.41 to 71.01]	78.47 s [78.08 to 79.86]	97.89 s [97.26 to 99.16]
	9770 MiB [9508 to 9829]	9909 MiB [9812 to 10112]	2085 MiB [1890 to 2205]
Segmentation (Transposed Conv.)	47.27 s [45.12 to 51.15]	–	64.68 s [62.76 to 66.32]
	12014 MiB [11598 to 12249]	–	1537 MiB [1399 to 1620]

Table 3. Computational performance (median time for N = 25 batches of 32 examples in seconds over N = 5 repetitions) and mean peak memory consumption (one batch of 32 examples in MiB, N = 6 repetitions) of the compared frameworks for the classification and segmentation benchmarks. Ranges in angled brackets. The Segmentation (Transposed Conv.) row showcases framework performance in a U-Net architecture using transposed convolutions. *Opacus* is incompatible with this layer type.

the expected privacy-utility trade-offs. Moreover, we compare our library's computational performance with alternative implementations of the algorithm offered by the *Opacus* and *Pyvacy* frameworks.

Chest radiography classification.

The classification model achieved a mean receiver-operator characteristic area-under-the-curve (ROC-AUC) of 0.848 (range 0.814 to 0.881) in the private setting and of 0.960 (range 0.946 to 0.971) in the non-private setting (*DeLong*-test $p < 0.001$, $N = 10$). GDP accounting yielded a privacy budget (ε) of 0.52 at a noise multiplier of 3.0 and an L_2 clipping norm of 1.0, a tighter result than 0.62, which would have resulted from the utilisation of RDP analysis ($\delta = 10^{-5}$). We observed that relaxing the privacy parameters (noise multiplier and clipping norm) resulted in a significant increase in classification performance of the private model (ROC-AUC in the relaxed privacy setting 0.882, range 0.868 to 0.899, *DeLong*-test vs. the strict privacy setting $p < 0.001$, $N = 10$) for an ε of 2.69 (GDP accounting) or 2.81 (RDP accounting). Even in the relaxed setting however, the model still significantly underperformed compared to non-private training (*DeLong*-test vs. non-private training $p < 0.001$, $N = 10$). These results are summarised in Table 1.

Semantic segmentation of computed tomography images.

In the semantic liver tissue segmentation task, the non-privately and privately trained models produced nearly identical results: The mean Dice coefficient achieved by the privately and the non-privately trained models was 0.943 (range 0.941 to 0.945), and 0.950 (range 0.948 to 0.951, N = 5), respectively. This segmentation performance of the privately trained model was attained at an ε of 0.12 (GDP) or 0.35 (RDP) and a δ-value of 10^{-5}, resulting from a noise multiplier of 5.0 and an L_2 clipping norm of 0.5, indicating that the provision of strict privacy guarantees was possible in this setting without a notable trade-off in model performance. Results are summarised in Table 2.

Computational performance comparison.

Table 3 presents a comparison of the computational performance and memory consumption of our framework versus the *Opacus* and *Pyvacy* libraries in the classification and segmentation settings. We found our framework to offer significantly faster computational performance in the segmentation setting compared to *Opacus* (*Student's* t-test $p < 0.001$) and *Pyvacy* ($p < 0.001$). *Opacus* significantly outperformed our framework ($p < 0.001$) and *Pyvacy* ($p < 0.001$) in the classification task. (All 25 batches of 32 examples over N = 5 repetitions).

Our framework required significantly less memory than *Opacus* in both the classification and segmentation setting (*Student's* t-test $p < 0.001$). *Pyvacy*, due to serial processing of the individual samples in each minibatch suffers from a drastically diminished computational performance, however requires significantly less memory than both other frameworks as a result of only needing to cache a single sample's gradients at a time (*Student's* t-test $p < 0.001$, all N = 6 repetitions).

Moreover, to exemplify our framework's compatibility, we benchmarked an additional U-Net architecture utilising transposed convolutions as described in the original work[21]. The *Opacus* framework is incompatible with transposed convolutions and could thus not be assessed. *Pyvacy*, while requiring less memory ($p < 0.001$), again was significantly slower per batch compared to *deepee* ($p < 0.001$).

Discussion

Here we present a novel technical implementation of the DP-SGD algorithm which we demonstrate and benchmark in the setting of medical image analysis. We found our technique's computational performance and memory consumption to be comparable to state-of-the-art frameworks without a requirement for user-side modifications. Our framework thus provides formal privacy guarantees regardless of the dataset, learning task and of model selection. Moreover, by leveraging the current state-of-the-art in DP analysis, we demonstrate tighter privacy bounds compared to previous DP accounting techniques. The two applications presented provide evidence for the usefulness of our DP-SGD algorithm in real-world medical image processing.

Medical imaging represents a domain in which privacy-utility trade-offs are especially problematic, as models that generalise well require large and diverse multi-centre datasets during training and must not divulge personal test data once deployed. Such demands are—for example—placed on ML models utilised for remote diagnosis-as-a-service[22], where expert-level algorithm performance is expected, while the model may be exposed to probing by malicious third parties. Formal security and secrecy mechanisms such as model encryption can only partially address this requirement, as even encrypted models have been found to leak sensitive information in previous work[23,24]. Similarly, distributed learning techniques such as federated learning, often touted as being "privacy-preserving" because the data does not leave its owner, have been proven ineffective against attackers who participate in the training protocol and are able to capture updates submitted by other participants[5,6]. Differentially private model training therefore stands as the only formal mechanism for privacy protection, able to shield models from feature reconstruction, model inversion and membership inference attacks[6,25]. Moreover, recent work demonstrates that DP can reduce the susceptibility of models to other adversarial interference such as *back-door attacks*[26], which can be attributed to the increased robustness of DP models imparted through the regularising properties of noise addition[27].

Inherent to these beneficial properties of DP model training is—however—also an unavoidable net reduction in model utility. We identify three key components of this utility penalty: (1) Diminished task-specific performance, e.g. in classification or segmentation tasks; (2) computational performance penalties through an increase in training time and memory consumption and (3) incompatibilities of the DP-SGD algorithm with the neural network architecture. Our work attempts to address all three of these points.

The use-cases chosen in our study, image classification and segmentation, represent two typical workflows in medical imaging analysis. Interestingly, we observed a marked performance decrease in the private classification task compared to non-private model training even under relaxed privacy guarantees. Semantic segmentation was possible under very strong privacy notions with unexpectedly strong performance. The only other work to report an ε-value in a medical image segmentation task[28] utilises a different DP technique, whose utilisation results in a high privacy expenditure of over 120 under the study's assumptions, compared to 0.12 in our work. No previous work—to our knowledge—reports ε-values for medical image classification. At present, it is not yet conclusively investigated to which extent the difficulty of the task, the choice of model and the specific training technique influence the privacy-utility trade-off. Future work will thus have to elucidate these relationships and expand on recent studies in this direction[13,14,29].

Besides these factors, more refined techniques for privacy accounting are able to offer an improved analysis of the DP mechanism and thus allow higher utility. In the medical imaging domain, the combination of high utility and low privacy budget is particularly important. As datasets are complex, highly sensitive and typically small, each individual in the dataset experiences a relatively higher privacy loss. A tight privacy analysis allows training the models for a longer time before the privacy budget is exhausted, enabling higher task-specific performance and therefore, a better diagnostic prediction. Our work utilises Gaussian Differential Privacy, a recently introduced DP formulation which—through a tight characterisation of the sub-sampled Gaussian noise mechanism utilised in DP-SGD—improves the outlook on the spent privacy budget compared to previous frameworks. It is expected that further advances, such as individual privacy accounting[30,31] will increase the granularity of privacy tracking further, allowing for the preservation of even higher utility during algorithm training.

Our main technical contribution is the introduction of a parallelised execution model for the DP-SGD algorithm within the *PyTorch* framework, which enables both fast performance and efficient memory utilisation. In addition, our technique-contrary to frameworks relying on the *a priori* specification of *per-sample* gradient calculations such as *Opacus*- is compatible by default with *any* neural network operation including (but not limited to) transformer architectures or transposed convolutions, as seen above. This disparity is discussed in[20], a line of work complementary to ours, whose authors utilise *just-in-time* compilation and vectorised execution to increase DP-SGD performance, albeit within a different machine learning framework. We moreover see a target for future work focused around automatic differentiation with inbuilt support for obtaining and manipulating *per-sample* gradients. After all, the requirement to calculate *per-sample* gradients in current DP-SGD frameworks stems from the inherent design philosophy of reverse-mode automatic differentiation systems, which are focused on efficiently obtaining gradients for minibatches but not for individual samples. We moreover note that techniques concerned with approximate gradient calculations[32] have some overlap with the objectives of DP-SGD, which inherently performs an "imprecise" gradient update step through noise addition, and could thus be utilised for increased performance, after considering their effect on privacy guarantees.

Similar to previous work[16], our work offers the capability to automatically modify the neural network architecture in case layers incompatible with DP-SGD are included. An example of this phenomenon in the current work is the deactivation of running statistics collection for Batch Normalisation layers. Moreover, our framework includes support for cryptographically secure random noise generation which is crucial to avoid vulnerabilities associated with default pseudo-random number generators[33].

We consider some limitations of our work: Our framework's focus is to provide a generic framework for DP-SGD and the examples presented represent a simplification of real-life use-cases intended to illustrate its utilisation in medical imaging. In the segmentation case-study in particular, we provide image-level privacy guarantees, whereas a real-life deployment would be adjusted to offer patient-level guarantees (that is, a "summary" of privacy guarantees derived from the utilisation of all images of a single patient). Moreover, DP techniques purpose-designed for high performance in classification, such as PATE[34] could yield improved privacy-utility trade-offs in the classification use-case compared to DP-SGD, however at the cost of not generalising well to other tasks such as segmentation[28] and an additional assumption of a publicly available dataset that cannot be reliably expected in a sensitive setting, such as medical imaging.

In conclusion, our work aims to facilitate the utilisation of differentially private deep learning in everyday practice. It is well-suited to privacy-sensitive tasks such as medical imaging analysis. We publicly release our framework and experiments in the hope that it will stimulate future research and lead to the design of improved algorithms and training techniques to enable privacy-preserving machine learning with improved algorithm utility in medical imaging and beyond.

Methods

Framework implementation details.

User-facing components. Our framework provides the following high-level user-facing components: (1) A collection of procedures to automatically modify the neural network architecture in case it contains layers which are incompatible for utilisation with DP-SGD. One example is the Batch Normalisation layer which maintains a (non-private) running average of statistics over more than one training example and is thus not compatible with the notion of *per-sample* gradient calculations, which are required in DP-SGD. (2) A data structure encapsulating the user-supplied model architecture, responsible for the main model training and evaluation loop. This *wrapper* internally maintains one copy of the user-supplied model per sample in the minibatch, performs a parallelised forward and backward pass over the minibatch and abstracts the gradient clipping and noise application of the DP-SGD procedure. (3) A *privacy accounting* mechanism for keeping track of the privacy spent at each training step and including a procedure to automatically interrupt the training if the privacy budget is exhausted. The system is supplemented by a cryptographically secure random number generator[35] suitable for use on the graphics processing unit and capable of parallelising the random noise generation step of the DP-SGD algorithm.

DP-SGD algorithm implementation. We implement the DP-SGD algorithm as described in[12]. In brief, the algorithm consists of the following steps:

1. Performing a forward pass on a minibatch of samples
2. Calculating the gradient of the loss with respect to each sample individually (*per-sample gradients*)
3. Normalising (*clipping*) the per-sample gradients to a predefined L_2-norm
4. Aggregating the per-sample gradients by averaging or summing over the minibatch axis
5. Adding calibrated Gaussian noise to the resulting gradient vector

In practice, step (2) of the above-mentioned procedure is the most time-consuming subroutine of the algorithm, as automatic differentiation systems are not designed with per-sample gradient computation in mind. To tackle this problem, our framework first creates a copy of the neural network for each sample in the minibatch and then performs step (1) of the algorithm above in parallel by dispatching one execution thread per minibatch sample. Thus, the backpropagation procedure yields per-sample gradients per definition (step (2) above). This approach has several benefits: It is computationally efficient as it is performed in parallel over the minibatch leveraging multi-threaded execution on e.g. the graphics processing unit (GPU). Moreover, memory only needs be allocated once for the neural network weights (as all copies share the same weights). Lastly, the process is entirely generic and can be used for any arbitrary neural network architecture without the requirement for user interaction. A similar technique to ours, albeit based on serial execution instead of a parallelised forward pass and only demonstrated for convolutional neural networks, is presented in[36], reportedly going back to (unpublished) work by Goodfellow et al.

Datasets.

Classification task. We evaluated our framework on a classification task on chest radiographs from the Paediatric Pneumonia dataset originally described in[37]. Originally, the task was formulated as three-class classification, however we merged the *viral* and *bacterial* pneumonia labels to obtain a binary classification task, in which the algorithm attempts to predict whether the radiograph shows signs of pneumonia or not. The dataset contains 1339 training images of healthy patients and 3824 images of patients that present evidence of pneumonia. The dataset is pre-split into a training (n = 5163) and a test set (n = 624). We further split the training set into 85% training data (n = 4389) and 15% validation data (n = 774). To account for class imbalance, we weighted the resulting loss by one minus the proportion of the dataset of the class. Data augmentation was performed using affine transformations (rotation, scaling, translation, shearing). Every occurence of an image from the same patient, regardless whether it was augmented or not, was counted against the total privacy expenditure. We trained the models for 20 epochs using the Adam optimiser in the non-private setting and the Stochastic Gradient Descent (SGD) optimiser in the private setting. Learning rates were determined using a learning rate finding algorithm[38] and set to 0.005 in both settings. Learning rate scheduling with halving of the learning rate on stagnation of the validation loss for two consecutive epochs was employed.

Semantic segmentation task. For the semantic segmentation task, we used the Medical Segmentation Decathlon (MSD) Liver segmentation dataset[39]. We split the available data into a training set (n = 5184), a validation set (n = 640) and a held-out test set (n = 2560), mindful to enforce strict patient independence between the training/validation sets and the test set. The task was re-formulated as a binary segmentation task, in which the liver tissue pixels (including tumours) are labelled as 1 and the background as 0. For augmentation purposes, affine transformations (rotation, translation, scaling, flipping) alongside random Gaussian noise were applied to the input images. Every occurence of an image from the same patient, regardless whether it was augmented or not, was counted against the total privacy expenditure. The model was trained for 20 epochs in the non-private setting. In the private setting, we limited the number of epochs to 5 in order to maintain a low privacy budget. Learning rates were determined using the same learning rate finding algorithm and set to 0.01, while utilising the Adam optimiser in both cases. Learning rate scheduling was performed in the same manner as for the classification task.

Model training.

For the classification task, we utilised the same model architecture in the private and non-private setting, namely a VGG-11[40] architecture with Batch Normalisation. However, in order to satisfy the assumptions essential for DP training, the collection of running statistics of Batch Normalisation layers was disabled for both non-private and DP training. For the segmentation task, we use a modified U-Net architecture[21] utilising VGG-11 with Batch Normalisation as a backbone[41]. Similarly to the classification task, the running statistics collection was disabled. The δ-parameter was set to 10^{-5} in all cases.

Computational performance and memory benchmarks.

For the purposes of computational performance benchmarking we measured the time to train for 25 steps with a minibatch size of 32 on the tasks we presented above, i.e., binary classification on 224x224 sized images and the segmentation of 256x256 images. Each measurement was repeated five times.

For memory utilisation benchmarking, a minibatch size of 32 images at a resolution of 256×256 was used, with a single channel for the classification benchmark and three channels for the segmentation benchmark. All benchmarks were conducted in triplicate to ensure stability between runs and repeated on two operating systems, *macOS 11.2.3* and *GNU Linux* on the 5.4.0-72 kernel (total N = 6 runs). Peak memory consumption was measured using the *Python* programming language (*CPython* v. 3.8.8) standard library module *resource*.

Statistical methods.

Areas under the ROC-curve were compared using the *DeLong*-test as described in[42]. Continuous variables were compared using the *Student's* t-test. *Bonferroni's* correction was used for three-way comparisons with the adjusted statistical significance threshold set to $p = 0.016$.

Accession codes

The *deepee* framework and code to reproduce the experiments is available at https://github.com/gkaissis/deepee. The paediatric pneumonia dataset is available from https://data.mendeley.com/datasets/rscbjbr9sj/3. The liver segmentation dataset is available from http://medicaldecathlon.com.

Received: 29 April 2021; Accepted: 18 June 2021
Published online: 29 June 2021

References

1. Wu, E. *et al.* How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals. *Nat. Med.*https://doi.org/10.1038/s41591-021-01312-x *(2021).*
2. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (IEEE, 2009).
3. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 1–7 (2020).
4. Sheller, M. J. *et al.* Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 1–12. https://doi.org/10.1038/s41598-020-69250-1 (2020).
5. Zhu, L., Liu, Z. & Han, S. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, 14747–14756 (2019).
6. Geiping, J., Bauermeister, H., Dröge, H. & Moeller, M. Inverting gradients—How easy is it to break privacy in federated learning? arXiv preprint arXiv:2003.14053 (2020).
7. He, Z., Zhang, T. & Lee, R. B. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, 148–162 (2019).
8. Kaissis, G. *et al.* End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.*https://doi.org/10.1038/s42256-021-00337-8 *(2021).*

9. Schwarz, C. G. *et al.* Identification of anonymous MRI research participants with face-recognition software. *N. Engl. J. Med.* **381**, 1684–1686. https://doi.org/10.1056/nejmc1908881 (2019).
10. Orekondy, T., Oh, S. J., Zhang, Y., Schiele, B. & Fritz, M. Gradient-leaks: Understanding and controlling deanonymization in federated learning. arXiv preprint arXiv:1805.05838 (2018).
11. Dwork, C. & Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407. https://doi.org/10.1561/0400000042 (2013).
12. Abadi, M. *et al.* Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318 (2016).
13. Avent, B., Gonzalez, J., Diethe, T., Paleyes, A. & Balle, B. Automatic discovery of privacy-utility pareto fronts. arXiv preprint arXiv:1905.10862 (2019).
14. Papernot, N., Chien, S., Song, S. & Thakurta, A. & Erlingsson, U. Architectures, initializations, and tuning for learning with privacy, making the shoe fit (2020).
15. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32* (eds Wallach, H. *et al.*) 8024–8035 (Curran Associates, Inc., 2019).
16. Opacus PyTorch library. Available from https://opacus.ai
17. Mironov, I. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 263–275 (IEEE, 2017).
18. Waites, C. *Pyvacy: Towards Practical Differential Privacy for Deep Learning* (Georgia Tech Library, 2019). https://github.com/ChrisWaites/pyvacy.
19. TensorFlowPrivacy. Available from https://github.com/tensorflow/privacy
20. Subramani, P., Vadivelu, N. & Kamath, G. Enabling fast differentially private SGD via just-in-time compilation and vectorization. arXiv preprint arXiv:2010.09063 (2020).
21. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, 234–241 (Springer, 2015).
22. La, H. J., Kim, M. K. & Kim, S. D. A personal healthcare system with inference-as-a-service. In *2015 IEEE International Conference on Services Computing*, 249–255 (IEEE, 2015).
23. Ziller, A. *et al.* Privacy-preserving medical image analysis. arXiv preprint arXiv:2012.06354 (2020).
24. Hayes, J., Melis, L., Danezis, G. & De Cristofaro, E. Logan: Membership inference attacks against generative models. arXiv preprint arXiv:1705.07663 (2017).
25. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18 (IEEE, 2017).
26. Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D. & Jana, S. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, 656–672 (IEEE, 2019).
27. Dwork, C. *et al.* Generalization in adaptive data analysis and holdout reuse. arXiv preprint arXiv:1506.02629 (2015).
28. Fay, D., Sjölund, J. & Oechtering, T. J. Decentralized differentially private segmentation with PATE. arXiv:2004.06567 (2020).
29. van der Veen, K. L., Seggers, R., Bloem, P. & Patrini, G. Three tools for practical differential privacy. arXiv:1812.02890 (2018).
30. Feldman, V. & Zrnic, T. Individual privacy accounting via a Renyi filter. arXiv preprint arXiv:2008.11193 (2020).
31. Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D. & Megías, D. Individual differential privacy: A utility-preserving formulation of differential privacy guarantees. *IEEE Trans. Inf. Forensics Secur.* **12**, 1418–1429 (2017).
32. Oktay, D., McGreivy, N., Aduol, J., Beatson, A. & Adams, R. P. Randomized automatic differentiation. arXiv preprint arXiv:2007.10412 (2020).
33. Garfinkel, S. L. & Leclerc, P. Randomness concerns when deploying differential privacy. In *Proceedings of the 19th Workshop on Privacy in the Electronic Society* (ACM, 2020). https://doi.org/10.1145/3411497.3420211.
34. Papernot, N. *et al.* Scalable private learning with pate. arXiv preprint arXiv:1802.08908 (2018).
35. Salmon, J. K., Moraes, M. A., Dror, R. O. & Shaw, D. E. Parallel random numbers: as easy as 1, 2, 3. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–12 (2011).
36. Rochette, G., Manoel, A. & Tramel, E. W. Efficient per-example gradient computations in convolutional neural networks. arXiv preprint arXiv:1912.06015 (2019).
37. Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).
38. Smith, L. N. *Cyclical learning rates for training neural networks*. arXiv:1506.01186 (2017).
39. Simpson, A. L. *et al.* A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019).
40. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
41. Yakubovskiy, P. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch (2020).
42. Molodianovitch, K., Faraggi, D. & Reiser, B. Comparing the areas under two correlated ROC curves: Parametric and non-parametric approaches. *Biometr. J.* **48**, 745–757. https://doi.org/10.1002/bimj.200610223 (2006).

Acknowledgements

This work was supported by the UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare (G.K. and D.R.) and the Technical University Munich/Imperial College London Joint Academy of Doctoral Studies (D.U.).

Author contributions

G.K. conceived and developed deepee and helped with experimental evaluation. A.Z. helped with deepee development, conceived and performed the experimental evaluation. A.Z., G.K. and D.U. wrote the initial manuscript. G.K. and D.U. revised the manuscript. R.B. and M.M. provided oversight for the medical imaging use-case. D.R. provided oversight for the technical implementation. R.B., M.M. and D.R. provided input on the revised manuscript. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to G.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.3 Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging

Synopsis:
Artificial intelligence (AI) models are increasingly used in the medical domain. However, as medical data is highly sensitive, special precautions to ensure its protection are required. The gold standard for privacy preservation is the introduction of differential privacy (DP) to model training. Prior work indicates that DP has negative implications on model accuracy and fairness, which are unacceptable in medicine and represent a main barrier to the widespread use of privacy-preserving techniques. In this work, we evaluated the effect of privacy-preserving training of AI models regarding accuracy and fairness compared to non-private training.We used two datasets: (1) A large dataset ($N = 193\,311$) of high quality clinical chest radiographs, and (2) a dataset ($N = 1\,625$) of 3D abdominal computed tomography (CT) images, with the task of classifying the presence of pancreatic ductal adenocarcinoma (PDAC). Both were retrospectively collected and manually labeled by experienced radiologists. We then compared non-private deep convolutional neural networks (CNNs) and privacy-preserving (DP) models with respect to privacy-utility trade-offs measured as area under the receiver operating characteristic curve (AUROC), and privacy-fairness trade-offs, measured as Pearson's r or Statistical Parity Difference.We find that, while the privacy-preserving training yields lower accuracy, it largely does not amplify discrimination against age, sex or co-morbidity. However, we find an indication that difficult diagnoses and subgroups suffer stronger performance hits in private training.Our study shows that – under the challenging realistic circumstances of a real-life clinical dataset – the privacy-preserving training of diagnostic deep learning models is possible with excellent diagnostic accuracy and fairness.

Contributions of thesis author: code development, experiment design and evaluation, paper writing.

communications medicine

Article

https://doi.org/10.1038/s43856-024-00462-6

Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging

Soroosh Tayebi Arasteh[1,6] ✉, Alexander Ziller[2,3,6] ✉, Christiane Kuhl[1], Marcus Makowski[2], Sven Nebelung[1], Rickmer Braren[2], Daniel Rueckert[3], Daniel Truhn[1,7] ✉ & Georgios Kaissis[2,3,4,5,7] ✉

Abstract

Background Artificial intelligence (AI) models are increasingly used in the medical domain. However, as medical data is highly sensitive, special precautions to ensure its protection are required. The gold standard for privacy preservation is the introduction of differential privacy (DP) to model training. Prior work indicates that DP has negative implications on model accuracy and fairness, which are unacceptable in medicine and represent a main barrier to the widespread use of privacy-preserving techniques. In this work, we evaluated the effect of privacy-preserving training of AI models regarding accuracy and fairness compared to non-private training.

Methods We used two datasets: (1) A large dataset ($N = 193{,}311$) of high quality clinical chest radiographs, and (2) a dataset ($N = 1625$) of 3D abdominal computed tomography (CT) images, with the task of classifying the presence of pancreatic ductal adenocarcinoma (PDAC). Both were retrospectively collected and manually labeled by experienced radiologists. We then compared non-private deep convolutional neural networks (CNNs) and privacy-preserving (DP) models with respect to privacy-utility trade-offs measured as area under the receiver operating characteristic curve (AUROC), and privacy-fairness trade-offs, measured as Pearson's r or Statistical Parity Difference.

Results We find that, while the privacy-preserving training yields lower accuracy, it largely does not amplify discrimination against age, sex or co-morbidity. However, we find an indication that difficult diagnoses and subgroups suffer stronger performance hits in private training.

Conclusions Our study shows that – under the challenging realistic circumstances of a real-life clinical dataset – the privacy-preserving training of diagnostic deep learning models is possible with excellent diagnostic accuracy and fairness.

Plain Language Summary

Artificial intelligence (AI), in which computers can learn to do tasks that normally require human intelligence, is particularly useful in medical imaging. However, AI should be used in a way that preserves patient privacy. We explored the balance between maintaining patient data privacy and AI performance in medical imaging. We use an approach called differential privacy to protect the privacy of patients' images. We show that, although training AI with differential privacy leads to a slight decrease in accuracy, it does not substantially increase bias against different age groups, genders, or patients with multiple health conditions. However, we notice that AI faces more challenges in accurately diagnosing complex cases and specific subgroups when trained under these privacy constraints. These findings highlight the importance of designing AI systems that are both privacy-conscious and capable of reliable diagnoses across patient groups.

The development of artificial intelligence (AI) systems for medical applications represents a delicate trade-off: On the one hand, diagnostic models must offer high accuracy and certainty, as well as treat different patient groups equitably and fairly. On the other hand, clinicians and researchers are subject to ethical and legal responsibilities towards the patients whose data is used for model training. In particular, when diagnostic models are published to third parties whose intentions are impossible to verify, care must be undertaken to ascertain that patient privacy is not compromised.

[1]Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany. [2]Institute of Diagnostic and Interventional Radiology, Technical University of Munich, Munich, Germany. [3]Artificial Intelligence in Healthcare and Medicine, Technical University of Munich, Munich, Germany. [4]Department of Computing, Imperial College London, London, United Kingdom. [5]Institute for Machine Learning in Biomedical Imaging, Helmholtz Munich, Neuherberg, Germany. [6]These authors contributed equally: Soroosh Tayebi Arasteh, Alexander Ziller.[7]These authors jointly supervised this work: Daniel Truhn, Georgios Kaissis. ✉e-mail: soroosh.arasteh@rwth-aachen.de; alex.ziller@tum.de; dtruhn@ukaachen.de; g.kaissis@tum.de

Privacy breaches can occur, e.g., through data reconstruction, attribute inference or membership inference attacks against the shared model[1]. Federated learning[2–4] has been proposed as a tool to address some of these problems. However, it has become evident that training data can be reverse-engineered from federated systems, rendering them just as vulnerable to the aforementioned attacks as centralized learning[5]. Thus, it is apparent that formal privacy preservation methods are required to protect the patients whose data is used to train diagnostic AI models. The gold standard in this regard is differential privacy (DP)[6].

Most, if not all, currently deployed machine learning models are trained without any formal privacy-preservation technique. It is especially crucial to employ such techniques in federated scenarios, where much more granular information about the training process can be extracted, or even the training process itself can be manipulated by a malicious participant[7,8]. Moreover, trained models can be attacked to extract training data through so-called model inversion attacks[9–11]. We also note that such attacks work better if the models have been trained on less data, which is especially concerning since even most FDA-approved AI algorithms have been trained on fewer than 1000 cases[12]. Creating a one-to-one correspondence between a successful attack and the resulting “privacy risk” requires a case-by-case consideration. The legal opinion (e.g., the GDPR) seems to have converged on the notion of singling out/ re-identification. Even from the aspect of newer legal frameworks, such as the EU AI act, which demand “risk moderation” rather than directly specifying “privacy requirements,”, DP can be seen as the optimal tool as it can quantitatively bound both the risk of membership inference (MI)[13,14] and data reconstruction[15]. Moreover, this was also shown empirically for both aforementioned attack classes[16–19]. It is also known that DP, contrary to de-identification procedures such as k-anonymity, provably protects against the notion of singling out[20,21].

DP is a formal framework encompassing a collection of techniques to allow analysts to obtain insights from sensitive datasets while guaranteeing the protection of individual data points within them. DP thus is a property of a data processing system which states that the results of a computation over a sensitive dataset must be approximately identical whether or not any single individual was included or excluded from the dataset. Formally, a randomized algorithm (mechanism) $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is said to satisfy (ε, δ)-DP if, for all pairs of databases $D, D' \in \mathcal{X}$ which differ in one row and all $S \subseteq \mathcal{Y}$, the following holds:

$$\Pr(\mathcal{M}(D) \in S) \leq e^{\varepsilon} \Pr(\mathcal{M}(D') \in S) + \delta, \tag{1}$$

where the guarantee is given over the randomness of $\mathcal{M}$ and holds equally when D and D' are swapped. In more intuitive terms, DP is a guarantee given from a data processor to a data owner that the risks of adverse events which can occur due to the inclusion of their data in a database are bounded compared to the risks of such events when their data is not included. The parameters ε and δ together form what is typically called a privacy budget. Higher values of ε and δ correspond to a looser privacy guarantee and vice versa. With some terminological laxity, ε can be considered a measure of the privacy loss incurred, whereas δ represents a (small) probability that this privacy loss is exceeded. For deep learning workflows, δ is set to around the inverse of the database size. We note that, although mechanisms exist where δ denotes a catastrophic privacy degradation probability, the sampled Gaussian mechanism used to train neural networks does not exhibit this behavior. The fact that quantitative privacy guarantees can be computed over many iterations (compositions) of complex algorithms like the ones used to train neural networks is unique to DP. This process is typically referred to as privacy accounting. Applied to neural network training, the randomization required by DP is ensured through the addition of calibrated Gaussian noise to the gradients of the loss function computed for each individual data point after they have been clipped in ℓ_2-norm to ensure that their magnitude is bounded[22], where the clipping threshold is an additional hyperparameter in the training process.

DP does not only offer formal protection, but several works have also empirically shown the connection between the privacy budget and the success of membership inference[16] and data reconstruction attacks[17,19,23]. We note that absolute privacy (i.e., zero risk) is only possible if no information is present[24]. This is, for example, the case in encryption methods, which are perfectly private as long as data is not decrypted. Note that training models e.g., via homomorphic encryption does, however, not offer such perfect privacy guarantees, as the information learned by the model is actually revealed at inference time through the model’s predictions. Thus, without the protection of differential privacy, no formal barrier stands between the sensitive data and an attacker (beyond potential imperfections of the attack algorithm, which are usually not controllable a priori). DP offers the ability to upper-bound the risk of successful privacy attacks while still being able to draw conclusions from the data. Determining the exact privacy budget is challenging, as it is a matter of policy. The technical perspective can provide insight into the appropriate budget level, as it is possible to quantify the risk of a successful attack at a given privacy budget compared to the model utility that can be achieved. The trade-offs between model utility and privacy preservation are also a matter of ethical, societal and political debate. The utilization of DP also creates two fundamental trade-offs: The first is a “privacy-utility trade-off,” i.e., a reduction in diagnostic accuracy when stronger privacy guarantees are required[25,26]. The other trade-off is between privacy and fairness. Intuitively, the fact that AI models learn proportionally less about under-represented patient groups[27] in the training data is amplified by DP, leading to demographic disparity in the model’s predictions or diagnoses[28]. Both of these trade-offs are delicate in sensitive applications, such as medical ones, as it is not acceptable to have wrong diagnoses or to discriminate against a certain patient group.

The need for the use of differential privacy (DP) has been illustrated by Packhäuser et al.[29], who showed that it is trivial to match chest x-rays of the same patient, which directly enables re-identification attacks; this was similarly shown in tabular databases by Narayanan et al.[30]. The training of deep neural networks on medical data with DP has so far not been widely investigated. Li et al.[31] investigated privacy-utility trade-offs in the combination of advanced federated learning schemes and DP methods on a brain tumor segmentation dataset. They find that DP introduces a considerable reduction in model accuracy in the given setting. Hatamizadeh et al.[23] illustrated that the use of federated learning alone can be unsafe in certain settings. Ziegler et al.[32] reported similar findings when evaluating privacy-utility trade-offs for a chest x-ray classification on a public dataset. These results also align with our previous work[17], where we demonstrated the utilization of a suite of privacy-preserving techniques for pneumonia classification in pediatric chest X-rays. However, the focus of this study was not to elucidate privacy-utility or privacy-fairness trade-offs, but to showcase that federated learning workflows can be used to train diagnostic AI models with good accuracy on decentralized data while minimizing data privacy and governance concerns. Moreover, we demonstrated that empirical data reconstruction attacks are thwarted by the utilization of differential privacy. In addition, the work did not consider differential diagnosis but only coarse-label classification into normal vs. bacterial or viral pneumonia.

In this work, we aim to elucidate the connection between using formal privacy techniques and the fairness towards underrepresented groups in the sensitive setting of medical use-cases. This is an important prerequisite for the deployment of ethical AI algorithms in such sensitive areas. However, so far, prior work is limited to benchmark computer vision datasets[33,34]. We thus contend that the widespread use of privacy-preserving machine learning requires testing under real-life circumstances. In the current study, we perform the first in-depth investigation into this topic. Concretely, we utilize a large clinical database of radiologist-labeled radiographic images, which has previously been used to train an expert-level diagnostic AI model, but otherwise not been curated or pre-processed for private training in any way. Furthermore, we analyze a dataset of abdominal 3D computed tomography (CT) images, where we classify the presence of a pancreatic ductal adenocarcinoma (PDAC). This mirrors the type of datasets available at clinical institutions. In this setting, we then study the extent of privacy-utility and privacy-fairness trade-offs in training advanced computer vision architectures.

To the best of our knowledge, our study is the first work to investigate the use of differential privacy in the training of complex diagnostic AI models on a real-world dataset of this magnitude (nearly 200,000 samples) and a 3D classification task, and to include an extensive evaluation of privacy-utility and privacy-fairness trade-offs.

Our results are of interest to medical practitioners, deep learning experts in the medical field and regulatory bodies such as legislative institutions, institutional review boards and data protection officers and we undertook specific care to formulate our main lines of investigation across the important axes delineated above, namely the provision of objective metrics of diagnostic accuracy, privacy protection and demographic fairness towards diverse patient subgroups.

Our main contributions can be summarized as follows: (1) We study the diagnostic accuracy ramifications of differentially private deep learning on two curated databases of medically relevant use-cases. We reach 97% of the non-private AUROC on the UKA-CXR dataset through the utilization of transfer learning on public datasets and careful choice of architecture. On the PDAC dataset, our private model at $\varepsilon = 8.0$ is not statistically significantly inferior compared to the non-private baseline. (2) We investigate the fairness implications of differentially private learning with respect to key demographic characteristics such as sex, age and co-morbidity. We find that – while differentially private learning has a mild fairness effect – it does not introduce significant discrimination concerns based on the subgroup representation compared to non-private training, especially at the intermediate privacy budgets typically used in large-scale applications.

Methods

Patient cohorts

We employed UKA-CXR[35,36], a large cohort of chest radiographs. The dataset consists of $N = 193{,}311$ frontal CXR images of 45,016 patients, all manually labeled by radiologists. The available labels include: pleural effusion, pneumonic infiltrates, and atelectasis, each separately for right and left lung, congestion, and cardiomegaly. The labeling system for cardiomegaly included five classes "normal," "uncertain," "borderline," "enlarged," and "massively enlarged." For the rest of the labels, five classes of "negative," "uncertain," "mild," "moderate," and "severe" were used. Data were split into $N = 153{,}502$ training and $N = 39{,}809$ test images using patient-wise stratification, but otherwise completely random allocation[35,36]. There was no overlap between the training and test sets. Supplementary Table 1 shows the statistics of the dataset, which are further visualized in Supplementary Figs. 1 and 2.

In addition, we used an in-house dataset at Klinikum Rechts der Isar of 1625 abdominal CT scans from unique, consecutive patients, of which 867 suffered from pancreatic ductal adenocarcinoma (PDAC) (positive) and 758 were a control group without a tumor (negative). We split the dataset into 975 train and 325 validation and test images respectively. During splitting we maintained the ratio of positive and negative samples in all subsets.

The experiments were performed in accordance with relevant national and international guidelines and regulations. Approval for the UKA-CXR dataset by the Ethical Committee of the Medical Faculty of RWTH Aachen University has been granted for this retrospective study (Reference No. EK 028/19). Analogously, for the PDAC dataset, the protocol was approved by the Ethics Committee of Klinikum Rechts der Isar (Protocol Number 180/17S). Both institutional review boards did not require informed consent from subjects and/or their legal guardian(s) as this was a retrospective study. The study was conducted in accordance with the Declaration of Helsinki.

Data pre-processing

We resized all images of the UKA-CXR dataset to (512×512) pixels. Afterward, a normalization scheme as described previously by Johnson et al.[37] was utilized by subtracting the lowest value in the image, dividing by the highest value in the shifted image, truncating values, and converting the result to an unsigned integer, i.e., in the range of [0,255]. Finally, we performed histogram equalization by shifting pixel values towards 0 or towards 255 such that all pixel values 0 through 255 have approximately equal frequencies[37].

We selected a binary classification paradigm for each label. The "negative" and "uncertain" classes ("normal" and "uncertain" for cardiomegaly) were treated as negative, while the "mild," "moderate," and "severe" classes ("borderline," "enlarged," and "massively enlarged" for cardiomegaly) were treated as positive.

For the PDAC dataset, we clipped the voxel density values of all CT scans to an abdominal window from −150 to 250 Hounsfield units and resized to a shape of $224 \times 224 \times 128$ voxels.

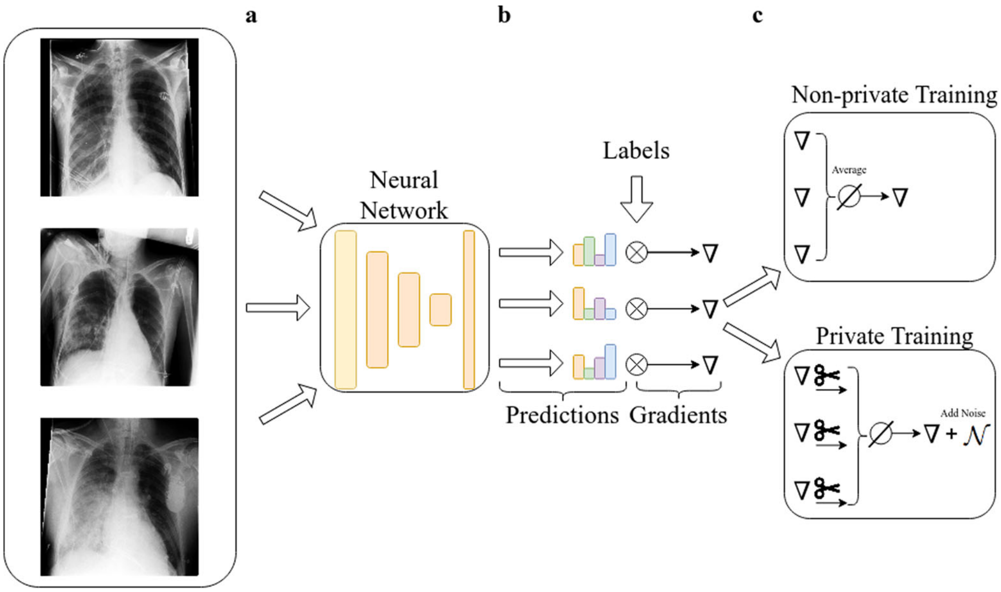


Fig. 1 | Differences between the private and non-private training process of a neural network. **a** Images from a dataset are fed to a neural network and predictions are made. **b** From the predictions and the ground truth labels, the gradient is calculated via backpropagation. ((**c**), upper panel) In normal training all gradients are averaged and an update step is performed. ((**c**), lower panel) In private training, each per-sample gradient is clipped to a predetermined ℓ_2-norm, averaged and noise proportional to the norm is added. This ensures that the information about each sample is upper-bounded and perturbed with sufficient noise.

Table 1 | Summary of dataset statistics and results

UKA-CXR																	
		Total		Male		Female		[0,30)		[30,60)		[60,70)		[70,80)		[80,100)	
Train	N	153,502		100,659		52,843		4279		42,340		36,882		48,864		21,137	
Test	N	39,809		25,360		14,449		1165		10,291		10,025		12,958		5370	
	Cardiomegaly	18,616		12,868		5748		334		3853		4714		6837		2876	
	Congestion	3275		2206		1069		50		817		906		991		510	
	Pl. Eff. R.	3275		2090		1185		52		709		847		1248		419	
	Pl. Eff. L.	2602		1636		966		70		589		632		894		417	
	Pn. Inf. R.	4847		3374		1473		184		1322		1367		1361		612	
	Pn. Inf. L.	3562		2381		1181		143		1087		949		959		423	
	Atel. R.	3920		2571		1349		127		1010		1056		1272		454	
	Atel. L.	3166		2010		1156		119		867		774		961		444	
	ε	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
AUROC	0.29	83.13	3.9	82.66	3.9	83.85	4.0	86.47	3.5	85.21	3.9	83.03	3.6	81.66	4.5	81.27	4.3
	0.54	84.00	3.8	83.61	3.8	84.61	3.9	86.43	3.1	85.96	3.7	83.96	3.5	82.69	4.5	82.15	4.2
	1.06	84.98	3.9	84.69	3.9	85.40	4.0	87.69	3.2	86.90	3.7	84.95	3.8	83.70	4.5	82.96	4.3
	2.04	85.80	3.9	85.52	3.9	86.19	3.9	88.77	3.3	87.53	3.8	85.88	3.8	84.47	4.4	83.85	4.3
	4.71	86.93	4.0	86.73	4.1	87.19	4.0	89.11	3.3	88.59	3.9	86.80	3.7	85.89	4.7	85.08	4.6
	7.89	87.36	4.1	87.12	4.2	87.66	4.1	89.72	4.1	88.97	3.9	87.26	3.9	86.30	4.7	85.48	4.8
	∞	89.71	3.8	89.46	3.9	90.06	3.8	91.64	3.5	90.99	3.4	89.73	3.8	88.73	4.4	88.18	4.5
PtD	0.29			−1.40	0.22	+1.40	0.22	+7.05	0.18	+0.98	0.73	+0.97	1.75	−1.73	0.36	−1.63	1.00
	0.54			−1.56	0.10	+1.56	0.10	+7.20	0.21	+0.80	0.52	+1.95	0.48	−2.65	0.31	−1.23	0.56
	1.06			−0.87	0.73	+0.87	0.73	+7.35	0.51	+2.56	0.67	+0.49	0.23	−1.92	0.78	−3.12	0.13
	2.04			+0.15	0.42	−0.15	0.42	+6.12	0.92	+1.80	0.39	+1.50	0.00	−2.80	0.30	−1.61	0.15
	4.71			−1.63	0.31	+1.63	0.31	+4.37	0.18	+2.15	0.70	+1.26	1.38	−2.27	0.50	−2.36	2.38
	7.89			−0.66	0.75	+0.66	0.75	+5.53	0.92	+1.27	0.04	+1.21	0.22	−1.33	0.06	−2.89	0.52
	∞			−0.34	0.47	+0.34	0.47	+4.00	0.60	+1.32	0.65	+0.21	0.66	−0.43	0.95	−2.67	0.20

PDAC															
		Total		Male		Female		Youngest 25%		Second 25%		Third 25%		Oldest 25%	
Train	N	975		552		423		231		290		228		226	
Test	N	325		197		127		86		85		79		75	
	Tumor	173		95		77		23		48		54		48	
	Control	152		102		50		63		37		25		27	
	ε	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
AUROC	0.29	86.84	4.0	88.11	4.6	85.47	2.5	87.92	9.1	85.87	3.0	84.44	1.2	89.15	7.2
	0.54	92.60	1.3	93.62	1.5	91.00	0.9	93.77	3.2	91.97	1.2	90.05	0.4	95.63	2.3
	1.06	95.58	0.9	96.70	0.9	93.52	1.3	96.57	1.6	94.84	1.3	93.83	1.1	98.43	0.9
	2.04	97.49	0.4	98.50	0.3	95.36	0.9	97.98	0.9	96.90	0.8	97.06	0.9	99.36	0.6
	4.71	98.31	0.2	99.19	0.1	96.38	0.7	98.48	0.3	97.84	0.2	98.30	0.4	99.97	0.0

Table 1 (continued) | Summary of dataset statistics and results

	ε	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
	5.0	98.33	0.2	99.20	0.1	96.41	0.7	98.48	0.4	97.86	0.1	98.37	0.4	100.00	0.0
	6.0	98.39	0.3	99.22	0.1	96.55	0.8	98.57	0.3	97.84	0.2	98.35	0.5	100.00	0.0
	7.0	98.41	0.3	99.22	0.1	96.60	0.8	98.62	0.3	97.88	0.1	98.25	0.5	100.00	0.0
	8.0	99.28	0.7	99.77	0.3	98.13	1.6	99.59	0.7	99.23	1.2	98.37	0.9	100.00	0.0
	∞	99.70	0.2	99.97	0.1	99.01	0.6	99.98	0.0	99.94	0.1	98.47	0.9	100.00	0.0
PtD	0.29			+3.27	5.38	−3.27	5.38	+9.03	1.32	+1.87	2.12	−9.54	4.33	−2.04	4.74
	0.54			+1.02	0.76	−1.02	0.76	+3.17	0.54	+0.34	1.44	−7.02	4.39	+3.42	3.28
	1.06			+1.29	1.27	−1.29	1.27	−0.18	3.85	+0.20	1.66	−3.58	3.53	+3.69	2.83
	2.04			+3.00	0.78	−3.00	0.78	−1.97	0.65	−3.16	2.47	+1.55	0.62	+4.00	3.47
	4.71			+4.58	1.33	−4.58	1.33	−3.29	1.23	−2.34	1.20	+1.47	1.46	+4.62	1.46
	5.0			+4.85	1.37	−4.85	1.37	−2.62	0.82	−2.73	1.16	+1.61	0.87	+4.18	1.90
	6.0			+4.41	0.53	−4.41	0.53	−2.10	2.06	−2.20	0.64	+1.05	2.10	+3.60	1.14
	7.0			+3.19	1.27	−3.19	1.27	−1.99	2.97	−3.68	1.21	+1.20	2.31	+4.93	2.02
	8.0			+3.28	2.61	−3.28	2.61	−2.45	1.51	−1.45	2.28	+1.58	2.44	+2.62	1.61
	∞			+2.81	2.38	−2.81	2.38	−1.21	1.59	−0.18	1.16	−0.33	1.71	+1.87	1.22

Diagnostic performance of patient sugroups for the UKA-CXR and PDAC datasets. We report the number of cases over subgroups and labels. All values refer to the test set. Total denotes the results on the entire test set. AUROC denotes the area under the receiver operating characteristic curve. PtD is the statistical parity difference of each subgroup. PDAC stands for presence of pancreatic ductal adenocarcinoma. μ are mean values, σ shows the standard deviation calculated over 1000 bootstrapping samples (UKA-CXR) respectively 3 independent model trainings (PDAC). All results are in percent.

Deep learning process

Network architecture. For both datasets, we employed the ResNet9 architecture introduced in ref. 38 as our classification architecture. For the UKA-CXR dataset, images were expanded to $(512 \times 512 \times 3)$ for compatibility with the neural network architecture. The final linear layer reduces the (512×1) output feature vectors to the desired number of diseases to be predicted, i.e., 8. The sigmoid function was utilized to convert the output predictions to individual class probabilities. The full network contained a total of 4.9 million trainable parameters. For the PDAC dataset, we used the conversion proposed by Yang et al.[39] to convert the model to be applicable to 3D data, which in brief applies 2D-convolutional filters along axial, coronal, and sagittal axes separately. Our utilized ResNet9 network employs the modifications proposed by Klause et al.[38] and by He et al.[40]. Batch Normalization[41] is incompatible with DP-SGD, as per-sample gradients are required, and batch normalization inherently intermixes information of all images in one batch. Hence, we used group normalization[42] layers instead with 32 groups to be compatible with DP processing. For the CXR dataset we pretrained the network on the MIMIC Chest X-ray JPG dataset v2.0.0 (MIMIC-CXR),[43] consisting of $N = 210{,}652$ frontal images. All training hyperparameters were selected empirically based on their validation accuracy, while no systematic/automated hyperparameter tuning was conducted.

Non-DP training. For the UKA-CXR dataset, the Rectified Linear Unit (ReLU)[44,45] was chosen as the activation function in all layers. We performed data augmentation during training by applying random rotation in the range of $[-10, 10]$ degrees and medio-lateral flipping with a probability of 0.50. The model was optimized using the NAdam[46] optimizer with a learning rate of $5 \cdot 10^{-5}$. The binary weighted cross-entropy with inverted class frequencies of the training data was selected as the loss function. The training batch size was chosen to be 128. In the PDAC dataset, we used an unweighted binary cross-entropy loss as well as the NAdam optimizer with a learning rate of $2 \cdot 10^{-4}$.

DP training. For UKA-CXR, we chose Mish[47] as the activation function in all layers. No data augmentation was performed during DP training as we found further data augmentation during training to be harmful to accuracy. All models were optimized using the NAdam[46] optimizer with a learning rate of $5 \cdot 10^{-4}$. The binary weighted cross-entropy with inverted class frequencies of the training data was selected as the loss function. The maximum allowed gradient norm (see Fig. 1) was chosen to be 1.5 and the network was trained for 150 epochs for each chosen privacy budget. Each point in the batch was sampled with a probability of $8 \cdot 10^{-4}$ (128 divided by $N = 153{,}502$). For the PDAC dataset, we chose a clipping norm of 1.0, $\delta = 0.001$ and a sampling rate of 0.31 (512/1 625). In both cases, the noise multiplier was calculated such that for a given number of training steps, sampling rate, and maximum gradient norm the privacy budget was reached on the last training step. For the UKA-CXR dataset, the indicated privacy guarantees are "per record" since some patients have more than one image, while for the PDAC datasets, they are "per individual."

Quantitative evaluation and statistical analysis

The area under the receiver operating characteristic curve (AUROC) was utilized as the primary evaluation metric. We report the average AUROC over all the labels for each experiment. The individual AUROC as well as all other evaluation metrics of individual labels are reported in the supplementary information (Supplementary Tables 2–8). For the UKA-CXR test set, we used bootstrapping with 1000 redraws for each measure to determine the statistical spread[48]. For calculating sensitivity, specificity, and accuracy, a threshold was chosen according to Youden's criterion[49], i.e., the threshold that maximized (true positive rate − false positive rate).

To evaluate the correlation between results of data subsets and their sample size, Pearson's r coefficient was used. To analyze fairness between subgroups, the statistical parity difference[50] was used which is defined as

$$P(\hat{Y} = 1 | C = \text{ Minority }) - P(\hat{Y} = 1 | C = \text{ Majority }) \qquad (2)$$

where $\hat{Y} = 1$ represents correct model predictions and C is the group in question. Intuitively, it is the difference in classification accuracy between the minority and majority class and thus is optimally zero. Values larger than zero mean that there is a benefit for the minority class, while values smaller than zero mean that the minority class is discriminated against.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

High classification accuracy is attainable despite stringent privacy guarantees

Table 1 shows an overview of our results for all subgroups. Supplementary Tables 2–8 show the per-diagnosis evaluation results for non-DP and DP training for different ε values. On the UKA-CXR dataset our non-private model achieves an AUROC of 89.71% over all diagnoses. It performs best on pneumonic infiltration on the right (AUROC=94%) while struggling the most to accurately classify cardiomegaly (AUROC=84%). Training with DP decreases all results slightly yet significantly (Hanley & McNeil-test p-value < 0.001, 1 000 bootstrapping redraws) and achieves an overall AUROC of 87.36%. The per-diagnosis performance ranges from 92% (pleural effusion right) to 81% AUROC (congestion). We next consider classification performance at a very strong level of privacy protection (i.e., at $\varepsilon < 1$). Here, at an ε-budget of only 0.29, our model achieves an average AUROC of 83.13% over all diagnoses. A visual overview is displayed in Fig. 2, which shows the average AUROC, accuracy, sensitivity, and specificity values over all labels.

On the PDAC dataset, we found that, while non-private training achieved almost perfect results on the test set the loss in utility for private training at $\varepsilon = 8$ is statistically non-significant (Hanley & McNeil-test p-value: 0.34, 3 independent experiments) compared to non-private training. Again, with lower privacy budgets, model utility decreases, but even at a very low privacy budget of $\varepsilon = 1.06$, we observe an average AUROC score of 95.58%.

Moreover, for UKA-CXR, the use of pre-training helps to boost model performance and reduce the amount of additional information the model needs to learn "from scratch" and consequently reduces the privacy budgets required (refer to Supplementary Fig. 3). This appears to primarily benefit the under-represented groups in the dataset. Conversely, non-private training, whether initialized with pre-training weights or trained from scratch, tends to yield comparable diagnostic results, as the latter network can leverage a greater amount of information. These findings are in line with the observations on the PDAC dataset (where no pretrained weights were available), namely that, at low privacy budgets, specific patient groups suffer a higher discrimination.

For the purpose of further generalization, we replicated the experiments using three other network architectures. All three models displayed a trend consistent with the utility penalties we observed for ResNet9 in both DP and non-DP training (see Supplementary Fig. 4). For further details, we refer to the supplementary information.

Diagnostic accuracy is correlated with patient age and sample size for both private and non-private models

Fig. 3 shows the difference in classification performance on the UKA-CXR dataset for each diagnosis between the non-private model evaluation and its

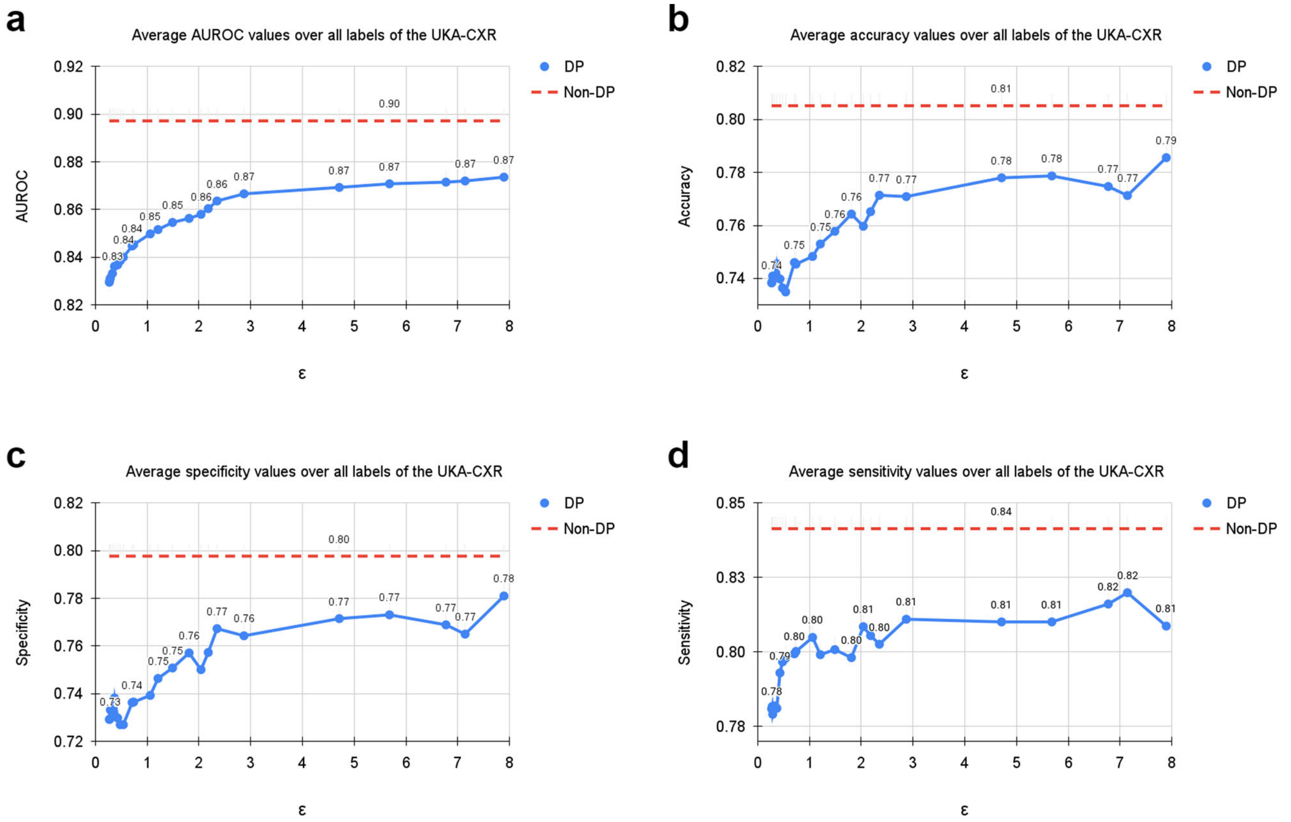


Fig. 2 | Average results of training with differential privacy (DP) with different ϵ values for $\delta = 6 \cdot 10^{-6}$. The curves show the average (**a**) area under the receiver operating characteristic curve (AUROC), (**b**) accuracy, (**c**) specificity, and (**d**) sensitivity values over all labels, including cardiomegaly, congestion, pleural effusion right, pleural effusion left, pneumonic infiltration right, pneumonic infiltration left, atelectasis right, and atelectasis left tested on N = 39,809 test images. The training dataset includes N = 153,502 images. Note, that the AUROC is monotonically increasing, while sensitivity, specificity and accuracy exhibit more variation. This is due to the fact that all training processes were optimized for the AUROC. Dashed lines correspond to the non-private training results. Source data are provided as a Source Data file.

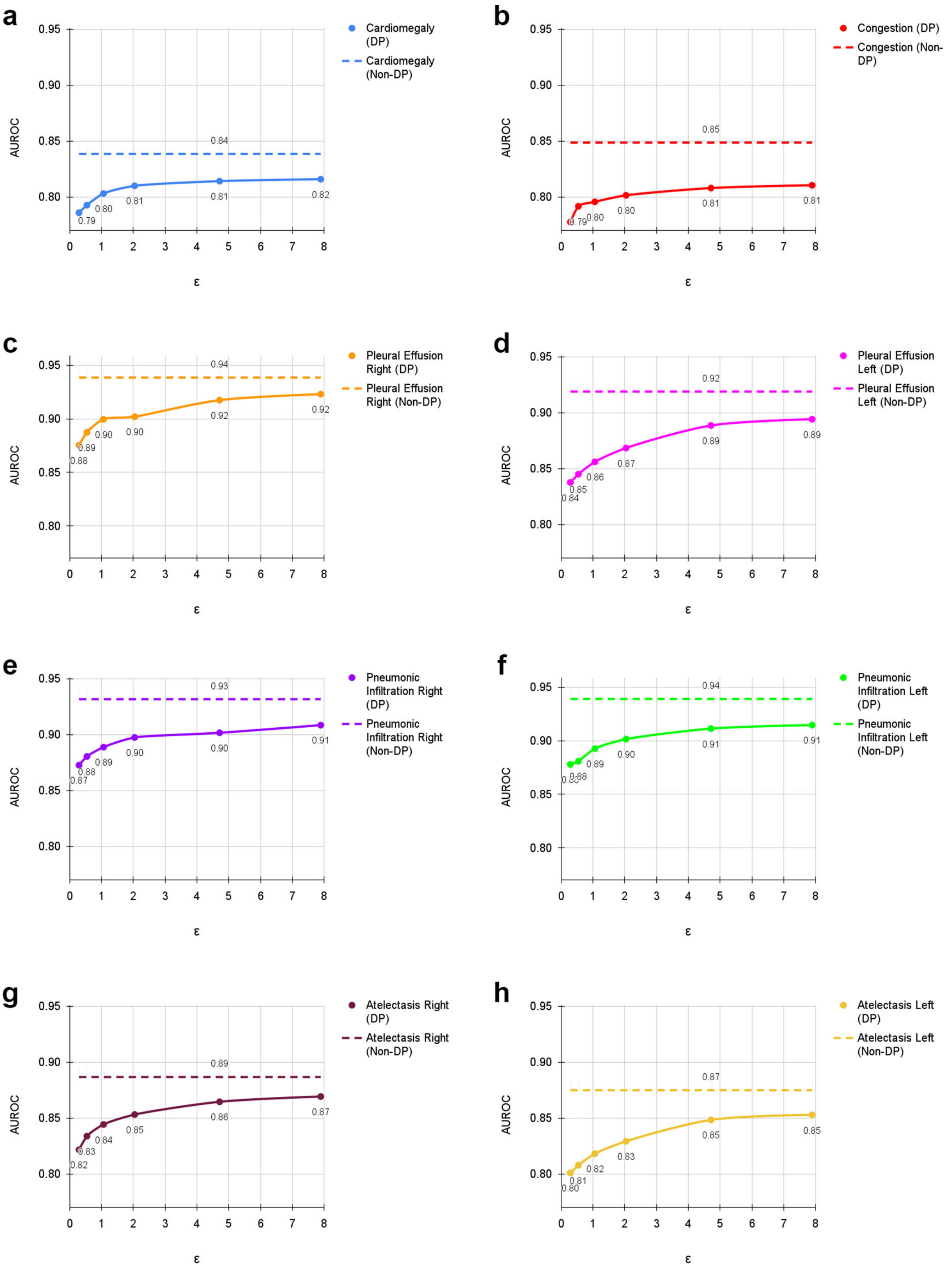


Fig. 3 | Evaluation results of training with differential privacy (DP) and without DP with different ϵ values for $\delta = 6 \cdot 10^{-6}$. The results show the individual area under the receiver operating characteristic curve (AUROC) values for (**a**) cardiomegaly, (**b**) congestion, (**c**) pleural effusion right, (**d**) pleural effusion left, (**e**) pneumonic infiltration right, (**f**) pneumonic infiltration left, (**g**) atelectasis right, and (**h**) atelectasis left tested on $N = 39{,}809$ test images. The training dataset includes $N = 153{,}502$ images. Dashed lines correspond to the non-private training results. Source data are provided as a Source Data file.

private counterpart compared to the sample size (that is, the number of available samples with a given label) within our dataset. At an ε = 7.89, the largest difference of AUROC between the non-private and privacy-preserving model was observed for congestion (3.82%) and the smallest difference was observed for pleural effusion right (1.55%, see Fig. 3). Of note, there is a visible trend (Pearson's r: 0.44) whereby classes in which the model exhibits good diagnostic performance in the non-private setting also suffer the smallest drop in the private setting. On the other hand, classes that are already difficult to predict in the non-private case deteriorate the most in terms of classification performance with DP (see Supplementary Fig. 9). Both non-private (Pearson's r: 0.57) and private (Pearson's r: 0.52) diagnostic AUROC exhibit a weak correlation with the number of samples available for each class (see Supplementary Fig. 9). However, the drop in AUROC between private and non-private training is not correlated with the sample size (Pearson's r: 0.06). On the PDAC dataset, patients with a tumor are overrepresented and in the non-private case diagnosed more accurately. Not surprisingly, the classification performance is thus also higher for private trainings except for the most restrictive privacy budget (see Supplementary Figs. 5–8).

Furthermore, we evaluated our models based on age range and patient sex (Table 1 and Figs. 4 and 5). Additionally, we calculated statistical parity difference for those groups to obtain a measure of fairness (Table 1). On the UKA-CXR dataset all models performed the best on patients younger than 30 years of age. It appears that, the older patients are, the greater the difficulty for the models to predict the labels accurately. Statistical parity difference scores are slightly negative for the age groups between 70 and 80 years and older than 80 years for all models, indicating that the models discriminate slightly against these groups. In addition, while for the aforementioned age groups the discrimination does not change with privacy levels, younger patients become more privileged as privacy increases. This finding indicates that – for models which are most protective of data privacy – young patients benefit the most, despite the group of younger patients being smaller overall. For patient sex, models show slightly better performance for female patients and slightly discriminate against male patients (Table 1). Statistical parity does not appear to correlate (Pearson's r: 0.13) with privacy levels.

On the PDAC dataset, we observed that, for all levels of privacy including non-private training, classification performance was worse for female patients compared to male patients, who are over-represented in the dataset. However, there is no trend observable between the privacy level and the parity difference. When analysing results of subgroups separated by patient age, we observed similarly to UKA-CXR that in all settings, statistical parity differences are on average better for younger patients compared to older ones. Just as in the UKA-CXR dataset, we found that the more restrictive the privacy budget is set, the stronger the privilege enjoyed by younger patients. We furthermore observed that the control group (i.e., no tumor) has an over-representation of both male patients and young patients, which consequently both exhibit better performance compared to the rest of the cohort. Conversely, female patients as well as older patients, have a higher chance of misclassification and are more abundant in the tumor group.

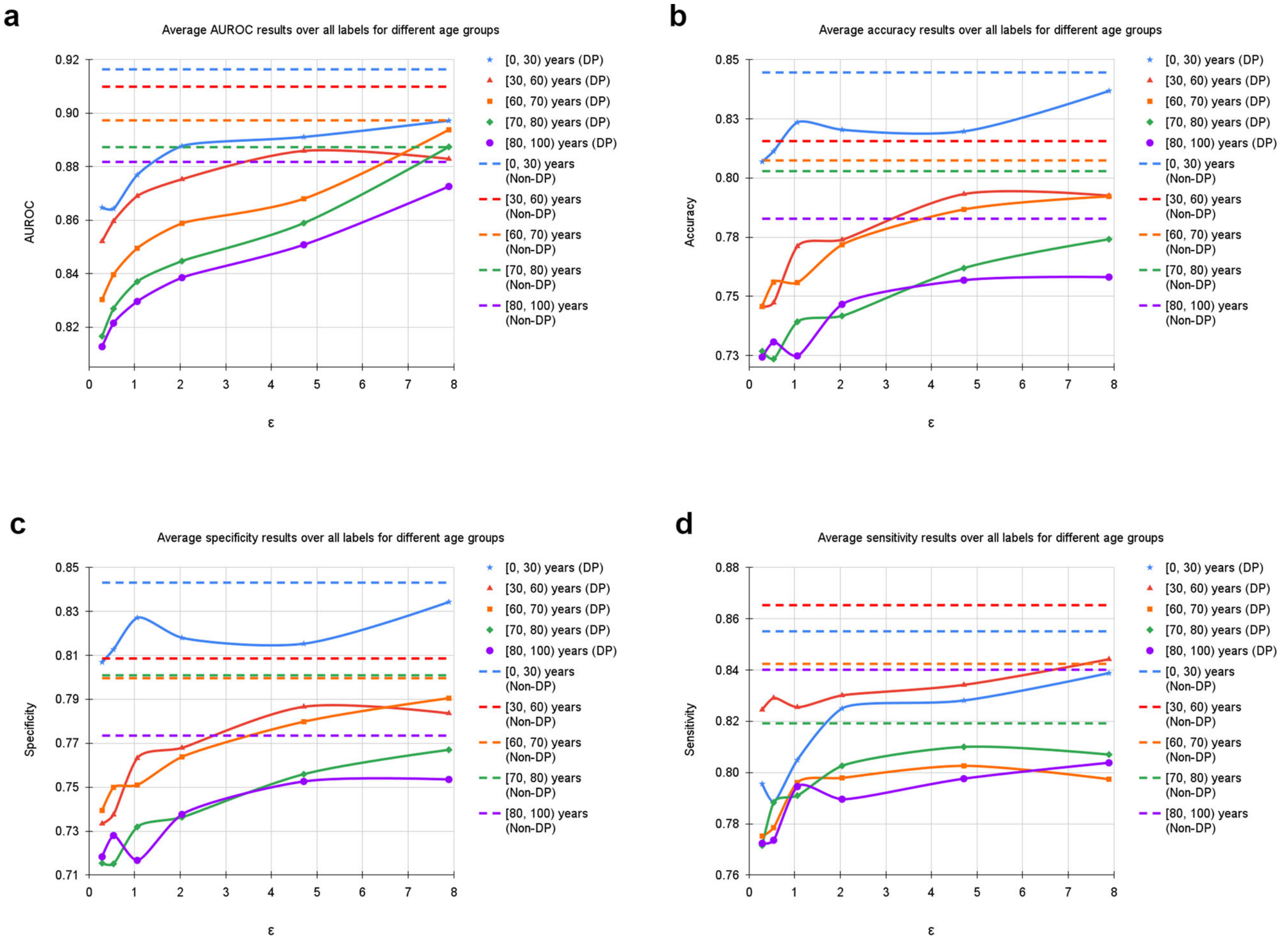


Fig. 4 | Average results of training with differential privacy (DP) with different ϵ values for $\delta = 6 \cdot 10^{-6}$, separately for samples of different age groups including [0, 30), [30, 60), [60, 70), [70, 80), and [80, 100) years. The curves show the average (**a**) area under the receiver operating characteristic curve (AUROC), (**b**) accuracy, (**c**) specificity, and (**d**) sensitivity values over all labels, including cardiomegaly, congestion, pleural effusion right, pleural effusion left, pneumonic infiltration right, pneumonic infiltration left, atelectasis right, and atelectasis left tested on $N = 39{,}809$ test images. The training dataset includes $N = 153{,}502$ images. Dashed lines in corresponding colors correspond to the non-private training results. Source data are provided as a Source Data file.

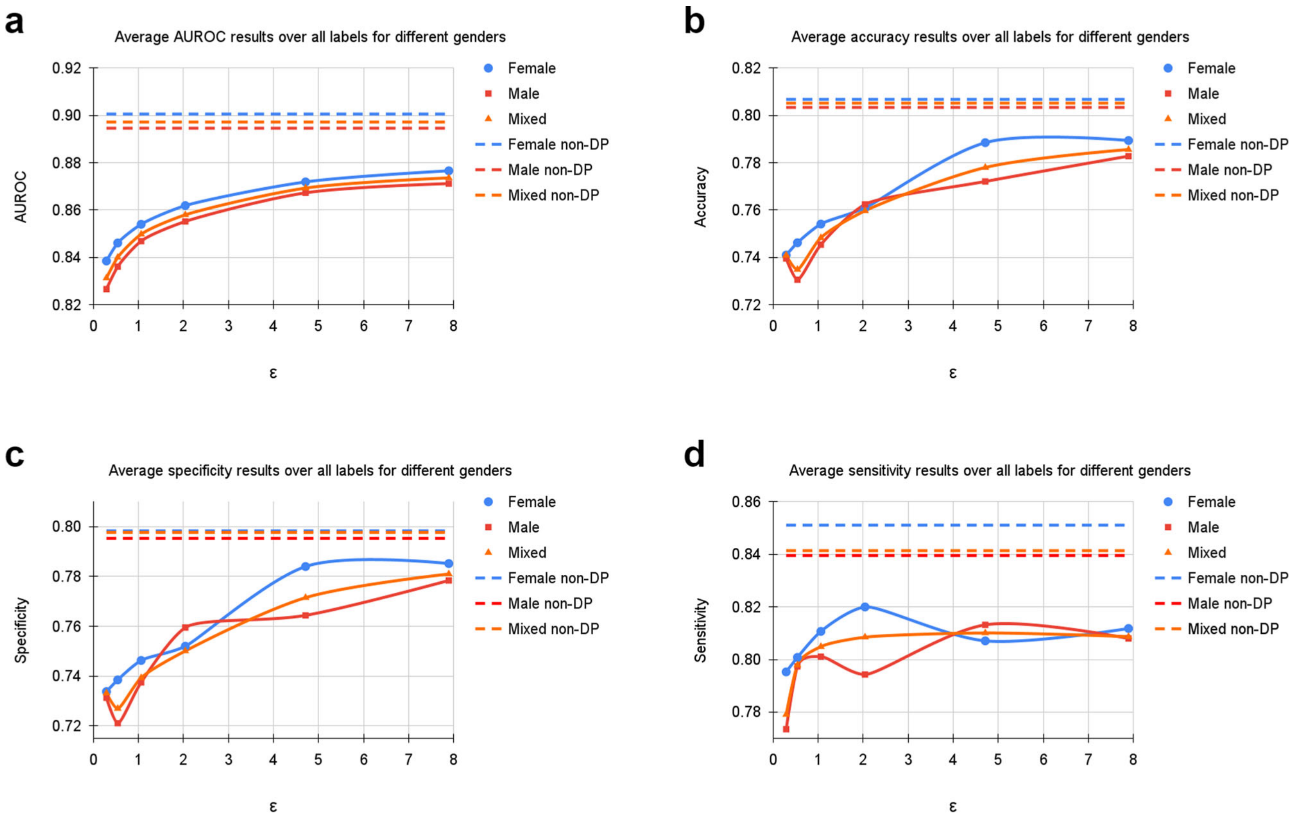


Fig. 5 | Average results of training with differential privacy (DP) with different ϵ values for $\delta = 6 \cdot 10^{-6}$, separately for female and male samples. The curves show the average (**a**) area under the receiver operating characteristic curve (AUROC), (**b**) accuracy, (**c**) specificity, and (**d**) sensitivity values over all labels, including cardiomegaly, congestion, pleural effusion right, pleural effusion left, pneumonic infiltration right, pneumonic infiltration left, atelectasis right, and atelectasis left tested on $N = 39{,}809$ test images. The training dataset includes $N = 153{,}502$ images. Note, that the AUROC is monotonically increasing, while sensitivity, specificity and accuracy exhibit more variation. This is due to the fact that all training processes were optimized for the AUROC. Dashed lines correspond to the non-private training results depicted as upper bounds. Source data are provided as a Source Data file.

Discussion

The main contribution of our paper is to analyse the impact of strong objective guarantees of privacy on the fairness enjoyed by specific patient subgroups in the context of AI model training on real-world medical datasets.

Across all levels of privacy protection, training with DP still yielded models exhibiting AUROC scores of 83% at the highest privacy level and 87% at an $\varepsilon = 7.89$ on the UKA-CXR dataset. The fact that the model maintained a relatively high AUROC even at $\varepsilon = 0.29$ is remarkable, and we are unaware of any prior work to report such a strong level of privacy protection at this level of model accuracy on clinical data. Our results thus exemplify that, through careful choice of architectures and best practices for the training of DP models, the use of model pretraining on a related public dataset, and the availability of sufficient data samples, privately trained models require only very small additional amounts of private information from the training dataset to achieve high diagnostic accuracy on the tasks at hand.

For the PDAC dataset, even though private models at $\varepsilon = 8.0$ are not significantly inferior compared to non-private counterparts, the effect of the lower amount of training samples is observable at more restrictive privacy budgets. Especially at $\varepsilon \leq 1.06$, the negative effect of private training on the discrimination of patients in certain age groups becomes noticeable. This underscores the requirement for larger training datasets, which the objective privacy guarantees of DP can enable through incentivizing data sharing.

Our analysis of the per-diagnosis performance of models that are trained with and without privacy guarantees shows that models discriminate against diagnoses that are underrepresented in the training set in both private and non-private training. This finding is not unusual and several examples can be found in[51]. However, the drop in performance between private and non-private training is uncorrelated to the sample size. Instead, the difficulty of the diagnosis seems to drive the difference in AUROC between the two settings. Concretely, diagnostic performance under privacy constraints suffers the most for those classes, which already have the lowest AUROC in the non-private setting. Conversely, diagnoses that are predicted with the highest AUROC suffer the least when DP is introduced.

Previous works investigating the effect of DP on fairness show that privacy preservation amplifies discrimination[33]. This effect is limited to very low privacy budgets in our study. Our models remain fair despite at the levels of privacy protection typically used for training state-of-the-art models in current literature[25], likely due to our real-life datasets' large size and/or high quality.

The effects we observed are not limited to within-domain models. Indeed, in a concurrent work, we investigated the effects of DP training on the domain generalizability of diagnostic medical AI models[52]. Our findings revealed that even under extreme privacy conditions, DP-trained models show comparable performance to non-DP models in external domains.

Our analysis of fairness related to patient age showed that older patients are discriminated against both in the non-private and private settings. On UKA-CXR, age-related discrimination remains approximately constant with stronger privacy guarantees. On the other hand, young patients enjoy overall lower model discrimination in the non-private and the private setting. Interestingly, young patients seem to profit more from stronger privacy guarantees, as they enjoy progressively more fairness privilege with increasing privacy protection level.

This holds despite the fact that patients under 30 represent the smallest fraction of the UKA-CXR dataset. The privilege of young patients is most likely due to a confounding variable, namely the lower complexity of imaging findings in younger patients due to their improved ability to cooperate during radiograph acquisition, resulting in better discrimination of the pathological finding on a more homogeneous background (i.e., "cleaner") radiographs which are easier to diagnose overall[35,53] (see Fig. 6). This hypothesis should be validated in cohorts with a larger proportion of young patients, and we intend to expand on this finding in future work. On the PDAC dataset, classification accuracy remains approximately on par between age subgroups except at very restrictive privacy budgets, where older patients begin to suffer discrimination, likely due to the aforementioned imbalance between control and tumor cases and the overall smaller dataset coupled with a lack of pre-training. The analysis of model fairness related to patient sex for UKA-CXR shows that female patients (which – similar to young patients – are an underrepresented group) enjoy a slightly higher diagnostic accuracy than male patients for almost all privacy levels and vice versa on the PDAC dataset. However, effect size differences were found to be small, so that this finding can also be explained by variability between models or by the randomness in the training process. Further investigation is thus required to elucidate the aforementioned effects.

Furthermore, there is no final conclusion for which fairness measure is preferable. In our study we focused on the statistical parity difference, however, there are other works proposing other measures. One, which recently received attention, is the underdiagnosis rate of subgroups[54]. We evaluated this for the PDAC dataset and found that in principle it shows the same trends as the statistical parity difference (see Supplementary Tables 9 and 10).

In conclusion, we analyzed the usage of privacy-preserving neural network training and its implications on utility and fairness for a relevant diagnostic task on a large real-world dataset. We showed that the utilization of specialized architectures and targeted model pre-training allows for high model accuracy despite stringent privacy guarantees. This enables us to train expert-level diagnostic AI models even with privacy budgets as low as $\varepsilon < 1$, which – to our knowledge – has not been shown before, and represents an important step towards the widespread utilization of differentially private models in radiological diagnostic AI applications. Moreover, our findings that the introduction of differential privacy mechanisms to model training does – in most cases – not amplify unfair model bias regarding patient age, sex or comorbidity signifies that – at least in our use case – the resulting models abide by important non-discrimination principles of ethical AI. We are hopeful that our findings will encourage practitioners and clinicians to introduce advanced privacy-preserving techniques such as differential privacy when training diagnostic AI models.

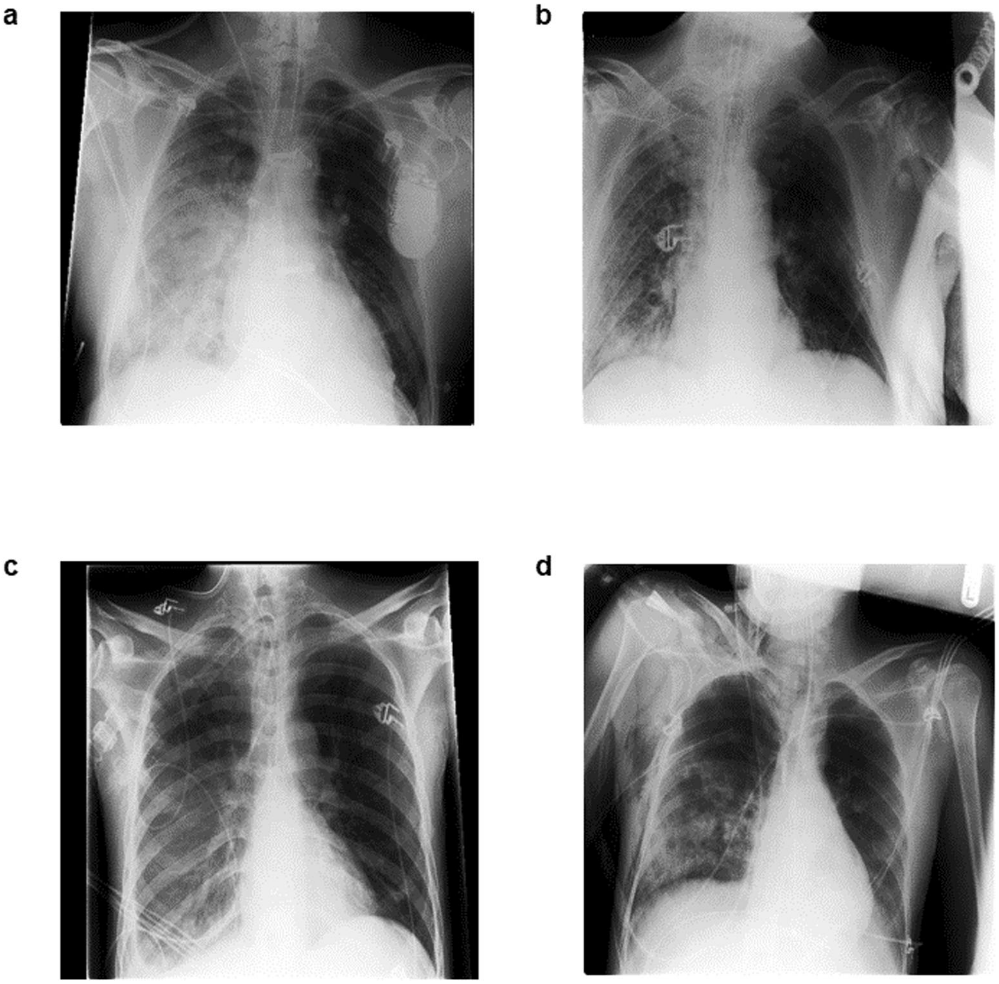


Fig. 6 | Illustrative radiographs from the UKA-CXR dataset. All examinations share the diagnosis of pneumonic infiltrates on the right patient side (=left image side). Diagnosis in older patients is often more challenging due to the more frequent presence of comorbidities and less cooperation during image acquisition which results in lower image quality (**a**) 76-year-old male patient, note the presence of a cardiac pacemaker that projects over part of the left lung. **b** 74-year-old male patient with challenging image acquisition: part of the lower right lung is not properly depicted. **c** 39-year-old male patient, the lungs are well inflated and pneumonic infiltrates can be discerned even though they are less severe. **d** 33-year-old male patient with challenging image acquisition, yet both lungs can be assessed (almost) completely.

Data availability

The UKA-CXR dataset is not publicly accessible, in adherence to the policies for patient privacy protection at the University Hospital RWTH Aachen in Aachen, Germany. Similarly, the PDAC dataset cannot be publicly shared due to patient privacy considerations, as it is an in-house dataset at Klinikum Rechts der Isar, Munich, Germany. Data access for both datasets can be granted upon reasonable request to the corresponding author. Source data presented in Figures are available as Supplementary Data 1.

Code availability

All source codes used for UKA-CXR for training and evaluation of the deep neural networks, differential privacy, data augmentation, image analysis, and preprocessing are publicly available at https://github.com/tayebiarasteh/DP_CXR. All code for the experiments was developed in Python 3.9 using the PyTorch 2.0 framework. The DP code was developed using Opacus 1.4.0[55]. Considering the utilization of equivalent computational resources, the time taken for the DP training to converge was approximately 10 times longer, in terms of total training time, than that required for the non-DP training with a similar network architecture. All code for the analyses on the PDAC dataset are available at https://github.com/TUM-AIMED/2.5DAttention. All source codes for both datasets are permanently archived on Zenodo and are accessible via[56] and[57].

Received: 6 April 2023; Accepted: 16 February 2024;
Published online: 14 March 2024

References

1. Usynin, D. et al. Adversarial interference and its mitigations in privacy-preserving collaborative machine learning. *Nat. Mach. Intell.* **3**, 749–758 (2021).
2. Konečny`, J., McMahan, H. B., Ramage, D. & Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527 (2016).
3. Konečny`, J. et al. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492 (2016).
4. McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 1273–1282 (PMLR, 2017).
5. Truhn, D. et al. Encrypted federated learning for secure decentralized collaboration in cancer image analysis. *Med. Image Anal.* (2024). https://doi.org/10.1016/j.media.2023.103059.
6. Dwork, C. & Roth, A. et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407 (2014).
7. Boenisch, F. et al. When the curious abandon honesty: Federated learning is not private. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, 175–199 (IEEE, 2023).
8. Fowl, L., Geiping, J., Czaja, W., Goldblum, M. & Goldstein, T. Robbing the fed: Directly obtaining private data in federated learning with modified models. In *International Conference on Learning Representations* (2021).
9. Wang, K.-C. et al. Variational model inversion attacks. *Adv. Neural Inf. Process. Syst.* **34**, 9706–9719 (2021).
10. Haim, N., Vardi, G., Yehudai, G., Shamir, O. & Irani, M. Reconstructing training data from trained neural networks. *Adv. Neural Inf. Processing Syst.* **35**, 22911–22924 (2022).
11. Carlini, N. et al. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 5253–5270 (2023).
12. Food, U. & Administration, D. Artificial intelligence and machine learning (ai/ml)-enabled medical devices. Webpage (2023). https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.
13. Wasserman, L. & Zhou, S. A statistical framework for differential privacy. *J. Am. Stat. Assoc.* **105**, 375–389 (2010).
14. Dong, J., Roth, A. & Su, W. J. Gaussian differential privacy. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* **84**, 3–37 (2022).
15. Kaissis, G., Hayes, J., Ziller, A. & Rueckert, D. Bounding data reconstruction attacks with the hypothesis testing interpretation of differential privacy. *Theory and Practice of Differential Privacy Workshop* (2023).
16. Nasr, M. et al. Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, 1631–1648 (2023).
17. Kaissis, G. et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* **3**, 473–484 (2021).
18. Hayes, J., Mahloujifar, S. & Balle, B. Bounding training data reconstruction in dp-sgd. arXiv preprint arXiv:2302.07225 (2023).
19. Balle, B., Cherubin, G. & Hayes, J. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, 1138–1156 (IEEE, 2022).
20. Cohen, A. & Nissim, K. Towards formalizing the gdpr's notion of singling out. *Proc. Nat. Acad. Sci.* **117**, 8344–8352 (2020).
21. Cohen, A. Attacks on deidentification's defenses. In *31st USENIX Security Symposium (USENIX Security 22)*, 1469–1486 (2022).
22. Abadi, M. et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318 (2016).
23. Hatamizadeh, A. et al. Do gradient inversion attacks make federated learning unsafe? *IEEE Trans. Med. Imaging* (2023).
24. Dwork, C. A firm foundation for private data analysis. *Commun. ACM* **54**, 86–95 (2011).
25. De, S., Berrada, L., Hayes, J., Smith, S. L. & Balle, B. Unlocking high-accuracy differentially private image classification through scale. arXiv preprint arXiv:2204.13650 (2022).
26. Kurakin, A. et al. Toward training at imagenet scale with differential privacy. arXiv preprint arXiv:2201.12328 (2022).
27. Tran, C., Fioretto, F., Van Hentenryck, P. & Yao, Z. Decision making with differential privacy under a fairness lens. In *IJCAI*, 560–566 (2021).
28. Cummings, R., Gupta, V., Kimpara, D. & Morgenstern, J. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 309–315 (2019).
29. Packhäuser, K. et al. Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest x-ray data. *Sci. Rep.* **12**, 14851 (2022).
30. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 111–125 (IEEE, 2008).
31. Li, W. et al. Privacy-preserving federated brain tumour segmentation. In *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*, 133–141 (Springer, 2019).
32. Ziegler, J., Pfitzner, B., Schulz, H., Saalbach, A. & Arnrich, B. Defending against reconstruction attacks through differentially private federated learning for classification of heterogeneous chest x-ray data. *Sensors* **22**, 5195 (2022).
33. Farrand, T., Mireshghallah, F., Singh, S. & Trask, A. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 Workshop on Privacy-preserving Machine Learning in Practice*, 15–19 (2020).
34. Bagdasaryan, E., Poursaeed, O. & Shmatikov, V. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems* **32**, https://proceedings.neurips.cc/paper_files/paper/2019/hash/fc0de4e0396fff257ea362983c2dda5a-Abstract.html (2019).

35. Khader, F. et al. Artificial intelligence for clinical interpretation of bedside chest radiographs. *Radiology* **307**, e220510 (2022).
36. Tayebi Arasteh, S. et al. Collaborative training of medical artificial intelligence models with non-uniform labels. *Sci. Rep.* **13**, 6046 (2023).
37. Johnson, A. E. et al. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
38. Klause, H., Ziller, A., Rueckert, D., Hammernik, K. & Kaissis, G. Differentially private training of residual networks with scale normalisation. *Theory and Practice of Differential Privacy Workshop, ICML* (2022).
39. Yang, J. et al. Reinventing 2d convolutions for 3d images. *IEEE J. Biomed. Health Inform.* **25**, 3009–3018 (2021).
40. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
41. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456 (pmlr, 2015).
42. Wu, Y. & He, K. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19 (2018).
43. Johnson, A. et al. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet* (2019).
44. Fukushima, K. Cognitron: A self-organizing multilayered neural network. *Biol. Cybern.* **20**, 121–136 (1975).
45. Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 807–814 (2010).
46. Dozat, T. Incorporating nesterov momentum into adam. In *International Conference on Learning Representations, Workshop Track* (2016).
47. Misra, D. Mish: A self regularized non-monotonic activation function. In *The 31st British Machine Vision Conference* (2020).
48. Konietschke, F. & Pauly, M. Bootstrapping and permuting paired t-test type statistics. *Stat. Comput.* **24**, 283–296 (2014).
49. Unal, I. Defining an optimal cut-point value in roc analysis: an alternative approach. *Comput. Math. Methods Med.* **2017** (2017).
50. Calders, T. & Verwer, S. Three naive bayes approaches for discrimination-free classification. *Data Mining Knowl. Discov.* **21**, 277–292 (2010).
51. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **54**, 1–35 (2021).
52. Tayebi Arasteh, S. et al. Securing collaborative medical AI by using differential privacy: Domain transfer for classification of chest radiographs. *Radiol. Artif. Intel.* **6**, e230212 (2024).
53. Wu, J. T. et al. Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. *JAMA Netw. Open* **3**, e2022779–e2022779 (2020).
54. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).
55. Yousefpour, A. et al. Opacus: User-friendly differential privacy library in pytorch (2021). https://arxiv.org/abs/2109.12298.
56. Arasteh, S. T. DP CXR. https://doi.org/10.5281/zenodo.10361657 (2023).
57. Ziller, A. 2.5d attention. https://doi.org/10.5281/zenodo.10361128 (2023).

Acknowledgements

STA is funded and partially supported by the Radiological Cooperative Network (RACOON) under the German Federal Ministry of Education and Research (BMBF) grant number 01KX2021. AZ and GK were supported by the German Ministry of Education and Research (BMBF) under Grant Number 01ZZ2316C (PrivateAIM). DT is supported from the Deutsche Forschungsgemeinschaft (DFG) (TR 1700/7-1) as well as by the German Federal Ministry of Education (TRANSFORM LIVER, 031L0312A; SWAG, 01KD2215B) and the European Union's Horizon Europe and innovation programme (ODELIA [Open Consortium for Decentralized Medical Artificial Intelligence], 101057091). RB was supported by Deutsches Konsortium für Translationale Krebsforschung (DKTK). GK and DR have been funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts through the Munich Centre for Machine Learning. DR was supported by ERC Grant Deep4MI (no. 884622). The funders played no role in the design or execution of the study. The authors of this work take full responsibility for its content.

Author contributions

The formal analysis was conducted by S.T.A., A.Z., D.T. and G.K. The original draft was written by S.T.A. and A.Z. and edited by D.T. and G.K. The experiments as well as the software development for UKA-CXR were performed by S.T.A. and for PDAC by A.Z. Statistical analyses were performed by A.Z. and S.T.A. D.T. and G.K. provided clinical and technical expertise. S.T.A., A.Z., C.K., M.M., S.N., R.B., D.R., D.T. and G.K. read the manuscript and agreed to the submission of this paper.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-024-00462-6.

Correspondence and requests for materials should be addressed to Soroosh Tayebi Arasteh, Alexander Ziller, Daniel Truhn or Georgios Kaissis.

Peer review information *Communications Medicine* thanks Valeriu Codreanu, Holger Roth and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

3.4 Reconciling Privacy and Accuracy in AI for Medical Imaging

Synopsis: Artificial intelligence (AI) models are vulnerable to information leakage of their training data, which can be highly sensitive, for example, in medical imaging. Privacy-enhancing technologies, such as differential privacy (DP), aim to circumvent these susceptibilities. DP is the strongest possible protection for training models while bounding the risks of inferring the inclusion of training samples or reconstructing the original data. DP achieves this by setting a quantifiable privacy budget. Although a lower budget decreases the risk of information leakage, it typically also reduces the performance of such models. This imposes a trade-off between robust performance and stringent privacy. Additionally, the interpretation of a privacy budget remains abstract and challenging to contextualize. Here we contrast the performance of artificial intelligence models at various privacy budgets against both theoretical risk bounds and empirical success of reconstruction attacks. We show that using very large privacy budgets can render reconstruction attacks impossible, while drops in performance are negligible. We thus conclude that not using DP at all is negligent when applying artificial intelligence models to sensitive data. We deem our results to lay a foundation for further debates on striking a balance between privacy risks and model performance.

Contributions of thesis author: study design, code development, experiment design and evaluation, paper writing.

nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00858-y

Reconciling privacy and accuracy in AI for medical imaging

Received: 22 November 2023

Accepted: 20 May 2024

Published online: 21 June 2024

Check for updates

Alexander Ziller[1], Tamara T. Mueller[1], Simon Stieger[1,2], Leonhard F. Feiner[1,3], Johannes Brandt[1], Rickmer Braren[1,3,4], Daniel Rueckert[1,5] & Georgios Kaissis[1,2,3,5]

Artificial intelligence (AI) models are vulnerable to information leakage of their training data, which can be highly sensitive, for example, in medical imaging. Privacy-enhancing technologies, such as differential privacy (DP), aim to circumvent these susceptibilities. DP is the strongest possible protection for training models while bounding the risks of inferring the inclusion of training samples or reconstructing the original data. DP achieves this by setting a quantifiable privacy budget. Although a lower budget decreases the risk of information leakage, it typically also reduces the performance of such models. This imposes a trade-off between robust performance and stringent privacy. Additionally, the interpretation of a privacy budget remains abstract and challenging to contextualize. Here we contrast the performance of artificial intelligence models at various privacy budgets against both theoretical risk bounds and empirical success of reconstruction attacks. We show that using very large privacy budgets can render reconstruction attacks impossible, while drops in performance are negligible. We thus conclude that not using DP at all is negligent when applying artificial intelligence models to sensitive data. We deem our results to lay a foundation for further debates on striking a balance between privacy risks and model performance.

The rapid rise of artificial intelligence (AI) applications in medicine promises to transform healthcare, offering improvements ranging from specific applications, such as more precise pathology detection or outcome prediction, to the promise of general medical AI[1–5]. However, recent results highlight a substantial vulnerability: AI models may disclose details of their training data. This can happen either inadvertently or be forced through attacks by malicious third parties, also called adversaries. Among the most critical attacks are data reconstruction attacks, where the adversary attempts to extract training data from the model or its gradients[6–17]. Such attacks harbour distinct risks. On one hand, a successful data reconstruction attack severely undermines the trust of patients whose data are exposed. This not only jeopardises the relationship between medical practitioners and patients, but probably also diminishes the willingness of patients to make their health data for the training of AI models or for other research purposes available. This is problematic since the success of AI models in medicine is dependent on the availability of large and diverse real-world patient datasets. On the other hand, a successful attack can also constitute a breach of patient data privacy regulations.

While privacy laws vary globally, the protection of health data is generally considered of high importance. For example, the European Union's General Data Protection Regulation declares the protection of personal data as a fundamental right. Notably, some of these laws deem the removal of personal identifiers (for example, name or

[1]Artificial Intelligence in Healthcare and Medicine, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany. [2]Institute of Machine Learning in Biomedical Imaging, Helmholtz Munich, Neuherberg, Germany. [3]Institute for Diagnostic and Interventional Radiology, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany. [4]German Cancer Consortium (DKTK), Munich partner site, Heidelberg, Germany. [5]Department of Computing, Imperial College London, London, UK. ✉e-mail: alex.ziller@tum.de

date of birth)—de-identification—sufficient protection. However, it has been demonstrated on several occasions that commonly used de-identification techniques such as anonymization, pseudonymization or *k*-anonymity are vulnerable to re-identification attacks[18–20]. This also holds true in the case of medical imaging data. For example, the facial contours of a patient can be obtained from a reconstructed magnetic resonance imaging scan even if their name has been removed from the record, thus enabling their re-identification from publicly available photographs[21]. Figuratively, this is analogous to considering passport photos without additional information not as personal data. Arguably, this highlights the tension between what is considered 'private' in a legal sense and what individuals consider acceptable in terms of informational self-determination. We thus contend that AI systems that process sensitive data should not only rely on de-identification techniques but also implement privacy-enhancing technologies (PETs), that is, technologies that furnish an objective or formal guarantee of privacy protection.

DP as the optimal privacy preservation

Among PETs, differential privacy (DP)[22] is considered the optimal protection for training AI models while moderating the privacy risk faced by participating patients due to its appealing properties: it provides a formal upper bound on the success of reconstructing data[23,24] and satisfies requirements imposed by regulations such as the General Data Protection Regulation concerning re-identification[19,25]. Moreover, the privacy guarantees of DP cannot be degraded through the use of side information or through post-processing (two notable vulnerabilities of traditional de-identification schemes). Last but not least, DP satisfies composability, that is, its guarantee degrades predictably when multiple DP algorithms are executed on the same dataset. This enables the concept of a 'privacy budget', which makes the cumulative re-identification risk quantifiable and can be set depending on policy or preference. We note that this ability to moderate risks stemming from AI applications is particularly beneficial, as it is also mandated by recent legal frameworks such as the European AI act[26]. These properties are leading to DP's increasing adoption in industry and government applications[27,28].

We remark that for a holistic workflow, additional PETs are advisable. Cryptographic techniques such as homomorphic encryption or secure multi-party computation can allow performing computations on data while ascertaining that only authorized instances can read the private information. However, these techniques are 'binary', that is, information is perfectly private (encrypted) or non-private (decrypted). In particular, at the latest at inference time, the information must be decrypted to be useful. In contrast, DP limits the probability that the output (gradient) can be correctly assigned to the input (data), which allows useful outputs at a guaranteed (but not perfect) level of privacy. Arguably, the most famous PET is federated learning, which provides a means to preserve data governance. However, without further protective measures, in particular DP, data can be reconstructed, and thus data governance is again not maintained. An overview can be found in ref. 29.

Despite these benefits, the effective and efficient implementation of DP in large-scale AI systems also presents a series of challenges. DP has been criticized for the fact that the choice of an appropriate privacy budget is delicate. Higher budgets correspond to less privacy protection and thus an increased risk of successful attacks, while lower budgets limit the information available for training. This introduces new challenges, namely a trade-off between privacy and model performance, that is diagnostic accuracy for a given use case. Furthermore, this trade-off also depends on the specific input data and learning task, which can vary drastically between scenarios. Arguably, concerns about reduced model performance are a probable reason why, despite its benefits, DP is not yet widely implemented in medical AI. After all, finding a trade-off between diagnostic accuracy and privacy represents a complex technical and ethical dilemma. This dilemma is best understood as DP is underlain by a worst-case set of assumptions. These assumptions, also called a threat model, include an adversary who is able to deeply manipulate and interfere with the dataset, the training process, model architecture and (hyper-)parameters, and has access to all parameters of the DP algorithm (mechanism). Moreover, the canonical DP adversary is not assumed to execute a data reconstruction attack but a much simpler type of attack, namely a membership inference attack, which attempts to determine whether a specific individual's data (which is available to the adversary) was included in the training dataset or not. Since there are only two possible outcomes of such an attack (member/non-member), membership inference must only reveal a single bit of information compared with a data reconstruction attack, which must successfully reveal a much larger record (for example, an image). Although worst-case assumptions are prudent for the theoretical modelling of adversaries, the DP threat model is unlikely to ever be encountered in practice. Moreover, the aforementioned membership inference attack in which the adversary has access to a target record and tries to determine whether it was used for training a specific model is arguably of very low practical relevance. Instead, data reconstruction attacks are probably perceived as a substantially more relevant privacy threat by patients. Moreover, realistic adversaries in the medical setting (where data is strongly guarded) can probably be assumed to not have access to the training data (as they would have little incentive to attack a model otherwise).

Table 1 | Overview of the capabilities of an adversary in the threat models analysed in this study

	Worst case	Relaxed	Realistic
Model architecture and weight	Yes	Yes	Yes
Hyper-parameter	Yes	Yes	Yes
Dataset access	Yes	Partially	No
Perfect reconstruction algorithm	Not applicable	Yes	No
Risk analysis	Theoretical	Theoretical	Empirical

In this Article, we investigate whether the aforementioned typical DP threat model might be too pessimistic for practical use cases and thus impose unnecessary privacy/performance trade-offs. To investigate this hypothesis, we study the privacy/performance characteristics of AI models trained on large-scale medical imaging datasets under more realistic threat models that still allow for strong privacy protection but represent a 'step down' from the worst-case assumptions of DP. Our main finding is that, even in complex medical imaging tasks, it is possible to train AI models with excellent diagnostic performance while still defending against data reconstruction attacks and thus a likely patient re-identification. We achieve this by training models under privacy budgets that would be considered too large to offer any protection against the threats considered under the worst-case DP threat model. This supports a recommendation for training AI models with DP protection by default. Therefore, although more restrictive privacy budgets than the ones used in our study remain relevant for use cases in which protection against membership inference is explicitly required, there exists an additional option: when high model performance is required but cannot be achieved without relinquishing membership inference protection, our findings offer a compromise whereby an important and relevant class of attacks can be defended against while fulfilling the requirement for high diagnostic accuracy.

As stated above, DP allows for a quantifiable reduction in the risk of privacy attacks associated with the training of AI models. In this work, we differentiate between three threat models, which we term worst case, relaxed, and realistic. DP, reconstruction risks and all threat models are described in detail in Supplementary Material A. An overview can be found in Table 1.

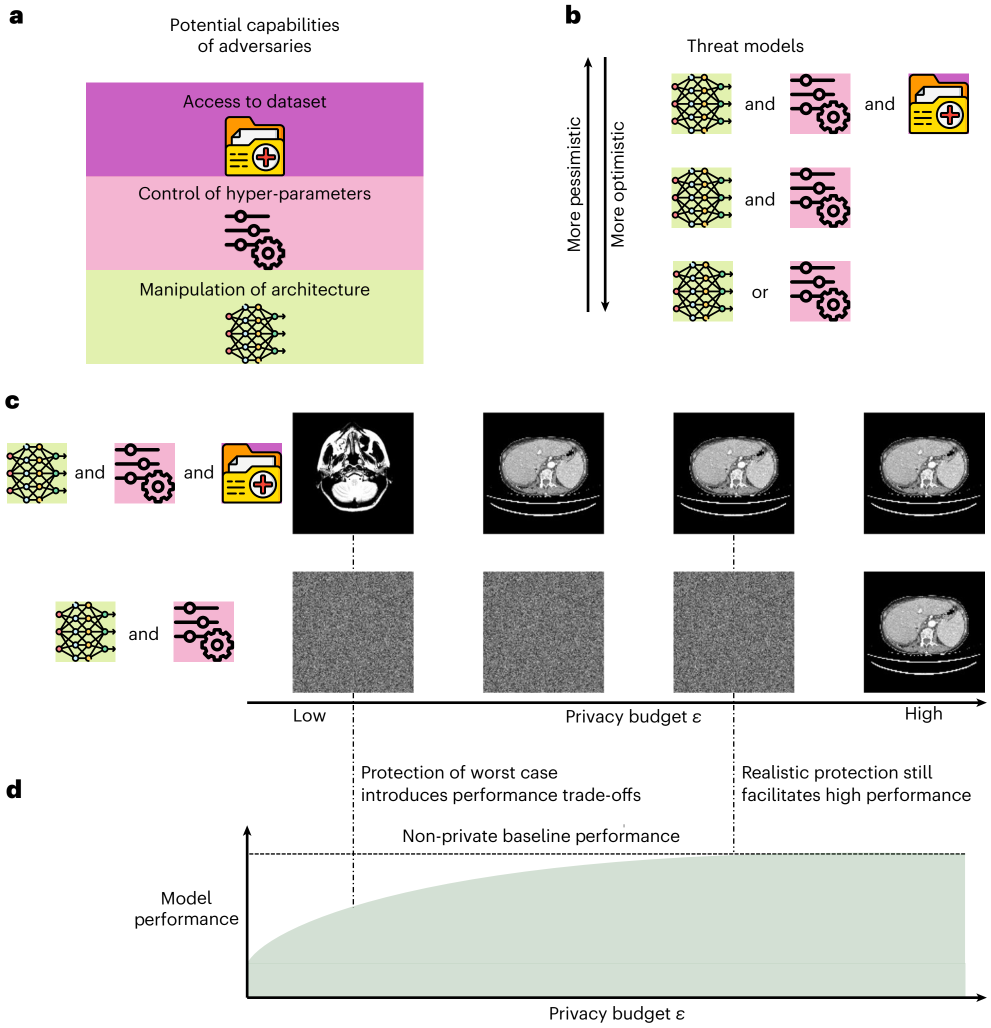


Fig. 1 | Comparison of a worst-case and a realistic threat model. a, Adversaries can have various capabilities depending on the setting. **b**, The combination of the adversary's capabilities defines the threat model. In a worst-case analysis, they have all capabilities. However, access to the database is a pessimistic, practically irrelevant scenario. **c**, The necessary privacy protection depends on the threat model. In a worst-case threat model, the adversary only needs to match the model and gradient to an image in the database. In a practically more relevant scenario, the image must be reconstructed from the model and gradient. Here, much less privacy protection is necessary. **d**, The more stringent the privacy protection is chosen, the higher the impacts on the model performance are. Thus, if a realistic threat model is considered appropriate, models can perform better.

The key contribution of our work is to investigate the realistic risks posed by a type of adversary who is still very powerful but can be reasonably assumed to exist in real-world medical AI model training use cases. An overview is displayed in Fig. 1. In the next section, we will show that perfectly defending against such adversaries is possible while maintaining a diagnostic model performance competitive with that of a model trained without any privacy protection.

Results

Set-up

Our evaluation focuses on how various privacy risks on multiple real-world characteristic datasets (compare Table 2) correlate with the algorithm's performance. We provide details on the datasets and our rationale for choosing these in Supplementary Material B1 and on the evaluation metrics in B2. First, we show the correlation of the AI performance on our datasets with privacy budgets. Second, we illustrate the implications of a certain privacy budget in a risk profile, summarizing the reconstruction risk under different threat models. We recall that a threat model corresponds to the set of assumptions over the attacker, where we give the theoretical bounds for a worst-case and a slightly relaxed adversary. Both are more pessimistic than any real-world scenario. Thus, we add a third threat model representing the worst 'realistic' case.

Table 2 | Overview of characteristics of our datasets

Dataset	Task	Small	Imbalanced	Multi-modal
RadImageNet	Classification		✓	✓
HAM10000	Classification	✓	✓	
MSD Liver	Segmentation	✓	✓	

In Table 3, we list the best possible AI model performance and corresponding reconstruction risk for all datasets and privacy budgets. The risk is three-tiered: (1) The upper bound of a worst-case adversary. This is the maximum risk under this setting and cannot be increased by post-processing or side information. (2) The upper bound of a minimally relaxed adversary as introduced in ref. 24. (3) The reconstruction success of the real-world adversary. We argue that—for practical use cases—protection against such a real-world attacker suffices. By listing

Table 3 | Comparison of performance to privacy risk over multiple datasets and privacy budgets

Privacy budget	Noise	Test MCC	Reconstruction risk		
ε at $\delta=8.0\times10^{-7}$	σ	Mean±s.d.	Worst case	Relaxed	Realistic
RadImageNet					
1	0.67	64.95±0.13%	0.00%	0.00%	0%
8	0.34	68.75±0.13%	0.04%	0.01%	0%
32	0.267	69.99±0.25%	13.18%	3.96%	0%
10^{12}	0.054	70.83±0.19%	100%	100%	0%
Non-private	0	71.83±1.86%	100%	100%	100%
HAM10000					
1	0.92	15.60±4.13%	0.03%	0.01%	0%
8	0.47	37.48±3.45%	1.22%	0.04%	0%
20	0.40	42.83±2.37%	22.30%	0.78%	0%
10^{9}	0.02	51.98±2.52%	100%	100%	0%
Non-private	0	51.66±1.38%	100%	100%	100%

		MSD Liver				
		Dice score liver	Dice score tumour	Reconstruction risk		
		Mean±s.d.	Mean±s.d.	Worst case	Relaxed	Realistic
1	9.97	42.84±1.83%	0.96±0.37%	1.66%	0.97%	0%
8	1.66	74.71±3.14%	3.01±0.96%	17.96%	3.68%	0%
20	0.96	79.06±2.17%	5.55±0.72%	74.24%	27.37%	0%
10^{9}	0.0054	91.20±0.23%	29.73±2.89%	100%	100%	0%
Non-private	0	91.58±0.41%	28.38±2.29%	100%	100%	100%

Test MCC denotes Matthew's correlation coefficient on the test dataset. For all performance metrics, we give the mean±s.d. over five runs with different random seeds. Reconstruction risk denotes the upper bounds for the risk of a successful reconstruction attack of a worst-case and minimally relaxed adversary, as well as the empirical success of one of the strongest 'realistic' attacks. An image is considered successfully reconstructed if the SSIM to any reconstruction is higher than 80%. Note that the noise multiplier σ is given for the empirical attack scenario where an adversary manipulated hyper-parameters in their favour. Noise multipliers for performance analysis are generally higher.

all three, we provide an overview of how the risk varies by changing assumptions about the adversary.

Performance trade-offs under varying privacy levels

Impacts on performance is substantial for small datasets. At first, we analyse the impact of a very restrictive (small) privacy budget of $\varepsilon = 1$ on the predictive AI performance on our datasets (Table 3). Across the board, we see that at these budgets, the impacts on the model performance are strong. Concretely, we find that on RadImageNet, a standard non-private AI model reaches 71.83% on average, while trained at such restrictive privacy guarantee we find an average Matthews' correlation coefficient (MCC) of 64.95%, which is still 90% of the non-private MCC score. The gap becomes much larger on the HAM10000 dataset, where the model performance, when trained with a very low privacy budget of $\varepsilon = 1$ is closely above the chance level at an MCC of 15.60%. Similarly, on the Medical Segmentation Decathlon (MSD) Liver dataset at restrictive privacy budgets, the average Dice score for the liver drops to 42.84% (non-private: 91.58%) and completely fails for the tumour with a Dice of 0.96%. This exemplifies the challenges of furnishing strong privacy protection when training AI models on small or difficult datasets.

Prediction quality under medium budgets depends on dataset. Next, we consider medium privacy budgets ranging from $\varepsilon = 8$ to $\varepsilon = 32$, which are typical choices in literature[30,31]. As ε is an exponential parameter (e^{ε}), larger values correspond to exponentially decreased privacy guarantees. For this reason, some argue that the guarantees provided by such medium budgets are meaningless[22,32].

At these privacy budgets, although the performance substantially increases compared with the extremely restrictive privacy budget, the private AI models never exactly match the non-private performance. On RadImageNet, the achieved result closely approaches the non-private baseline: at a privacy budget of $\varepsilon = 32$, the MCC is 69.99% versus 71.83% in the non-private case. Also, for HAM10000, performance is strongly improved at 42.83% MCC, yet still decreased by 9% compared with the non-private result. Lastly, in MSD Liver, the liver as a larger organ can now be learned up to a reasonable Dice score of 79.06% at $\varepsilon = 20$. However, it remains far from the non-private performance. The prediction quality of the tumour, which is a much smaller and more complex structure, is especially concerning. This leads to a poor segmentation quality and only achieves an average Dice score of 5.55%, which is unsuitable for real-world applications. Again, we note that performance trade-offs especially impact smaller and imbalanced datasets.

Performance trade-offs vanish under large privacy budgets. For very large privacy budgets, we observe that the gap between private and non-private performance disappears. We recall that HAM10000 and MSD Liver as small datasets are extremely challenging under restrictive DP conditions. When increasing the privacy budget to $\varepsilon = 10^{9}$, no statistically significant difference to the non-private model can be detected (*P* values: HAM10000: 0.36; and MSD Liver dataset liver: 0.10 and tumour: 0.29, Student's *t*-test). Only on RadImageNet, although the non-private model is still statistically significantly superior (*P* value: 0.001), the private model at an $\varepsilon = 10^{12}$ achieves 99% of the non-private baseline performance.

It is unsurprising that increasing the privacy budget mitigates the negative implications on the model performance. Hence, the question that must be asked is what level of privacy is necessary for a specific setting. This cannot be answered generally and must be carefully

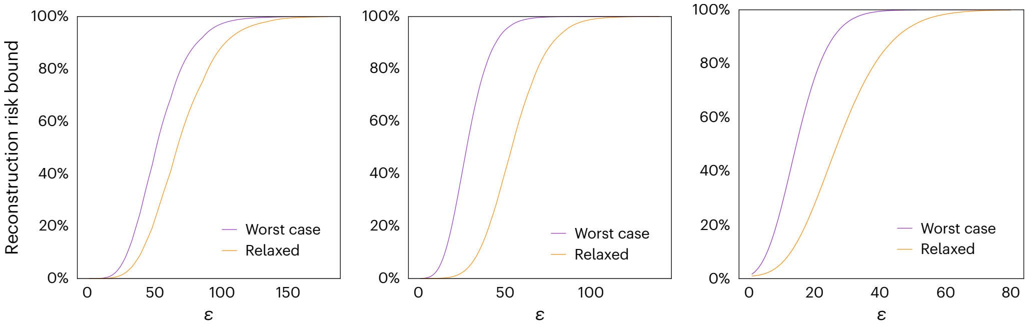


Fig. 2 | Theoretical reconstruction bounds for a worst-case and slightly relaxed adversary. From left to right: RadImageNet, HAM10000 and MSD Liver. We see that the mathematical upper bound for a reconstruction risk of a minimally relaxed threat model (orange) is already substantially lower compared with a worst-case setting (purple).

considered for each use case. Important for these considerations is which risks are associated with a certain privacy budget, which we analyse next.

Worst-case bounds require small privacy budgets

Although too pessimistic for most use cases, worst-case analyses have the advantage of a formal guarantee, that is, an absolute upper bound on the risk in this scenario. When analysing the theoretical worst-case (highest) success of reconstruction attackers, we find that for the large RadImageNet dataset for budgets $\varepsilon \leq 8$, the risk is <0.05%. However, already at $\varepsilon = 32$, the theoretical probability of the original data being reconstructed is 15%. Here, the smaller datasets are again at higher risk. While at $\varepsilon = 1$ the risk remains low, it strongly increases at $\varepsilon = 8$ for HAM10000 (0.03% to 1.22%) and MSD Liver (1.66% to 17.96%). At $\varepsilon = 20$ theoretically, up to 74.24% of all data samples of the MSD Liver dataset can be reconstructed.

However, even minimally relaxing the threat model assumptions decreases the risk associated with these privacy budgets drastically. We recall that under this relaxed threat model, the only change compared with the worst case is that the attacker does not know the sample that is reconstructed beforehand. Yet, for theoretical analysis, there is still the assumption that the reconstruction algorithm is either perfect or fails and the risk which is then calculated is the maximum rate where the attacker correctly decides if the reconstruction they obtained was indeed the dataset sample in question. This threat model is still too pessimistic for any real-world use case and the analysis is mostly for theoretical purposes. Still, such a minimal relaxation already gives a much more favourable risk profile, especially for medium privacy budgets. Exemplarily, the risk associated with $\varepsilon = 20$ diminishes from over 20% to less than 1% for the HAM10000 dataset. Similarly, the risk for the MSD dataset at $\varepsilon = 8$ decreases from 18% to 4%. A visualization of the risk difference in worst-case and relaxed threat models can be found in Fig. 2.

Empirical protection even at large privacy budgets

The previously discussed theoretical analyses show rapidly growing risks associated with small and medium privacy budgets. However, as discussed before, we argue that these analyses are too strict for any ‘realistic’ use case. Hence, we ask what the worst case of any practical scenario is and determine it to be a federated learning set-up, where a central server coordinates the learning on the data of distributed clients, which follow each training command sent by the server. This implies that the server can freely choose any network architecture and hyper-parameters. Note that any client who performs a simple check would notice such a malicious server. For such cases, attacks have been shown in literature, which analytically can recover the model input perfectly[8,9]. Moreover, it has been shown that these attacks can be transferred to corrupted pre-trained models[17]. We employ these attacks as empirical risk assessments. To measure the reconstruction success, we use the structural similarity (SSIM) score, which is a standard metric for image similarity[33].

In contrast to the aforementioned theoretical risk bounds, we find that, for practical attacks, even privacy budgets considered meaningless ($\varepsilon > 10^9$) can provide effective protection against reconstruction. In Fig. 3, left, we plot how many dataset images are below an increasing SSIM error per privacy budget. It can be thought of as the cumulative distribution function of reconstruction errors. We observe that, for all datasets without the addition of DP constraints, nearly all images can be reconstructed perfectly. As soon as some privacy guarantee is introduced, even very generous budgets at an $\varepsilon \approx 10^9$ provide empirical protection against the reconstruction of data samples. Furthermore, confirming previous works[8,34], our threat model is still extremely powerful. A server without the control of hyper-parameters but still over the model architecture already imposes a substantially lower reconstruction risk. If the server does not set the batch size to one but is set to the real training batch size, for example, on the RadImagenet dataset even in the non-private case we could only reconstruct less than 5% of all images at a batch size of 3,328. We note that such large privacy budgets, which are near-universally shunned as being meaningless, still offer empirical protection. In other words, even a ‘pinch of privacy’ has drastic effects in practical scenarios. Complemented by the finding that performance trade-offs nearly disappear in these settings, this signifies a potential compromise between protection and usability.

Discussion

In this study, we explore the relationship between privacy risks and AI performance in sensitive applications such as medical imaging. Currently, practitioners are confronted with trade-offs between AI performance, privacy protection and computational efficiency, where no solution has so far been able to accomplish all of these goals. Previous work showed that DP training profits much more than standard AI training from a higher number of training steps[30]. By increasing privacy budgets, practitioners can reach similar trade-offs with fewer training steps, which further allows a broader use for practitioners without substantial compute resources. Moreover, prior work also showed that pre-training on a 4 billion image dataset allows models to transfer to private datasets[35]. However, in practice this is typically infeasible due to limited access to such large datasets or the computational resources to train such a model. Furthermore, such data scales only exist for natural two-dimensional images but not yet for three-dimensional images, which are typical in medical imaging. Therefore, often the choice remains for practitioners to prioritise privacy and sacrifice performance or to put sensitive data at risk of being leaked. Currently, there is no clear method to balance these two objectives, leaving practitioners without guidance. To make informed decisions on these trade-offs, broad discourse involving ethicists, lawmakers

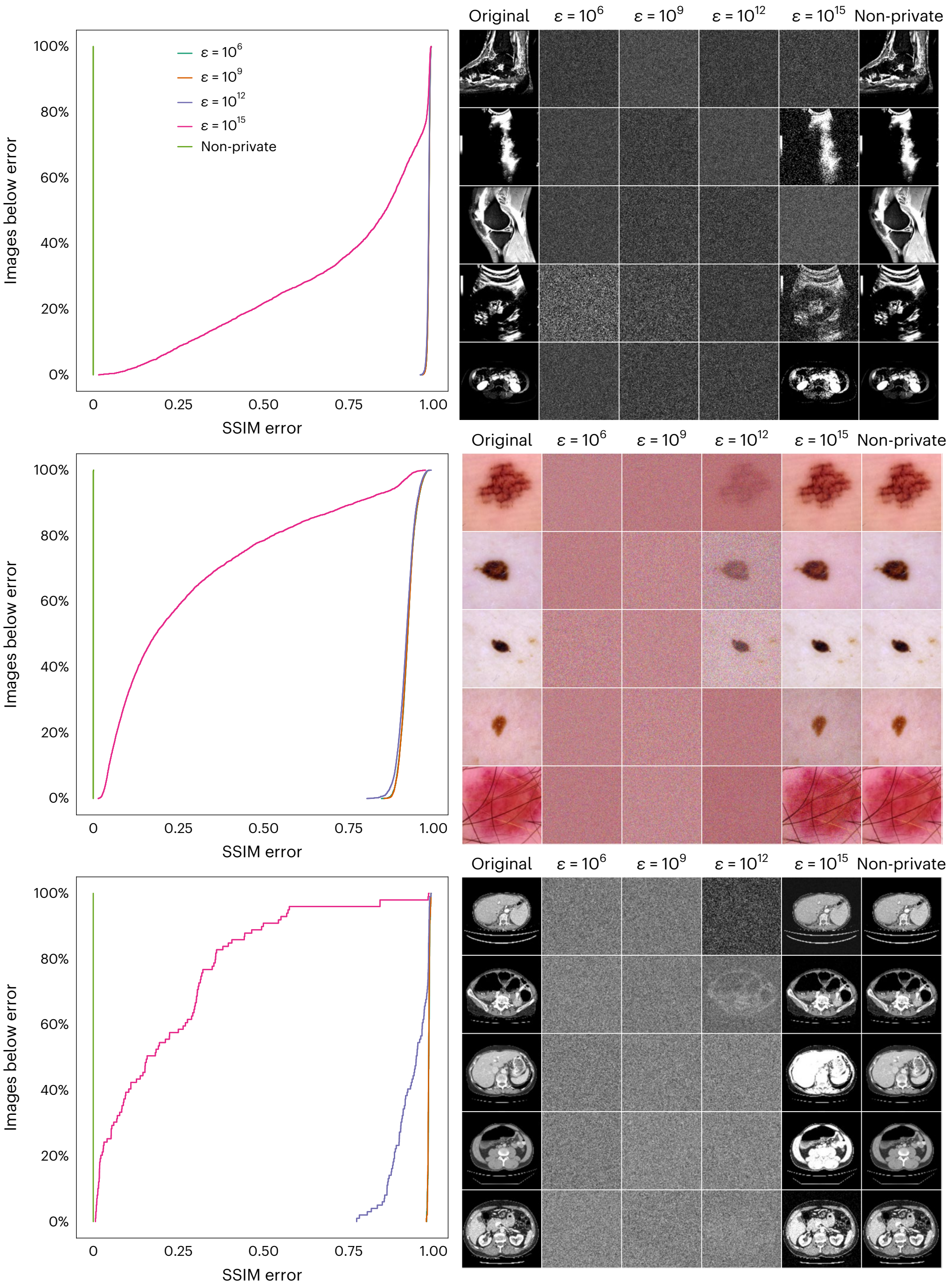


Fig. 3 | Reconstruction threat analysis for three datasets. Each row shows one dataset. From top to bottom: RadImageNet, HAM10000 and MSD Liver. Left: the cumulative number of images that have, in an empirical reconstruction, a SSIM difference lower than the value on the x axis. Note that it is the SSIM reconstruction error and thus perfect at 0 and worst at 1. Exemplarily, we see that on the MSD dataset at a reconstruction error of 10% all non-private (green) images, 39% at $\varepsilon = 10^{15}$ (pink) and none at more restrictive privacy guarantees can be reconstructed Right: the top five images with the best reconstruction score and their corresponding best reconstruction at various privacy budgets.

and the general population is crucial. A prerequisite of this dialogue is understanding the risks associated with specific privacy budgets and the potential trade-offs in AI performance. Our study across three representative medical imaging datasets lays the foundation for this conversation. We find that real-world data reconstruction risks can be averted without performance trade-offs. In fact, privacy–performance trade-offs have so far always been based on worst-case assumptions, which do not overlap with realistic training settings. We postulate that it is more critical to prevent data reconstruction in real-world settings, and show that for workflow de-risking, large privacy budgets suffice. Even more, we find that the trade-off between privacy risks and model performance vanishes when using such large but protective privacy budgets.

It is known from previous works[23,36–38] that PETs formally protect AI models in sensitive contexts from reconstruction attacks. While we note that our results are empirical, it is apparent that DP training with minimal guarantees still provides better protection than non-private training. Considering this finding, it seems negligent to train AI models without any form of formal privacy guarantee. We note that the threat model we consider is probably still stronger than attackers encountered in practical attack scenarios. In a slightly different threat model, where an adversary only has black-box access to the final trained weights of a model but has an image prior containing the true target point, ref. 23 found that large privacy budgets in the order of the dimensionality of the data suffice to prevent reconstruction attacks. Similarly, ref. 32 found that against reconstruction attacks, noise multipliers which otherwise would be seen as vacuous, suffice. Furthermore, ref. 39 studied the reconstruction of discrete data and found that privacy budgets can be much larger than previously thought to effectively defend against reconstruction attacks. However, for our threat model, we find even much larger privacy budgets than the aforementioned to suffice and, without a theoretical lower bound, the possibility exists that future attacks could achieve success closer to the upper bound. Owing to this, we explicitly warn readers to take our results as a carte blanche to use arbitrarily high privacy budgets. The truth lies in the middle: if the alternative is to not use any privacy at all, rather use DP with a very high budget.

We remark that the effectiveness of the DP protection against attacks at a fixed clipping norm, batch size, training duration and training set size depends only on the noise multiplier. This is a consequence of how DP budgets are accounted. For example, in the Rényi-DP (RDP) accountant[40] used in our work, one step is $(\alpha, q^2 \frac{2\alpha C^2}{\sigma^2})$-RDP for appropriate values of the parameters α the order of the Rényi divergence, q, the subsampling rate (that is, batch size divided by training set size), C, the clip gradient norm and σ, the noise multiplier. However, our empirical results suggest that for all other factors being constant, even small noise multipliers, which imply very large privacy budgets, are sufficient to protect against reconstruction attacks and facilitate high-performing AI models. We also observed that the AI performance loss introduced by DP tends to be smaller on larger datasets due to less injected noise per sample and more information to achieve a certain privacy budget at consistent hyper-parameters. Yet, many medical datasets are inherently small. This can have negative consequences for the applicability of such networks in clinical practice. For models to be effectively trained on such challenging datasets, when pre-training is not possible for reasons of data availability or computational resources, our techniques reach a limit indicating a potential need to either accept elevated privacy risks or obtain access to more data. The solution to both problems might go hand in hand with more robust mathematical guarantees safeguarding data privacy. In such a scenario, we anticipate that patients may be more inclined to share their data, thereby allowing large-scale medical AI training. In such a scenario, the privacy–performance trade-offs presented might even be more favourable than our findings indicate. This would be complemented by a workflow where multiple PETs are employed to enable various aspects to privacy. For example, a system using federated learning to assert the data governance remains at the original hospital, secure aggregation to conceal contributions from different sites and DP to limit the private information of single patients demonstrated in previous works[36] would provide a holistic workflow.

We note that our choice of datasets and architectures is motivated by medical imaging settings. In those settings, typically computational resources are limited and data are scarce. In fact, we are convinced that the widespread use of such methods will only ensue once they can be used by the majority of practitioners who typically lack access to large computing clusters. Hence, we carefully designed our study to cover typical and representative medical problems to provide a holistic analysis with trade-offs in computational resources. Under these considerations, we limited ourselves to a few model architectures that are known to be trained efficiently (ResNet, DenseNet and U-Net) and datasets that represent a broad range of typical problems.

An additional technical limitation stems from the fact that the authors of the RadImagenet dataset[41] mention that some patients contributed multiple images. However, we have no information about image-to-patient correspondence. As we calculate the privacy guarantees over the dataset per image, the per-patient privacy guarantee depends on the number of images one patient contributed and might be lower.

In conclusion, we show that even the use of nominally loose privacy guarantees still provides substantially better protection than standard AI training, while achieving comparable performance. This can facilitate a compromise between provable risk management and performance trade-offs, which previously prevented the breakthrough of DP. Further research should be directed towards analysing various threat models beyond the worst case. Only by illuminating the risks of multiple scenarios, the basis for a broad discussion among ethicists, policymakers, patients and other stakeholders is provided regarding how to trade-off privacy and performance as fundamental goals of AI in sensitive applications.

Methods

In this section, we report all the details necessary for our experiments on training models in a differentially private way on our datasets as well as the procedures to analyse risk profiles. Furthermore, we describe the rationale for several choices in our study design and describe hyper-parameters necessary for reproducibility.

Data

In Supplementary Material A, we describe characteristics of typical medical datasets. We note, that these characteristics partially amplify the negative performance impact by the constraints introduced by DP. Broadly speaking, at a constant clipping norm the amount of introduced noise during the DP process determines the negative impact on the AI performance. At any privacy budget, the injected noise increases if more training steps are performed or if a higher sampling rate, that is, the ratio between batch size and dataset size, is used. However, the batch size is typically irrespective of the dataset size, which implies that smaller datasets typically have higher sampling rates. Furthermore, they often require more training epochs, that is, the amount of times the entire dataset was (on average) presented to the network. As a consequence, the amount of noise that is injected when training on small datasets compared with larger ones is increased and higher performance penalties are expected. Furthermore, DP bounds the magnitude any single sample on the training. This is important for training with imbalanced datasets with underrepresented classes, which often suffer an additional performance loss[42].

For detailed descriptions of the datasets we refer to the original publications[41,43–45]. In the following, we describe modifications we performed and the effects on the data distribution.

For the HAM10000 dataset[43], we merged classes into whether there is indication for immediate treatment, which is still a medically important distinction. By this we convert the multi-class classification problem into a highly imbalanced binary classification problem. We categorized them here as follows:

Treatment indication	
Immediate	**Not immediate**
Actinic keratoses and intra-epithelial carcinomas	Melanocytic nevi
Basal cell carcinomas	Benign keratinocytic lesions
Melanomas	Dermatofibromas
	Vascular lesions

In total, this dataset has 10,015 images, of which 1,954 are labelled for immediate treatment and 8,061 are not.

Model training

All of our experiments were performed using an NAdam optimizer, which is extremely robust to learning rate changes allowing us to keep a consistent learning rate of $2e^{-3}$. Input data were always normalized with the mean and standard deviation of all images in the training set. For each dataset, we perform a hyper-parameter search, where we evaluate for one privacy level ($\varepsilon = 8$) and the non-private training the optimal setting for architecture, batch size, loss weighting and augmentation. In the non-private case, we perform an early stopping strategy to determine the number of epochs. In the private case, this is not possible as the number of epochs directly influences the amount of added noise. However, previous works showed that longer training almost always yields better results[30]. Yet, to limit training time, we also search for the point of saturation. Also for reasons of computational complexity, we assume that the optimal settings for these parameters transfer to all other privacy regimes. Furthermore, we limit the choice of architectures to a ResNet-9 with ScaleNorm and a WideResNet40-4, which have in previous literature been proven to be especially suited for differentially private training[30,46]. In the segmentation case, we limit ourselves to a standard U-Net[47,48], where we optimize the number of channels on the bottleneck. We then evaluate for each privacy setting separately the optimal clipping norm. Again for reasons of computational complexity, we evaluate this after one epoch and assume it transfers to longer trainings. Finally, we train for each setting five models with different random seeds and report the mean and standard deviation of the respective performance metric.

All our models are trained from 'scratch', that is, we have not pre-trained on any other dataset. This is because there is no 'good choice' of a dataset for pre-training. ImageNet, which for most computer vision tasks is the standard, is not very effective for medical imaging tasks[41]. Large public databases for pre-training are scarce and only available for a few tasks. Furthermore, pre-training on non-public medical databases is unacceptable, as it risks leaking the information from the pre-training data, which could be just as private[49,50].

We used the Opacus[51] library for accounting the privacy loss. In particular, we used an RDP accountant, as it provides numerically the most stable implementation. We used an extension of the objax library[52] as implementation for the DP-Stochastic Gradient Descent algorithm.

We open source the program code used for this paper at https://github.com/a1302z/RePrAAIMI.

RadImagenet. As described in the 'Model training' section, we analysed the architecture, number of epochs, batch size, loss and multiplicity for the non-private and one private setting ($\varepsilon = 8$). For the non-private case, we found a WideResNet40-4 using an unweighted loss function, a batch size of 16 and random vertical (probability of augmentation (P_{aug}) = 0.2) and horizontal flips ($P_{\text{aug}} = 0.1$) as augmentation to yield the best results. To determine the number of epochs, we used an early stopping strategy with a patience of five epochs and 0.1% improvement threshold. For the private case, a ResNet-9 trained for 50 epochs, using an unweighted loss function, using an augmentation multiplicity of four again with random vertical ($P_{\text{aug}} = 0.2$) and horizontal ($P_{\text{aug}} = 0.2$) flips with a batch size of 3,328 yielded best results. The clipping norm was tuned for each budget separately and was set as follows:

ε	1	8	32	1e^{12}
Clip norm	6.46	5.66	5	3.75

HAM10000. For the modified HAM10000 dataset, we found the ResNet-9 to perform best in private and non-private settings. In the non-private case, we trained with a weighted loss function at a batch size of 32 using random vertical flips ($P_{\text{aug}} = 0.5$) as augmentation. We trained using an early stopping strategy using a patience of 50 epochs at a minimal improvement threshold of 0.1%. For the private case, we used an unweighted loss function at a batch size of 2,048 and trained for 100 epochs. We used the same augmentations as in the non-private case for a privacy level of $\varepsilon = 10^9$, for all others, we did not use augmentations. Clipping norms are as follows:

ε	1	8	20	1e^{9}
Clip norm	18	8.5	9.5	9

MSD Liver. For the MSD Liver dataset, we found for both private and non-private cases a U-Net with 16 channels and no augmentations to perform best. In the non-private case we used a weighted loss function (background: 0.1; liver: 0.4; tumour: 0.5) and trained at a batch size of two. Again, we employed an early stopping strategy with a patience of 50 epochs and a minimal improvement threshold of 0.1%. In the private case, we trained at a batch size of one for 500 epochs. For privacy budgets $\varepsilon \leq 20$ we used an unweighted loss function, for higher privacy budgets we used the same weighting as in the non-private case.

ε	1	8	20	1e^{9}
Clip norm	0.0004	0.046	0.0015	0.33

Reconstruction risk analysis

In our empirical reconstruction attacks, there is no clear way to evaluate whether a specific sample was reconstructed. For each input batch consisting of N samples, we receive M reconstructions. We evaluate this by calculating the pairwise distance between all data samples and reconstructions and assigning each input the reconstruction with the lowest distance. However, this approach loses meaning in the case of images, which have no structure but are entirely dark. This is the case for the RadImagenet dataset, where we put a constraint that only data samples are considered that contain more than 10% non-zero pixels.

We evaluate the practical reconstruction success by using a principle demonstrated in previous literature[8,9] adapted to our use case. The network architecture is slightly modified by prepending two linear layers in front of the actual network architecture. The first takes all input image pixels as input and projects them to an intermediate representation of N bins. In our experiments, we set $N = 10$. This intermediate representation is afterwards projected again to the number of all pixels and re-sized to the original image shape. To each of the outputs, the mean of the intermediate representations is added. Afterwards, it can be processed as usual by the remaining neural network. As our adversary is assumed to have control over all hyper-parameters, they can set the batch size to one and by that enforce that no reconstruction of two images overlap. If now a gradient is calculated over the network, which is non-zero for the weights W_i and biases b of the first linear layer, the input x can be analytically recovered by $x = \nabla_{W_i}\mathcal{L} \oslash \frac{\partial \mathcal{L}}{\partial b}$, where $\oslash$ is the element-wise division. We note that, for this attack, it is irrelevant what network architecture comes after this imprint block. We used implementations provided by ref. 53.

The reconstruction error, which we use as basis for the risk analysis in this paper, is the minimum reconstruction error between a data sample to any reconstruction that was derived from a gradient containing the data sample.

Choice of privacy budgets

For our experiments on the utility trade-off, we chose several privacy budgets. We note that this choice was arbitrary. For all experiments, we used a $\delta = 8 \times 10^{-7}$. For all settings, we evaluated $\varepsilon = 1$ and $\varepsilon = 8$, which are standard values in the literature[30,31,46]. Furthermore, we calculate the theoretical reconstruction bound of the worst case and relaxed threat models. As the already included privacy budgets at $\varepsilon = 1$ and $\varepsilon = 8$ already showcase very low reconstruction bounds, we add one more privacy level for all datasets, where a large amount of samples is already at risk of being reconstructed. In addition, we report a privacy budget $\varepsilon = 10^{3N}, N \in \mathbb{N}$, where the characteristic reconstruction robustness curve is still similar to random noise.

Environmental impact

Lastly, we would like to give a rough estimate of the climate impact of this study. We assume the average German power mix that as of 2021 according to the German Federal Environment Agency corresponds to 475 g CO_2e kWh^{-1} (ref. 54) Only the final RadImagenet trainings (no hyper-parameter optimization) ran on eight NVIDIA A40s, where we assume a power consumption of 250 W on average, each for almost 4 days, five privacy levels and five repetitions. Hence, this amounts to around 960 kWh and thus more than 450 kg of CO_2e. This almost equals a return flight from Munich to London. Hence, we tried to limit our hyper-parameter searches to the necessary. In total, we assume that this study produced at least 2 tons of CO_2e.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets used in this study are published and publicly available. Access to RadImageNet[41] must be requested at https://www.radimagenet.com/. The HAM10000 dataset[43] is available at https://doi.org/10.7910/DVN/DBW86T. The MSD Liver dataset[44,45] is available at http://medicaldecathlon.com/ and https://doi.org/10.1038/s41467-022-30695-9.

Code availability

Our program code is available at https://github.com/a1302z/RePrAAIMI and permanently archived under https://doi.org/10.5281/zenodo.11184978 ref. 55. Furthermore, we created a modified version of[53], which is available at https://github.com/a1302z/objaxbreaching and https://doi.org/10.5281/zenodo.11184998 ref. 56.

References

1. Lång, K. et al. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* **24**, 936–944 (2023).
2. Wang, G. et al. Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. *Nat. Med.* **29**, 2007–2018 (2023).
3. Al-Zaiti, S. S. et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. *Nat. Med.* **29**, 1804–1813 (2023).
4. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
5. Yao Jiang, L. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
6. Geiping, J., Bauermeister, H., Dröge, H. & Moeller, M. Inverting gradients—how easy is it to break privacy in federated learning? *Adv. Neural Inf. Process. Sys.* **33**, 16937–16947 (2020).
7. Yin, H. et al. See through gradients: image batch recovery via gradinversion. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16337–16346 (2021).
8. Fowl, L., Geiping, J., Czaja, W., Goldblum, M. & Goldstein, T. Robbing the fed: directly obtaining private data in federated learning with modified models. In *Tenth International Conference on Learning Representations* (2022).
9. Boenisch, F. et al. When the curious abandon honesty: federated learning is not private. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)* 175–199 (IEEE, 2023).
10. Wang, Kuan-Chieh et al. Variational model inversion attacks. *Adv. Neural Inf. Process. Syst.* **34**, 9706–9719 (2021).
11. Haim, N., Vardi, G., Yehudai, G., Shamir, O. & Irani, M. Reconstructing training data from trained neural networks. *Adv. Neural Inf. Process. Syst.* **35**, 22911–22924 (2022).
12. Carlini, N. et al. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)* 5253–5270 (2023).
13. Buzaglo, G. et al. Deconstructing data reconstruction: multiclass, weight decay and general losses. In *Thirty-Seventh Conference on Neural Information Processing Systems* (2023).
14. Hatamizadeh, A. et al. Do gradient inversion attacks make federated learning unsafe? *IEEE Trans. Med. Imaging* **42**, 2044–2056 (2023).
15. Chen, H., Zhu, T., Zhang, T., Zhou, W. & Yu, P. S. Privacy and fairness in federated learning: on the perspective of tradeoff. *ACM Comput. Surv.* **56**, 1–37 (2023).
16. Usynin, D., Rueckert, D. & Kaissis, G. Beyond gradients: exploiting adversarial priors in model inversion attacks. *ACM Trans. Priv. Secur.* **26**, 1–30 (2023).
17. Feng, S.& Tramèr, F. Privacy backdoors: stealing data with corrupted pretrained models. In *International Conference on Machine Learning (ICML)* (2024).
18. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* 111–125 (IEEE, 2008).
19. Cohen, A. & Nissim, K. Towards formalizing the GDPR's notion of singling out. *Proc. Natl Acad. Sci. USA* **117**, 8344–8352 (2020).
20. Cohen, A. Attacks on deidentification's defenses. In *31st USENIX Security Symposium (USENIX Security 22)* 1469–1486, (2022).
21. Schwarz, C. G. et al. Identification of anonymous mri research participants with face-recognition software. *N. Engl. J. Med.* **381**, 1684–1686 (2019).
22. Dwork, C. et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407 (2014).
23. Balle, B., Cherubin, G. & Hayes, J. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)* 1138–1156 (IEEE, 2022).
24. Kaissis, G., Hayes, J., Ziller, A. & Rueckert, D. Bounding data reconstruction attacks with the hypothesis testing interpretation of differential privacy. *CoRR* abs/2307.03928 (2023).
25. Nissim, K. Privacy: from database reconstruction to legal theorems. In *Proc. 40th ACM SIGMOD–SIGACT–SIGAI Symposium on Principles of Database Systems* 33–41 (2021).
26. *Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, document 52021PC0206* (European Parliament and of the Council, 2021).
27. Foote, A. D., Machanavajjhala, A. & McKinney, K. Releasing earnings distributions using differential privacy: disclosure avoidance system for post-secondary employment outcomes (PSEO). *J. Priv. Confidential.* **9**, 2 (2019).
28. Aktay, A. et al. Google COVID-19 community mobility reports: anonymization process description (version 1.1). Preprint at https://arxiv.org/abs/2004.04145 (2020).
29. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
30. De, S., Berrada, L., Hayes, J., Smith, S. L. & Balle, B. Unlocking high-accuracy differentially private image classification through scale. Preprint at https://arxiv.org/abs/2204.13650 (2022).
31. Sander, T., Stock, P. & Sablayrolles, A. Tan without a burn: scaling laws of dp-sgd. In *International Conference on Machine Learning* 29937–29949 (PMLR, 2023).
32. Stock, P., Shilov, I., Mironov, I. & Sablayrolles, A. Defending against reconstruction attacks with Rényi differential privacy. *CoRR* abs/2202.07623 (2022).
33. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).

34. Usynin, D., Rueckert, D., Passerat-Palmbach, J. & Kaissis, G. Zen and the art of model adaptation: low-utility-cost attack mitigations in collaborative machine learning. *Proc. Priv. Enhancing Technol.* **2022**, 274–290 (2022).
35. Berrada, L. et al. Unlocking accuracy and fairness in differentially private image classification. Preprint at https://arxiv.org/abs/2308.10888 (2023).
36. Kaissis, G. et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach.Intell.* **3**, 473–484 (2021).
37. Ziegler, J., Pfitzner, B., Schulz, H., Saalbach, A. & Arnrich, B. Defending against reconstruction attacks through differentially private federated learning for classification of heterogeneous chest x-ray data. *Sensors* **22**, 5195 (2022).
38. Hayes, J., Mahloujifar, S. & Balle, B. Bounding training data reconstruction in DP-SGD. In Proc. *37th Conference on Neural Information Processing Systems* (OpenReview.net, 2023).
39. Guo, C., Sablayrolles, A. & Sanjabi, M. Analyzing privacy leakage in machine learning via multiple hypothesis testing: a lesson from fano. In *International Conference on Machine Learning* 11998–12011 (PMLR, 2023).
40. Mironov, I., Talwar, K. & Zhang, L. Rényi differential privacy of the sampled Gaussian mechanism. Preprint at https://arxiv.org/abs/1908.10530 (2019).
41. Mei, X. et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiol. Artif. Intell.* **4.5**, e210315 (2022).
42. Bagdasaryan, E., Poursaeed, O. & Shmatikov, V. Differential privacy has disparate impact on model accuracy. *Adv. Neural Inf. Process. Syst.* **32**, (2019).
43. Tschandl, P., Rosendahl, C. & Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 1–9 (2018).
44. Simpson, A. L. et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. Preprint at https://arxiv.org/abs/1902.09063 (2019).
45. Antonelli, M. et al. The medical segmentation decathlon. *Nat. Commun.* **13**, 4128 (2022).
46. Klause, H., Ziller, A., Rueckert, D., Hammernik, K. & Kaissis, G. Differentially private training of residual networks with scale normalisation. In *Theory and Practice of Differential Privacy Workshop* (ICML, 2022).
47. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015.* Part III 18, 234–241 (Springer, 2015).
48. Çiçek, Özgün, Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D u-net: learning dense volumetric segmentation from sparse annotation. In *Proc. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016.* Part II 19, 424–432 (Springer, 2016).
49. Abascal, J., Wu, S., Oprea, A. & Ullman, J. Tmi! finetuned models spill secrets from pretraining. In *The Second Workshop on New Frontiers in Adversarial Machine Learning* (2023).
50. Tramèr, F., Kamath, G. & Carlini, N. Considerations for differentially private learning with large-scale public pretraining. Preprint at https://arxiv.org/abs/2212.06470 (2022).
51. Yousefpour, Ashkan, et al. Opacus: user-friendly differential privacy library in PyTorch. Preprint at https://arxiv.org/abs/2109.12298 (2021).
52. Objax. *Objax Developers* https://github.com/google/objax (2022).
53. Wen, Y., Geiping, J. & Fowl, L. Breaching. *GitHub* https://github.com/JonasGeiping/breaching (2023).
54. Icha, P., Lauf, T. & Kuhs, G. Entwicklung der spezifischen Treibhausgas-Emissionen des deutschen Strommix in den Jahren 1990–2021. *Umweltbundesamt Dessau-Roß*lau (2022).
55. Ziller, A., Kaissis, G. & Stieger, S. a1302z/repraaimi. *Zenodo* https://doi.org/10.5281/zenodo.11184978 (2024).
56. Ziller, A. objaxbreaching. *Zenodo* https://doi.org/10.5281/zenodo.11184998 (2024).

Acknowledgements

The authors acknowledge the following sources of funding for this work: A.Z. and G.K. are supported by the German Ministry of Education and Research (BMBF) under grant number 01ZZ2316C (PrivateAIM). T.T.M. and D.R. are supported supported by the ERC Grant Deep4MI (no. 884622). L.F.F. and R.B. are supported by the Federal Ministry of Education and Research (BMBF), Project 'NUM 2.0' (grant no. 01KX2121). R.B. is funded by the Federal Ministry of Education and Research (BMBF, grant no. 01ZZ2315B), Bavarian Cancer Research Center (BZKF, Lighthouse AI and Bioinformatics), German Cancer Consortium (DKTK, Joint Imaging Platform). G.K. and D.R. are funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts through the Munich Centre for Machine Learning. G.K. is supported by the Bavarian Collaborative Research Project PRIPREKI of the Free State of Bavaria Funding Programme 'Artificial Intelligence–Data Science' and from the German Academic Exchange Service (DAAD) under the Kondrad Zuse School of Excellence for Reliable AI (RelAI). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.
Furthermore, the authors thank J. Hayes for providing an implementation of the prior aware attack presented in ref. 38.

Author contributions

A.Z. conceptualized this study, wrote the program code, performed all experiments and prepared the paper. T.T.M. assisted in the preparation of the paper. S.S. assisted in the design of the program code. L.F.F. wrote program code for an efficient reconstruction matching and segmentation loss. J.B. helped to prepare the HAM10000 dataset for our purposes. R.B. and D.R. provided oversight. G.K. helped conceptualize this study and in the preparation of the paper, wrote code for the theoretical risk bounds and provided oversight. All authors revised the paper.

Funding

Open access funding provided by Technische Universität München.

Competing interests

The authors declare no competing interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00858-y.

Correspondence and requests for materials should be addressed to Alexander Ziller.

Peer review information *Nature Machine Intelligence* thanks Holger Roth, Yiyu Shi, Tian Xia and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature portfolio

Corresponding author(s): Alexander Ziller, NATMACHINTELL-A231110141A

Last updated by author(s): May 14, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ ☐ A description of all covariates tested

☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

No software was used for data collection.

Data analysis

Permanently archived and cited computer code available at https://doi.org/10.5281/zenodo.11184978 and https://doi.org/10.5281/zenodo.11184998.
Our code was developed in Python V3.10.
We used Objax V1.6.0 for training our AI models.
For accounting the privacy loss we used Opacus V1.4.0.
Other packages that were used in this study are pytorch V2.0 scikit-learn V1.3.0, scikit-image V0.21.0, pandas V2.0.3, OpenCV V4.7.0, breaching V0.1.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets used in this study are published and publicly available. Access to RadImageNet [28] must be requested at https://www.radimagenet.com/. The HAM10000 dataset [29] is available at https://doi.org/10.7910/DVN/DBW86T. The MSD Liver dataset [30,31] is available at http://medicaldecathlon.com/ and https://doi.org/10.1038/s41467-022-30695-9.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were predetermined by the public datasets used in this study. The datasets were chosen in order to represent typical medical AI workflows, with one multi-modal dataset, one small imbalanced dataset and a segmentation 3D dataset.
Data exclusions	None
Replication	Findings are deterministic with given data and seed for pseudo-random number generator.
Randomization	RadImagenet has a predetermined split into train, validation and test set, which ascertains that there is no data leakage with one patient in multiple cohorts. For HAM10000, we split randomly with a stratification by the class label in order to ensure that train, validation and test set approximately had the same distribution of classes. This was done using the train_test_split function of scikit-learn. For MSD 10000 we split randomly on a patient level.
Blinding	There are no groups which can be blinded. Therefore blinding is not applicable to this study.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For*

studies involving existing datasets, please describe the dataset and source.

Sampling strategy

Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.

Data collection

Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.

Timing

Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Non-participation

State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.

Randomization

If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.

Research sample

Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.

Sampling strategy

Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

Data collection

Describe the data collection procedure, including who recorded the data and how.

Timing and spatial scale

Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Reproducibility

Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.

Randomization

Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

Blinding

Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work? □ Yes □ No

Field work, collection and transport

Field conditions

Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).

Location

State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).

Access & import/export

Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).

Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	*State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

□ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
□	□	Public health
□	□	National security
□	□	Crops and/or livestock
□	□	Ecosystems
□	□	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
□	□	Demonstrate how to render a vaccine ineffective
□	□	Confer resistance to therapeutically useful antibiotics or antiviral agents
□	□	Enhance the virulence of a pathogen or render a nonpathogen virulent
□	□	Increase transmissibility of a pathogen
□	□	Alter the host range of a pathogen
□	□	Enable evasion of diagnostic/detection modalities
□	□	Enable the weaponization of a biological agent or toxin
□	□	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

□ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

□ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

□ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

□ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

□ All plots are contour plots with outliers or pseudocolor plots.

□ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*

Software | *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*

Cell population abundance | *Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*

Gating strategy | *Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

□ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type | *Indicate task or resting state; event-related or block design.*

Design specifications | *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*

Behavioral performance measures | *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s) | *Specify: functional, structural, diffusion, perfusion.*

Field strength | *Specify in Tesla*

Sequence & imaging parameters | *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition | *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI □ Used □ Not used

Preprocessing

Preprocessing software | *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization | *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template | *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal | *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring | *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings | *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested | *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: □ Whole brain □ ROI-based □ Both

Statistic type for inference
(See Eklund et al. 2016) | *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction | *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
□	□ Functional and/or effective connectivity
□	□ Graph analysis
□	□ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	*Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*
Graph analysis	*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*
Multivariate modeling and predictive analysis	*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*

4 Concluding Remarks

In this thesis, we have evaluated several aspects when combining PETs and AI in medicine. In particular, we have investigated a real-world privacy-preserving workflow, efficient implementations of DP-SGD, the implications on subgroup fairness in two real-world medical tasks and lastly, the impact of relaxed threat models on the privacy budget. With this, we hope to contribute towards a trustworthy, ethically, and legally compliant use of AI in medicine. In the following, we discuss our findings, particularly in terms of how they fit into the bigger picture of privacy-preserving AI and recent developments since the publication of the manuscripts. Furthermore, we give an outlook on potential future developments and research directions.

4.1 Discussion

End-to-end privacy preserving deep learning on multi-institutional medical imaging We demonstrated the use of a holistic AI pipeline facilitated by the use of various PETs, most notably FL, SMPC, and DP. It allows for distributed training, where the contributions of each hospital are concealed, and the privacy of all patients is preserved. Although the use of these technologies leads to reduced performance for models trained at each site, we could demonstrate that, for the use case of paediatric chest x-ray classification, when training this on data from all sites, it outperforms local models as well as expert radiologists. This is based on the common assumption that PETs enable access to more data, which can compensate for the utility penalties of these technologies. However, to the best of our knowledge, this assumption has not been scientifically evaluated. It, therefore, requires a legal and sociological investigation of whether and to what extent this assumption corresponds to the actual situation.

Medical imaging deep learning with differential privacy We developed a framework for training AI models under DP conditions, which was competitive to state-of-the-art frameworks at the time of publication. Moreover, our approach does not require a mathematical specification for any neural network layer but instead is natively compatible with all types of layers, which adhere to the assumptions imposed by DP-SGD. Both are important aspects for the practical use of DP. Since the publication, several new developments have been added to the Opacus framework [76]. Specifically, the inclusion of functorch [74] allows the just-in-time compilation of DP-SGD training with arbitrary layers, which can lead to substantial improvements in computational efficiency.

Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging As alluded to in Section 2.1, there are other important aspects for the trustworthy and ethical use of AI, especially in sensitive areas, such as medicine. One of them is that AI models adhere to a notion of fairness and, by that, do not discriminate against certain subgroups. Prior work has found evidence that DP could exacerbate the underperformance of AI models on underrepresented subgroups. In our investigation we found that not necessarily the representation but the difficulty of correctly predicting subgroups could be decisive. In particular, we found that the underperformance of AI models on subgroups which already have the lowest prediction performance in non-private trainings is further exacerbated with increasing privacy guarantees. This could be a paradigm shift when assessing the interaction between privacy and fairness. In particular, this could imply that when construing datasets for a fair and privacy-preserving AI model, not all subgroups should be equally represented, but subgroups which are harder to diagnose should be overrepresented. Furthermore, it could be necessary to have deviating privacy guarantees for different subgroups, i.e. subgroups on which AI performance is overly affected sacrifice privacy budget for models, which perform equally well as on other subgroups with higher privacy guarantees. It is open for future research to investigate the exact relationship of subgroup performance and the influence of DP training, and if this might even allow for a notion of "diagnosis difficulty".

Reconciling Privacy and Accuracy in AI for Medical Imaging Perhaps the most important unresolved drawback of using DP for the training of AI models is the induced privacy-utility trade-off. This describes the effect of stricter privacy preservation leading to stronger negative impacts on model performance. Hence, an important question is what an appropriate level of privacy is for a specific setting. This question is largely underexplored. So far, most works chose ε-values in a range between 1 and 10, with 8 being the default value. While the range is rooted in the risk against a MIA of a theoretical worst-case adversary, the exact value is typically not further substantiated. In this work, we addressed the question of what level of privacy is necessary to protect against "realistic" scenarios. Empirically, we found that even very large –so far considered meaningless– privacy budgets can suffice. However, in opposition to worst-case MIA risks, this is not based on irrefutable theoretical results, and it is conceivable that there are factors which are not considered in this study that could lead to higher reconstruction risks. Hence, this calls for a complementary theoretical study of the considered setting. At the time of this thesis, this complementary study is available as a preprint [102] and remains to be published in a peer-reviewed venue.

4.2 Outlook

We have so far discussed prospects and challenges which come along with the use of PETs. In this section, we give an informal, speculative outlook on how relevant open questions concerning privacy-preserving workflows for the training of medical AI could evolve.

Will PETs Become The Standard for Medical AI? Looking at the promises that the use of PETs brings, one may wonder why these technologies are not yet established standards in such a sensitive area as medical AI. There are several reasons for that: Perhaps the most trivial one is that there is a certain effort required by practitioners to use such technologies. The principle of least effort states that animals, people and well-designed machines, and therefore presumably also programmers, choose the path of least resistance [103]. This can be overcome by two options: Decreasing the resistance, i.e. simplifying the use of PETs, or blocking other paths, e.g., by legislating the use of PETs. Hence, we promote the incorporation of these technologies into existing widely used frameworks for AI training such as PyTorch [74], which would arguably set down the hurdles for the use of PETs drastically. Moreover, if this is complemented by advantages for the practitioners, such as *easier access* or *more data*, PETs could soon be standard in many pipelines. Yet, as we have outlined in Section 2.2.4, especially DP requires a basic understanding of the technology and is incompatible with certain standard AI workflows. A second key reason why PETs are not yet widely used, which not even imposing a legal requirement resolves, is most likely the privacy-utility trade-off. While we could show that adapting the privacy budget to the actual threat model can mitigate this trade-off, there will be situations in which a restrictive privacy guarantee is required. Potentially, this trade-off could also be solved with access to larger medical databases. However, as we discuss in the next paragraph, even if PETs allow for the access to more training data, it comes along with several open questions. Given that it is currently unforeseeable whether large training datasets will be unlocked, the trade-off in situations where a restrictive privacy protection is mandated can only be resolved by technical innovations for the privacy-preserving training of AI models. One such potential innovation are the recent developments for DP-Follow-the-regularized-leader (DP-FTRL) as an alternative to DP-SGD [104, 105, 106, 107]. As the name suggests, instead of performing a gradient descent algorithm, it is based on Follow-the-regularized-leader (FTRL) algorithm [108]. Opposed to gradient descent algorithms, which are typically trained on a pre-defined static dataset, FTRL is an online learning algorithm. In the paradigm of online learning algorithms, the AI network is presented sequentially with data samples and, at each step, calculates the best predictor from all previous steps. By this, it is also naturally compatible with FL as no aggregation strategies have to be performed. Based on DP-FTRL, it was shown that strong predictors can be trained under restrictive levels of privacy. However, to the best of our knowledge, no broad comparison between DP-SGD and DP-FTRL –especially on datasets comparable to the requirements for medical AI– has been published yet.

Will PETs Lead To More Training Data? While providing technical implementations of private AI systems is often imposed by legal requirements, it also comes alongside a hope: The hope is that guaranteed privacy can unlock large amounts of previously inaccessible data for the training of AI models. These increased amounts of data leave AI practitioners dreaming of unprecedented capabilities for medical AI. Even more, models trained on data from various regions of the world, generated by all types of medical equipment, covering all sorts of diagnostic modalities, having access to even the rarest conditions. In short, generalist models which can be applied to any patient and diagnosis.

At first glance, the hope for unlocking large amounts of training data is justified: It has been shown that, unlike anonymization procedures, the use of DP –under certain assumptions– fulfils a necessary (but not sufficient) condition for regulations such as the GDPR [20]. Also, empirically, it was shown that users are more inclined to share their data if there are robust technical protective measurements in place [109]. However, there is a catch, namely the privacy budget. It is unclear at what level of privacy DP still complies with the legal requirements. Theoretically, the budget could be chosen to be an incredibly large but non-infinite number. While this gives *a* guarantee, it could be an extremely weak, in the extreme case, a potentially meaningless guarantee. Finding a compromise cannot be solved on a technical or mathematical level and will be the task of ethical, legal and political debate. However, if such debates agree on a certain privacy budget, it implies that this budget must not be exceeded to fulfil legal requirements. As a consequence, medical training data becomes a consumable good. This comes along with a multitude of open and largely undiscussed questions: Is there only one chance to train a generalist AI model on all available data, or should the available data be distributed for several training sessions? Both are valid options. Training on all available data would yield the model with the most information available. On the other hand, one could argue that there is enough data to train several almost-perfect models. This would also mitigate the next arising question: What if new developments allow better AI models, but the privacy budget is already exceeded? Deep learning-based AI models have only begun about a decade ago to be a large research field. Developments in the field are often outdated shortly after publication. So far, it is not foreseeable that this trend will stop. Hence, training on all available data and privacy budget would imply that until fresh large quantities of data are generated no newer and likely superior models can be trained. As alluded to in the first question, a possible solution could be to split up the data and retrain the current state-of-the-art approach after a fixed schedule, e.g., once per year, with all the data that was generated within this timeframe. However, finding the trade-off between retraining and collecting data will be delicate here, as a low frequency would provide more training data but could miss important improvements in the AI development. Nevertheless, even if this trade-off is resolved, it is still unclear who is commissioned to train such a model? Choosing the highest bidding would follow economic models. However, the resulting AI model has the potential to improve all current diagnostic procedures, find previously unknown correlations, and expand the understanding of unresolved medical problems. Yet, *with great power, there must also come great responsibility* [110]. Thus, it may be in the public interest to leave control over such a model to a non-commercial entity.

4.3 Conclusion

In conclusion, this thesis broadly investigated the potentials, challenges, and possible solutions for a privacy-preserving workflow of AI in medicine and healthcare. We demonstrated the practicality of a privacy-preserving workflow composed of various PETs and showed that the resulting private models can compete and even outperform non-private locally trained models as well as expert radiologists. Moreover, we implemented an approach that, at the time of publication, was natively compatible with any type of DP-conformant neural network layer and competitive to state-of-the-art frameworks in terms of runtime. Both are important factors for the practical use of privacy-preserving technologies. Furthermore, we investigated the interaction of private AI models and their fairness in the classification of subgroups. Opposed to prior works, we found it is not necessarily the representation but rather the difficulty of the task that is driving the negative performance impact under stricter privacy levels. This could have potential impacts on the collection of data for the training of privacy-preserving and fair AI models. Lastly, we raised the question of how privacy budgets are set and found that, currently, typical privacy budgets are based on worst-case assumptions. We showed that when relaxing these assumptions to a more realistic scenario, the privacy-utility trade-off is largely mitigated and, in some cases, even vanishes entirely. With these contributions, we hope to facilitate a more widespread breakthrough for privacy-preserving techniques for the AI training on sensitive datasets such as medical data. However, there are open legal and ethical questions that need to be part of a societal discussion about the future of AI models in medicine and healthcare applications.

Bibliography

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. 'Imagenet: A large-scale hierarchical image database'. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al. 'Gpt-4 technical report'. In: *arXiv preprint arXiv:2303.08774* (2023).

[3] R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth et al. 'Gemini: a family of highly capable multimodal models'. In: *arXiv preprint arXiv:2312.11805* (2023).

[4] A. Satariano. 'ChatGPT is banned in Italy over privacy concerns'. In: *The New York Times*. 1 (Mar. 2023).

[5] K. Lång, V. Josefsson, A.-M. Larsson, S. Larsson, C. Högberg, H. Sartor, S. Hofvind, I. Andersson and A. Rosso. 'Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study'. In: *The Lancet Oncology* 24.8 (2023), pp. 936–944.

[6] G. Wang, X. Liu, K. Wang, Y. Gao, G. Li, D. T. Baptista-Hon, X. H. Yang, K. Xue, W. H. Tai, Z. Jiang et al. 'Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution'. In: *Nature Medicine* (2023), pp. 1–12.

[7] S. S. Al-Zaiti, C. Martin-Gill, J. K. Zègre-Hemsey, Z. Bouzid, Z. Faramand, M. O. Alrawashdeh, R. E. Gregg, S. Helman, N. T. Riek, K. Kraevsky-Phillips et al. 'Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction'. In: *Nature Medicine* (2023), pp. 1–10.

[8] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl et al. 'Large language models encode clinical knowledge'. In: *Nature* (2023), pp. 1–9.

[9] L. Y. Jiang, X. C. Liu, N. P. Nejatian, M. Nasir-Moin, D. Wang, A. Abidin, K. Eaton, H. A. Riina, I. Laufer, P. Punjabi et al. 'Health system-scale language models are all-purpose prediction engines'. In: *Nature* (2023), pp. 1–6.

[10] P. Rajpurkar, E. Chen, O. Banerjee and E. J. Topol. ‘AI in health and medicine’. In: *Nature medicine* 28.1 (2022), pp. 31–38.

[11] J. Geiping, H. Bauermeister, H. Dröge and M. Moeller. ‘Inverting gradients-how easy is it to break privacy in federated learning?’ In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 16937–16947.

[12] H. Yin, A. Mallya, A. Vahdat, J. M. Alvarez, J. Kautz and P. Molchanov. ‘See through gradients: Image batch recovery via gradinversion’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 16337–16346.

[13] K.-C. Wang, Y. Fu, K. Li, A. Khisti, R. Zemel and A. Makhzani. ‘Variational model inversion attacks’. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 9706–9719.

[14] N. Haim, G. Vardi, G. Yehudai, O. Shamir and M. Irani. ‘Reconstructing training data from trained neural networks’. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 22911–22924.

[15] L. Fowl, J. Geiping, W. Czaja, M. Goldblum and T. Goldstein. ‘Robbing the fed: Directly obtaining private data in federated learning with modified models’. In: *Tenth International Conference on Learning Representations* (2022).

[16] F. Boenisch, A. Dziedzic, R. Schuster, A. S. Shamsabadi, I. Shumailov and N. Papernot. ‘When the curious abandon honesty: Federated learning is not private’. In: *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2023, pp. 175–199.

[17] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito and E. Wallace. ‘Extracting training data from diffusion models’. In: *32nd USENIX Security Symposium (USENIX Security 23)*. 2023, pp. 5253–5270.

[18] G. Buzaglo, N. Haim, G. Yehudai, G. Vardi, Y. Oz, Y. Nikankin and M. Irani. ‘Deconstructing Data Reconstruction: Multiclass, Weight Decay and General Losses’. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[19] J. Dong, A. Roth and W. J. Su. ‘Gaussian differential privacy’. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.1 (2022), pp. 3–37.

[20] A. Cohen and K. Nissim. ‘Towards formalizing the GDPR’s notion of singling out’. In: *Proceedings of the National Academy of Sciences* 117.15 (2020), pp. 8344–8352.

[21] A. Cohen. ‘Attacks on Deidentification’s Defenses’. In: *31st USENIX Security Symposium (USENIX Security 22)*. 2022, pp. 1469–1486.

[22] B. Balle, G. Cherubin and J. Hayes. ‘Reconstructing training data with informed adversaries’. In: *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2022, pp. 1138–1156.

[23] J. Hayes, S. Mahloujifar and B. Balle. ‘Bounding Training Data Reconstruction in DP-SGD’. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[24] A. Y. Ng, C. J. Oberije, É. Ambrózay, E. Szabó, O. Serfőző, E. Karpati, G. Fox, B. Glocker, E. A. Morris, G. Forrai et al. ‘Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer’. In: *Nature Medicine* 29.12 (2023), pp. 3044–3049.

[25] Y.-R. Wang, K. Yang, Y. Wen, P. Wang, Y. Hu, Y. Lai, Y. Wang, K. Zhao, S. Tang, A. Zhang et al. ‘Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging’. In: *Nature Medicine* (2024), pp. 1–10.

[26] J.-N. Eckardt, C. Röllig, K. Metzeler, P. Heisig, S. Stasik, J.-A. Georgi, F. Kroschinsky, F. Stölzel, U. Platzbecker, K. Spiekermann et al. ‘Unsupervised meta-clustering identifies risk clusters in acute myeloid leukemia based on clinical and genetic profiles’. In: *Communications Medicine* 3.1 (2023), p. 68.

[27] U. Food and D. Administration. *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

[28] A. Jobin, M. Ienca and E. Vayena. ‘The global landscape of AI ethics guidelines’. In: *Nature machine intelligence* 1.9 (2019), pp. 389–399.

[29] S. D. Warren and L. D. Brandeis. ‘Right to privacy’. In: *Harv. L. Rev.* 4 (1890), p. 193.

[30] S. M. Jourard. ‘Some psychological aspects of privacy’. In: *Law & Contemp. Probs.* 31 (1966), p. 307.

[31] A. Westin. *Privacy and Freedom*. Athenum, New York, 1967.

[32] H. Nissenbaum. ‘Privacy as contextual integrity’. In: *Wash. L. Rev.* 79 (2004), p. 119.

[33] D. J. Solove. ‘Conceptualizing privacy’. In: *Calif. L. Rev.* 90 (2002), p. 1087.

[34] D. J. Solove. ‘A taxonomy of privacy’. In: *U. Pa. L. Rev.* 154 (2005), p. 477.

[35] D. J. Solove. *Understanding privacy.* Harvard University Press, May, 2008.

[36] A. Ziller, T. T. Mueller, R. Braren, D. Rueckert and G. Kaissis. ‘Privacy: An axiomatic approach’. In: *Entropy* 24.5 (2022), p. 714.

[37] L. Sweeney. ‘k-anonymity: A model for protecting privacy’. In: *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05 (2002), pp. 557–570.

[38] C. G. Schwarz, W. K. Kremers, T. M. Therneau, R. R. Sharp, J. L. Gunter, P. Vemuri, A. Arani, A. J. Spychalla, K. Kantarci, D. S. Knopman et al. ‘Identification of anonymous MRI research participants with face-recognition software’. In: *New England Journal of Medicine* 381.17 (2019), pp. 1684–1686.

[39] J. Buolamwini and T. Gebru. ‘Gender shades: Intersectional accuracy disparities in commercial gender classification’. In: *Conference on fairness, accountability and transparency.* PMLR. 2018, pp. 77–91.

[40] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove and A. Rieke. ‘Discrimination through optimization: How Facebook’s Ad delivery can lead to biased outcomes’. In: *Proceedings of the ACM on human-computer interaction* 3.CSCW (2019), pp. 1–30.

[41] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan. ‘A survey on bias and fairness in machine learning’. In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35.

[42] R. J. Chen, J. J. Wang, D. F. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai and F. Mahmood. ‘Algorithmic fairness in artificial intelligence for medicine and healthcare’. In: *Nature biomedical engineering* 7.6 (2023), pp. 719–742.

[43] L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen and M. Ghassemi. ‘Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations’. In: *Nature medicine* 27.12 (2021), pp. 2176–2182.

[44] F. Meissen, S. Breuer, M. Knolle, A. Buyx, R. Müller, G. Kaissis, B. Wiestler and D. Rückert. '(Predictable) performance bias in unsupervised anomaly detection'. In: *Ebiomedicine* 101 (2024).

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

[46] S. Warnat-Herresthal, H. Schultze, K. L. Shastry, S. Manamohan, S. Mukherjee, V. Garg, R. Sarveswara, K. Händler, P. Pickkers, N. A. Aziz et al. 'Swarm learning for decentralized and confidential clinical machine learning'. In: *Nature* 594.7862 (2021), pp. 265–270.

[47] S. Pati, U. Baid, B. Edwards, M. Sheller, S.-H. Wang, G. A. Reina, P. Foley, A. Gruzdev, D. Karkada, C. Davatzikos et al. 'Federated learning enables big data for rare cancer boundary detection'. In: *Nature communications* 13.1 (2022), p. 7346.

[48] M. Oldenhof, G. Ács, B. Pejó, A. Schuffenhauer, N. Holway, N. Sturm, A. Dieckmann, O. Fortmeier, E. Boniface, C. Mayer et al. 'Industry-scale orchestrated federated learning for drug discovery'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 13. 2023, pp. 15576–15584.

[49] R. L. Rivest, A. Shamir and L. Adleman. 'A method for obtaining digital signatures and public-key cryptosystems'. In: *Communications of the ACM* 21.2 (1978), pp. 120–126.

[50] X. Yi, R. Paulet, E. Bertino, X. Yi, R. Paulet and E. Bertino. *Homomorphic encryption*. Springer, 2014.

[51] A. Acar, H. Aksu, A. S. Uluagac and M. Conti. 'A survey on homomorphic encryption schemes: Theory and implementation'. In: *ACM Computing Surveys (Csur)* 51.4 (2018), pp. 1–35.

[52] M. Naehrig, K. Lauter and V. Vaikuntanathan. 'Can homomorphic encryption be practical?' In: *Proceedings of the 3rd ACM workshop on Cloud computing security workshop*. 2011, pp. 113–124.

[53] A. Al Badawi, C. Jin, J. Lin, C. F. Mun, S. J. Jie, B. H. M. Tan, X. Nan, K. M. M. Aung and V. R. Chandrasekhar. 'Towards the alexnet moment for homomorphic encryption: Hcnn, the first homomorphic cnn on encrypted data with gpus'. In: *IEEE Transactions on Emerging Topics in Computing* 9.3 (2020), pp. 1330–1343.

[54] R. Cramer, I. B. Damgård et al. *Secure multiparty computation.* Cambridge University Press, 2015.

[55] Y. Lindell. 'Secure multiparty computation'. In: *Communications of the ACM* 64.1 (2020), pp. 86–96.

[56] C. Dwork, A. Roth et al. 'The algorithmic foundations of differential privacy'. In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407.

[57] G. Kaissis, M. Knolle, F. Jungmann, A. Ziller, D. Usynin and D. Rueckert. 'A unified interpretation of the gaussian mechanism for differential privacy through the sensitivity index'. In: *Journal of Privacy and Confidentiality* 12.1 (2022).

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

[59] Q. Geng and P. Viswanath. 'The optimal noise-adding mechanism in differential privacy'. In: *IEEE Transactions on Information Theory* 62.2 (2015), pp. 925–951.

[60] N. Fernandes, A. McIver and C. Morgan. 'The laplace mechanism has optimal utility for differential privacy over continuous queries'. In: *2021 36th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS).* IEEE. 2021, pp. 1–12.

[61] Y. Zhu, J. Dong and Y.-X. Wang. 'Optimal accounting of differential privacy via characteristic function'. In: *International Conference on Artificial Intelligence and Statistics.* PMLR. 2022, pp. 4782–4817.

[62] I. Mironov. 'Rényi differential privacy'. In: *2017 IEEE 30th computer security foundations symposium (CSF).* IEEE. 2017, pp. 263–275.

[63] I. Mironov, K. Talwar and L. Zhang. 'Rényi differential privacy of the sampled gaussian mechanism'. In: *arXiv preprint arXiv:1908.10530* (2019).

[64] J. Neyman and E. S. Pearson. 'IX. On the problem of the most efficient tests of statistical hypotheses'. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706 (1933), pp. 289–337.

[65] G. Kaissis, S. Kolek, B. de Balle Pigem, J. Hayes and D. Rueckert. 'Beyond the Calibration Point: Mechanism Comparison in Differential Privacy'. In: *International Conference on Machine Learning*. PMLR. 2024.

[66] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow and K. Talwar. 'Semi-supervised knowledge transfer for deep learning from private training data'. In: *Fifth International Conference on Learning Representations* (2017).

[67] S. Song, K. Chaudhuri and A. D. Sarwate. 'Stochastic gradient descent with differentially private updates'. In: *2013 IEEE global conference on signal and information processing*. IEEE. 2013, pp. 245–248.

[68] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar and L. Zhang. 'Deep learning with differential privacy'. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318.

[69] S. Gopi, Y. T. Lee and L. Wutschitz. 'Numerical composition of differential privacy'. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11631–11642.

[70] M. Knolle, R. Dorfman, A. Ziller, D. Rueckert and G. Kaissis. 'Bias-Aware Minimisation: Understanding and Mitigating Estimator Bias in Private SGD'. In: *Theory and Practice of Differential Privacy* (2023).

[71] H. B. McMahan, D. Ramage, K. Talwar and L. Zhang. 'Learning differentially private recurrent language models'. In: *The Sixth International Conference on Learning Representations* (2018).

[72] J. Zhang, T. He, S. Sra and A. Jadbabaie. 'Why gradient clipping accelerates training: A theoretical justification for adaptivity'. In: *The Seventh International Conference on Learning Representations* (2019).

[73] J. Qian, Y. Wu, B. Zhuang, S. Wang and J. Xiao. 'Understanding gradient clipping in incremental gradient methods'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1504–1512.

[74] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al. 'Pytorch: An imperative style, high-performance deep learning library'. In: *Advances in neural information processing systems* 32 (2019).

[75] I. Goodfellow. 'Efficient per-example gradient computations'. In: *arXiv preprint arXiv:1510.01799* (2015).

[76] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao, G. Cormode and I. Mironov. 'Opacus: User-Friendly Differential Privacy Library in PyTorch'. In: *arXiv preprint arXiv:2109.12298* (2021).

[77] R. Cummings, V. Gupta, D. Kimpara and J. Morgenstern. 'On the compatibility of privacy and fairness'. In: *Adjunct publication of the 27th conference on user modeling, adaptation and personalization.* 2019, pp. 309–315.

[78] E. Bagdasaryan, O. Poursaeed and V. Shmatikov. 'Differential privacy has disparate impact on model accuracy'. In: *Advances in neural information processing systems* 32 (2019).

[79] T. Farrand, F. Mireshghallah, S. Singh and A. Trask. 'Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy'. In: *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice.* 2020, pp. 15–19.

[80] C. Tran, F. Fioretto, P. Van Hentenryck and Z. Yao. 'Decision Making with Differential Privacy under a Fairness Lens.' In: *IJCAI.* 2021, pp. 560–566.

[81] A. Sanyal, Y. Hu and F. Yang. 'How unfair is private learning?' In: *Uncertainty in Artificial Intelligence.* PMLR. 2022, pp. 1738–1748.

[82] P. Mangold, M. Perrot, A. Bellet and M. Tommasi. 'Differential privacy has bounded impact on fairness in classification'. In: *International Conference on Machine Learning.* PMLR. 2023, pp. 23681–23705.

[83] M. Yang, M. Ding, Y. Qu, W. Ni, D. Smith and T. Rakotoarivelo. 'Privacy at a Price: Exploring its Dual Impact on AI Fairness'. In: *arXiv preprint arXiv:2404.09391* (2024).

[84] B. Kulynych, Y.-Y. Yang, Y. Yu, J. Błasiok and P. Nakkiran. 'What you see is what you get: Principled deep learning via distributional generalization'. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 2168–2183.

[85] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu and S. Jana. 'Certified robustness to adversarial examples with differential privacy'. In: *2019 IEEE symposium on security and privacy (SP).* IEEE. 2019, pp. 656–672.

[86] V. Feldman. 'Does learning require memorization? a short tale about a long tail. corr abs/1906.05271 (2019)'. In: *arXiv preprint arXiv:1906.05271* (2019).

[87] S. Ioffe and C. Szegedy. ‘Batch normalization: Accelerating deep network training by reducing internal covariate shift’. In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.

[88] Y. Wu and K. He. ‘Group normalization’. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[89] R. Nasirigerdeh, R. Torkzadehmahani, D. Rueckert and G. Kaissis. ‘Kernel Normalized Convolutional Networks’. In: *Transactions on Machine Learning Research* (2024).

[90] R. Nasirigerdeh, J. Torkzadehmahani, D. Rueckert and G. Kaissis. ‘Kernel normalized convolutional networks for privacy-preserving machine learning’. In: *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE. 2023, pp. 107–118.

[91] N. Papernot and T. Steinke. ‘Hyperparameter tuning with renyi differential privacy’. In: *The Tenth International Conference on Learning Representations* (2022).

[92] S. P. Benthall. *Context, Causality, and Information Flow: Implications for Privacy Engineering, Security, and Data Economics*. University of California, Berkeley, 2018.

[93] M. Nasr, S. Songi, A. Thakurta, N. Papernot and N. Carlini. ‘Adversary instantiation: Lower bounds for differentially private machine learning’. In: *2021 IEEE Symposium on security and privacy (SP)*. IEEE. 2021, pp. 866–882.

[94] S. Feng and F. Tramèr. ‘Privacy Backdoors: Stealing Data with Corrupted Pretrained Models’. In: *International Conference on Machine Learning*. PMLR. 2024.

[95] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier and H. Jégou. ‘White-box vs black-box: Bayes optimal strategies for membership inference’. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5558–5567.

[96] D. Usynin, A. Ziller, M. Makowski, R. Braren, D. Rueckert, B. Glocker, G. Kaissis and J. Passerat-Palmbach. ‘Adversarial interference and its mitigations in privacy-preserving collaborative machine learning’. In: *Nature Machine Intelligence* 3.9 (2021), pp. 749–758.

[97] 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)*. IEEE. 2017, pp. 3–18.

[98] G. Kaissis, A. Ziller, S. Kolek, A. Riess and D. Rueckert. 'Optimal privacy guarantees for a relaxed threat model: Addressing sub-optimal adversaries in differentially private machine learning'. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[99] J. Ho, A. Jain and P. Abbeel. 'Denoising diffusion probabilistic models'. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[100] J. Song, C. Meng and S. Ermon. 'Denoising diffusion implicit models'. In: *Ninth International Conference on Learning Representations* (2021).

[101] L. Melis, C. Song, E. De Cristofaro and V. Shmatikov. 'Exploiting unintended feature leakage in collaborative learning'. In: *2019 IEEE symposium on security and privacy (SP)*. IEEE. 2019, pp. 691–706.

[102] A. Ziller, A. Riess, K. Schwethelm, T. T. Mueller, D. Rueckert and G. Kaissis. 'Bounding Reconstruction Attack Success of Adversaries Without Data Priors'. In: *arXiv preprint arXiv:2402.12861* (2024).

[103] G. K. Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, 1949.

[104] P. Kairouz, B. McMahan, S. Song, O. Thakkar, A. Thakurta and Z. Xu. 'Practical and private (deep) learning without sampling or shuffling'. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5213–5225.

[105] Z. Xu, Y. Zhang, G. Andrew, C. A. Choquette-Choo, P. Kairouz, H. B. McMahan, J. Rosenstock and Y. Zhang. 'Federated learning of gboard language models with differential privacy'. In: *arXiv preprint arXiv:2305.18465* (2023).

[106] C. A. Choquette-Choo, A. Ganesh, T. Steinke and A. Thakurta. 'Privacy Amplification for Matrix Mechanisms'. In: *arXiv preprint arXiv:2310.15526* (2023).

[107] C. A. Choquette-Choo, A. Ganesh, R. McKenna, H. B. McMahan, J. Rush, A. Guha Thakurta and Z. Xu. '(Amplified) Banded Matrix Factorization: A unified approach to private training'. In: *Advances in Neural Information Processing Systems* 36 (2024).

[108] B. McMahan. 'Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization'. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 525–533.

[109] R. Cummings, G. Kaptchuk and E. M. Redmiles. '" I need a better description": An Investigation Into User Expectations For Differential Privacy'. In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2021, pp. 3037–3052.

[110] S. Lee and S. Ditko. *Amazing Fantasy #15*. New York: Marvel Comics, 1962.

Appendices

A End-to-end privacy preserving deep learning on multi-institutional medical imaging - Supplementary Information

nature portfolio

https://doi.org/10.1038/s42256-021-00337-8

Supplementary information

End-to-end privacy preserving deep learning on multi-institutional medical imaging

End-to-end privacy preserving deep learning on multi-institutional medical imaging

Supplementary Material

1 Hyperparameter Optimisation

As described in the main manuscript, PriMIA provides the ability to perform centrally coordinated hyperparameter optimisation provided by the *Optuna* framework [1]. By default, the *Tree-Structured Parzen Estimator* algorithm [2] is utilised, but this functionality is user-adjustable. By default, the framework will search over the following parameters: Learning rate, optimiser parameters, learning rate scheduler restarts, weight decay, class-weighting for the loss, federated averaging, augmentation parameters, including the probabilities of each individual augmentation, synchronisation rate and class-weighting for federated averaging. These options are user-adjustable in case more or fewer parameters are desired.

The *Optuna* library we utilise offers a number of visualisations, including parameter importance and parallel coordinate plots [3] of the hyperparameter optimisation procedure. We demonstrate an example of the procedure used to determine the optimal parameters for the FL model presented in the main manuscript in Figure 1. Moreover, we offer the option to save the individual hyperparameter optimisation trials to a relational database back-end, which can be used to resume runs without repeating fruitless hyperparameter combinations. Lastly, the framework offers a number of *pruning* options, that is, early stopping of a run if it evolves in an unpromising direction compared to previous runs. Further documentation can be found at `https://optuna.readthedocs.io/en/v2.0.0/`.

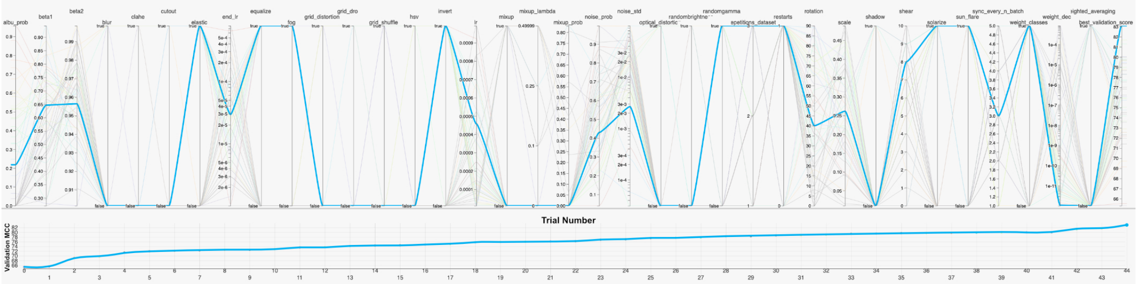


Figure 1: **Parallel coordinate plot of the hyperparameter optimisation run leading to the FL model used in the main manuscript (DP-/SecAgg+).** The top panel shows parallel vertical axes representing the individual variables. Faint lines represent the individual trials, connecting the corresponding variable values. The cyan line represents the "best" trial, and connects the optimal hyperparameters found within the run. The bottom panel shows the evolution in MCC over the trials of the optimisation procedure ordered by ascending MCC. Importantly, the ascending ordering of the bottom panel's X-axis does **not** represent the order in which the trials were performed. In this case, the optimal MCC accuracy was encountered earlier during the procedure (trial 26). The ascending ordering of the trials is applied *post-hoc* for visual clarity.

2 Statistical evaluation and inter-rater/model agreement

The McNemar test was used to test for differences between the predictions of human observers and/or the models trained using FL/ centrally as defined in the main manuscript. Results can be found in Table 1.

	Federated DP-/SecAgg+	Federated DP-/SecAgg-	Centrally Trained	Federated DP+/SecAgg+	Expert 1
Federated DP-/SecAgg-	1.00				
Centrally Trained	0.48	1.00			
Federated DP+/SecAgg+	**0.04**	**0.02**	**0.01**		
Expert 1	**0.03**	**0.02**	**0.01**	0.27	
Expert 2	**0.02**	**0.01**	**0.01**	0.24	0.88

Table 1: **McNemar test results on Test Set 1.** *Federated*: Model trained with federated learning. *DP+/-*: Model trained with (+) or without (-) differentially private gradient descent. *SecAgg+/-*: Model trained with (+) or without (-) secure aggregation. *Centrally Trained*: Model trained on the entire dataset on a single machine. *Expert 1/2*: Human experts. **Bold** text signifies p<0.05. Duplicated values omitted.

As seen in the table, the FL DP-/SecAgg- and FL DP-/SecAgg+ models did not produce significantly different predictions from each-other or from the centrally trained model (p-values >0.05) but significantly outperformed human observers (p-values <0.05). The DP+ model performed significantly below the models trained without DP and the centrally trained model (p-values <0.05), but on-par with human observers (p-values >0.05).

Cohen's κ was used to assess inter-model/observer agreement as described in the main manuscript. Numbers close to 1.0 indicate high agreement. Results are shown in Table 2.

	Federated DP-/SecAgg+	Federated DP-/SecAgg-	Centrally Trained	Federated DP+/SecAgg+	Expert 1
Federated DP-/SecAgg-	0.99				
Centrally Trained	0.98	0.99			
Federated DP+/SecAgg+	0.94	0.93	0.91		
Expert 1	0.57	0.59	0.63	0.55	
Expert 2	0.60	0.62	0.60	0.59	0.51

Table 2: **Cohen's κ between models and observers on Test Set 1.** Duplicate values omitted.

Inter-observer agreement (according to McHugh [4]) was almost perfect between the models, moderate to strong between models and observers and moderate between observers. This indicates high reliability of the model's predictions on out-of-sample data and therefore good generalisation performance.

3 Statistical evaluation of training and inference benchmarks

The statistical evaluation of the results shown in Figure 3 of the main manuscript was performed using one-way analysis of variance (ANOVA) for results in panels **A** to **D** followed by pairwise Student's t-tests, and Student's t-test for results in Panel **E**. Statistically significant results were found for Panels **A**, **B** and **C**. In Panel **A**, the difference between DP-/SecAgg+ and DP+/SecAgg- was the only non-significant result (p=0.08), all other pairwise comparisons were significant at the p<0.001 level. In Panel **B**, all pairwise comparisons were significant at the p<0.001 level. In Panel **C**, the only significant results were found in the middle subpanel (DP-/SecAgg+), where all pairwise comparisons were significant at the p<0.001 level except the comparison between 5 and 8 workers (p=0.02). In panel **E**, all pairwise comparisons were significant at the <0.001 level.

4 Statistical evaluation of gradient-based attacks

The statistical evaluation of the results shown in Figure 4, panel **B** of the main manuscript was performed using one-way ANOVA followed by pair-wise Student's t-tests. All results were significant at the <0.001 level.

5 Auxiliary reconstruction attack figures

As mentioned in the main manuscript, DP negated the attacks against the paediatric pneumonia dataset, consistent with the guarantees DP provides. Results from the unprotected centrally trained model, the DP-/SecAgg+ and the DP+/SecAgg+ models are shown in Figure 4 of the main manuscript. Figure 2 shows the attack against the centrally trained model with DP as well as the DP+/SecAgg- model. The DP training procedure again negates the attacks against the dataset, whereas the centrally trained model without DP is susceptible to catastrophic privacy loss.

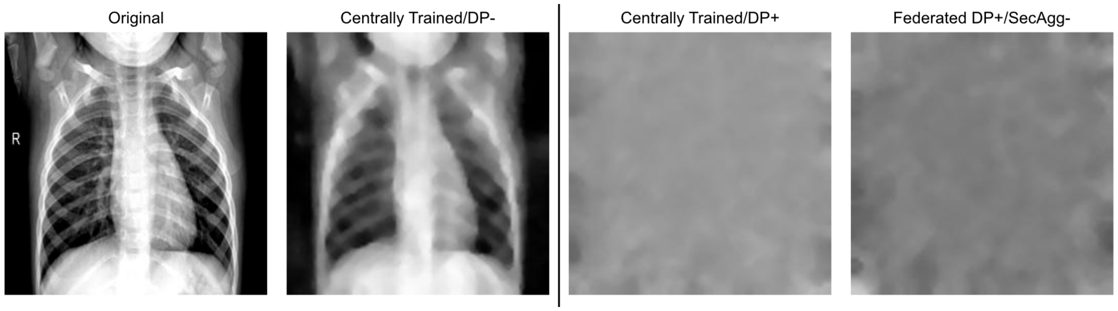


Figure 2: **Results from the gradient-based inversion attack described in the main manuscript**. Attacks were performed against the centrally trained model trained with DP (*Centrally Trained/DP+*) and the FL model trained with DP but without SecAgg (*DP+/SecAgg-*). Even in the *best-case* scenario of local training, DP negates the effects of the attack. The original image and centrally trained model without DP (identical to the main manuscript) are shown for reference on the left hand side.

6 Attacks against the MedNIST dataset

As described in the section on model inversion attacks and shown in Figure 5 of the main manuscript, we conducted additional experimentation using the MedNIST dataset[1]. The dataset includes 58954 images in 6 categories with a size of 64x64 pixels. CT images in the *AbdomenCT* and *ChestCT* categories were histogram normalised using the PIL library[2]. We utilised the ResNet18 architecture described in the main manuscript. Models were trained without either DP training or secure aggregation (DP-/SecAgg-) and with DP training and secure aggregation (DP+/SecAgg+). For the DP-/SecAgg- training, we modelled an adversary intercepting a gradient update from a single data owner at batch size 1 immediately at the beginning of training, representing a worst-case privacy loss scenario. The DP+/SecAgg+ model was also trained at a batch size of 1, however -due to SecAgg-, the *effective batch size* was 3 (at the used synchronisation parameter of 1, models are averaged after one training step). We did not train the DP model to convergence, hence no final ϵ value for the model is provided. Gradient norms were clipped to 1.0, the noise multiplier and sampling rates were set to 1.0. By application of the Rényi Differential Privacy Accountant [5], one epoch of this training (19652 steps at a sampling rate of 0.00509%) would have resulted in a privacy loss of ϵ=0.64 at a δ of 10^{-5} and an optimal α-value (divergence order) of 9.0.

7 Liver Segmentation Case Study

To showcase PriMIA's flexibility in configuring the framework to work with different medical imaging modalities, we here provide an additional case study in which PriMIA is used to train a semantic segmentation model using FL. For this, the modifications required to the PriMIA source code are: adding the desired model architecture to the codebase, modifying the training procedure to load the appropriate model and modifying the loss function and evaluation metrics. Individual training settings (augmentation etc.) are set in the central configuration file. The number of clients for training is derived automatically by the number of unique IP addresses present in the relevant configuration file.

The task chosen was liver segmentation in computed tomography (CT) imaging. The Medical Segmentation Decathlon dataset[3] liver dataset was chosen, which contains 131 training and 70 test 3D volumes. The training volumes were split into a 90% training and 10% validation set. Images were pre-processed by conversion to 32-bit floating point numbers in the following way: DICOM headers were parsed and acquisition pixel values were converted to Hounsfield Units. The mean and standard deviation of the training dataset was used to normalise the training and validation datasets. Images were resized to 256x256 pixels by nearest neighbour interpolation. Hounsfield units were clipped to the range (-150, 250) to yield an abdominal parenchymal window setting. The liver and tumour masks were merged yielding a two-class segmentation task. The UNet model architecture [6] was utilised, but the backbone was modified to use an untrained ResNet18 network using the *SegmentationModels PyTorch* library [7] and modified so that the final layer's logits pass through a sigmoid activation function to be bounded between 0 and 1. A batch size of 32 was selected and the Adam optimiser was used with a log-linearly decreasing learning rate of 10^{-3} to 10^{-4} with one warm restart. L_2-regularistaion was applied with a coefficient of 10^{-3}. No image augmentation was used. The synchronisation parameter was set to 40. The *Dice* loss-function was used for training and the *Dice* score was employed for evaluation [8]. The model was trained for 100 epochs and cached after every epoch; the model with the highest Dice score was used for evaluation on the test set. The best model achieved a

[1] https://github.com/Project-MONAI/tutorials/blob/master/2d_classification/mednist_tutorial.ipynb

[2] https://doi.org/10.5281/zenodo.596518

[3] http://medicaldecathlon.com

Dice Score of 0.94 on the separate test set. An exemplary segmentation mask produced by the model is shown in Figure 3.

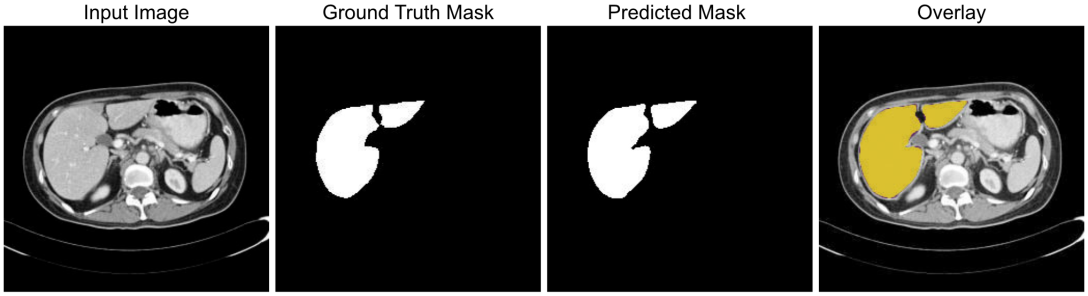


Figure 3: **Exemplary result from the liver FL segmentation case study.** The input image is a computed tomography slice of the abdomen on the level of the liver. The *Ground Truth* segmentation mask and the *Predicted* segmentation masks are shown. The *Overlay* image shows the predicted mask in yellow and the Ground Truth mask in purple. Non-overlapping regions (purple shine-through) indicate model errors.

8 DP model training and additional considerations

PriMIA includes two procedures in which DP mechanisms are used: The DP secure aggregation of dataset statistics and DP neural network training. We note that for our case studies and analysis, we assume that the datasets on the FL nodes are disjoint. We furthermore assume that each patient is included only once in each training dataset. We thus provide *per-patient* privacy guarantees, which in our case is equivalent to *per-record* guarantees. Lastly we point out that we will refer to *differentially private stochastic gradient descent* (DP-SGD) even in cases the Adam optimiser was used, as the subroutine described is independent of the choice between Adam and Stochastic Gradient Descent. (In brief, momentum-based algorithms retain a weighted moving average of previous gradient updates. However, since the gradients stored are privatised, the utilisation of these updates is privacy-neutral due to *closure under post-processing*, see below).

8.1 DP secure aggregation of dataset statistics

As described in the main manuscript, performing remote inference requires the incoming data to be rescaled with the mean and standard deviation (=statistics) of the training dataset. Training using PriMIA is not limited to imaging data. When non-imaging data are included (e.g. age), such statistics can contain sensitive information which should not be disclosed. In the case of FL, the statistics are aggregated over the federation. To prevent the leakage of sensitive data, a differentially private mean and standard deviation query procedure is used on each individual node. Here, a user-defined ϵ-value is used to determine the magnitude of Laplacian noise added to the mean and standard deviation before it is released by the formula $V_{DP} \leftarrow V + Lap(\frac{L}{\epsilon})$, where V is the value in question, L and is the L_1-sensitivity of the query. This procedure (*Laplacian mechanism*) is ϵ-DP (Proof in [9], Section 3.3). The secure aggregation (i.e. averaging) of the thus *privatised* values does not impact privacy guarantees as it represents *post-processing* of non-overlapping databases (Proof in [9], Proposition 2.1).

8.2 Differentially private neural network training

8.2.1 Technical overview of the DP training procedure

We first describe the rationale behind the DP-SGD algorithm [10] and point out specific considerations applicable to PriMIA. Neural network training with gradient descent can be made to preserve (ϵ, δ)-DP ([9], Definition 2.4) by applying the following modifications:

1. Bounding the per-example L_2 sensitivity of the gradient calculation in the backpropagation step by *clipping* the global gradient norm[4]

2. Adding random Gaussian noise to the bounded gradient before performing a gradient descent step

Each gradient descent step can therefore be considered a differentially private release of a gradient, which can then be freely post-processed (e.g. averaged with other gradients). See [12], Theorem 1 for the proof.

The DP-SGD procedure therefore has two configuration parameters: The clipping norm and a *noise multiplier*, which scales the noise added (in addition to the noise required to obfuscate the gradient signal). These parameters can be selected by the user in PriMIA. Notably, selecting the parameters does not require interacting with the dataset, as the privacy guarantees can be calculated in advance and are independent of the data. A convenience script is provided in PriMIA for this purpose.

In practice, this procedure, in particular subroutine (1), introduces added complexity. Reverse-mode automatic differentiation systems do not provide access to the *per-example* gradients of a minibatch by default, but only to an accumulated (either averaged or summed) gradient. The following solutions have been introduced to tackle this issue:

- Goodfellow [13] introduces a technique which allows efficient per-example gradient computation, this procedure is however limited to specific types of neural network layers and cannot be universally applied. A variant of this technique (see comments in [14]) is used by current DP-SGD frameworks such as Opacus[5]. The downside of this technique is that the introduction of newer layer types (e.g. recurrent neural network layers) requires layer-specific modifications for compatibility. We therefore decided against employing this technique in PriMIA, aiming for a generic implementation. The benefit of the technique is that it only introduces minimal computational overhead, as it leverages the native capabilities of the automatic differentiation system.

- A universal technique for per-example gradient computation which is independent of layer type and can be applied to any differentiable layer is the individual processing of samples from a minibatch in separate forward and backpropagation passes. Each computed gradient is then individually processed and stored in a temporary data store. After all examples from a minibatch have been computed, the per-example gradients are averaged or summed (in our case averaged), the gradient descent step taken and the data store flushed. The benefit of this technique is its abovementioned flexibility and simplicity. PriMIA uses this technique for per-example gradient computation. Its main drawback is the increased time complexity. Newer techniques utilising *just-in-time* compilation and vectorised mapping [15] reduce this overhead significantly. However, at the time of this manuscript's preparation, PyTorch, on which PriMIA depends, does not yet support vectorised mapping for gradient calculations. Of note, both above-mentioned techniques introduce increased storage requirements for the per-sample gradients in a temporary data-store.

[4] We note that this subroutine relies on a specific method of clipping referred to as *global norm clipping* as described in [11]

[5] `https://opacus.ai`

- A modification of the second technique was proposed in [16]. Here, instead of every sample being processed separately, so-called *microbatches* are used, which are subsets of the minibatch. Processing such microbatches has the benefit of allowing (sometimes substantially) faster processing. However, proportionally larger magnitude of Gaussian noise has to be added to hide the contribution of the microbatch (for a theoretical examination of the relationship between noise magnitude and privacy guarantees see e.g. [17]). The microbatching technique represents a trade-off between computational speed and algorithm utility. PriMIA includes the option to utilise this technique via the *microbatch size* configuration parameter.

An additional consideration comes from the calculation of the *privacy budget*. Abadi et al. [10] introduce the *Moments Accountant* technique which provides a more realistic outlook on the true privacy cost of the DP-SGD procedure compared to naive methods of calculation based on *composition* ([9], Section 3.5). However, the utilisation of this accounting technique, relies on *subsampling amplification* guarantees provided by so-called *Poisson* sampling of the minibatches. The utilisation of random shuffling followed by serial processing of the samples which is common in deep learning does not satisfy this sampling scheme. PriMIA therefore introduces a separate *Poisson batch sampler* for the DP training procedure. An analysis of various sampling schemes and their guarantees can be found in [18]. Of note, PriMIA utilises a newer version of the Moments Accountant algorithm (*Rényi Differential Privacy Accountant* [5]) which is more numerically stable but otherwise provides identical guarantees.

Lastly, we point out that DP-SGD requires the modification of neural network architectures in case they contain layers which track state over more than one minibatch during the forward pass (which is non-private). An example are Batch Normalisation (BN) layers [19]. As these layers track a moving average of several minibatches, their gradients cannot be clipped *per-example* as -by definition- it is already averaged over multiple examples. In practice, BN layers can be naively replaced by Group Normalisation (GN) [20] layers, which is the strategy followed in PriMIA. More principled solutions to this challenge entail the design of differentially private normalisation layers. Some work in this regard has recently been demonstrated [21].

8.2.2 Practical DP neural network training considerations

For the case study presented in the main manuscript, we assumed the Paediatric Pneumonia Dataset to be private. In the setting of DP, this entails selecting a *privacy budget*. This budget becomes depleted by interacting with the dataset and after it is exhausted, no further interaction with the dataset is allowed. To reconcile the complication of training FL models on private data with this notion, [10] propose the utilisation of a publicly available dataset for tuning the parameters of the DP-SGD algorithm and pre-training a model. Thus, efficient transfer learning can be employed to reach convergence in a smaller amount of training steps, a large proportion of the model can be *frozen* (that is, made untrainable), therefore greatly diminishing the amount of noise added and training becomes faster. Furthermore, hyperparameter optimisation runs are avoided, which would quickly deplete the privacy budget.

In the current study, we used the ChestX-ray8 dataset [22], containing ca. 100,000 chest radiographies with the following modifications: We merged the classification labels to yield a two-class classification problem of "normal" vs. "abnormal" radiograph. Furthermore, we split the test set of 25595 images into a subset of 11211 *validation* images and a subset of 14384 *test* images.

Individual images were processed and distributed to FL nodes as detailed in the main manuscript. We used a δ value of $1.9 * 10^{-4}$ for privacy calculation during training as the paediatric pneumonia dataset contains approximately 5000 images (see [9], page 18 for rationale). For training on the public dataset, we arbitrarily set a privacy budget of $\epsilon = 10$ after which training was automatically aborted. We performed a brute-force grid search over 15 linearly spaced values of the clipping norm parameter from 0.8 to 1.2 and of the noise multiplier parameter from 1.0 to 3.0. We trained a total of three models in this fashion:

- A ResNet18 model with BN layers replaced by GN and randomly initialised with the He uniform initialiser (trained *from scratch*)
- A ResNet18 model pretrained on ImageNet with BN layers replaced by GN after training, the final linear layer replaced by a newly initialised linear layer with 512 units (identical model to the main manuscript except the use of GN)
- A ResNet18 model non-privately pretrained to convergence on the ChestX-ray8 dataset. Following [10], we later used the median gradient norm of this non-private training as the clipping norm of the privately trained model on the paediatric pneumonia dataset

Through this evaluation, we empirically determined the model pre-trained on the ChestX-ray8 dataset to yield a beneficial privacy-utility *Pareto* frontier compared to ImageNet pre-training and -especially- training *from scratch*. The results of this evaluation can be found in Figure 4.

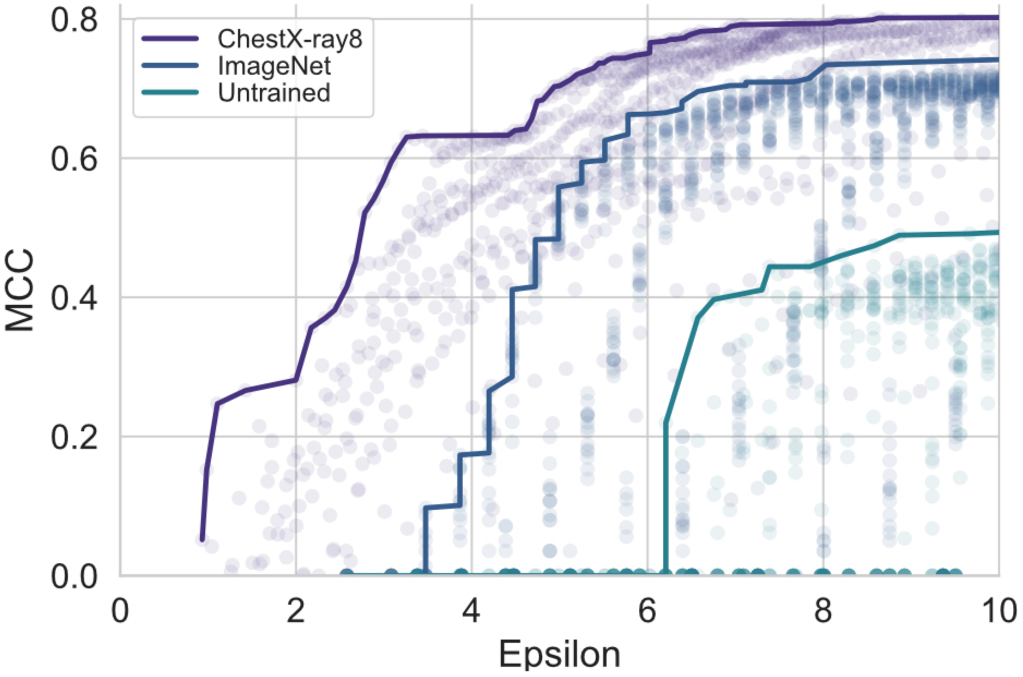


Figure 4: **Empirical *Pareto*-curves of the models trained *from scratch* (Untrained), pre-trained on ImageNet and pre-trained on the ChestX-ray8 dataset.** Curves represent the *privacy-utility-trade-off* for various models by demonstrating the highest classification performance reached at various ϵ thresholds. The Y-axis shows the MCC achieved on the separate test set. The model pretrained on ChestX-ray8 consistently *Pareto*-dominates the other models. MCC gains beyond an ϵ-value of 6.0 are minimal. At $\epsilon = 6.0$, the model achieves an MCC of 0.78.

Based on these results, we selected an ϵ=6.0 as a cut-off for private model training. We utilised the model pre-trained on the ChestX-ray8 dataset for training on the paediatric pneumonia dataset and *froze* (rendered untrainable) all but the last linear layer. At the selected clipping norm of 1.07 and noise multiplier of 3.0, the privately trained model on the paediatric pneumonia dataset achieved an MCC of 0.78 at the ϵ cut-off of 6.0 as described in the main manuscript, indicating efficient utilisation of the pre-training procedure.

Additional considerations

We note that further improvements in the form of optimal learning rate selection, choice of optimiser, etc. could have yielded improved performance of our trained model. Moreover, our procedure relies on purely independent noise sources for each institution which provides stronger-than-required privacy guarantees without accounting for it, thus likely harming utility. Results from newer works demonstrate the utilisation of a central noise source in addition to independent noise to mitigate the utility penalty of our approach [23]. We leave the exploration of such techniques as an interesting direction for future work.
Importantly, the privacy guarantees offered by our DP implementation are not specific to PriMIA but are related to dataset size, the DP-SGD algorithm itself as well as the privacy analysis technique and should be chosen based on user requirements. The privacy guarantee chosen in our case study was arbitrary and represented a good empirical cut-off between privacy and utility, with a bias towards higher model utility. Users of PriMIA can choose a different ϵ-value for their individual use-case and suited to the privacy guarantees they wish to offer to the patients included in their dataset. We note also that the choice of numerical value of ϵ can be complex and the individual privacy requirements are domain dependent [24].

We chose to utilise DP-SGD since it represents an algorithm with few assumptions which naturally fits the deep learning model development process and the flexibility of our framework for different modelling use-cases. Alternative approaches to DP model training have been proposed by Papernot et al., who introduced the Private Aggregation of Teacher Ensembles (PATE) method [25]. The strong assumptions required for PATE render it unsuitable for our framework. For one, PATE was not primarily designed as a collaborative training method and requires strictly disjoint subsets of data, which cannot be assumed in FL workflows. Training the student requires black-box access to predictions of a non-privately trained model, which in itself poses a privacy risk [26]. DP-SGD makes no such assumption. Furthermore, the PATE algorithm was designed for classification, while our framework aims at providing generic medical imaging analysis functionality such as segmentation, as described in the section above. The work by Fay et al. [27] attempted to perform federated segmentation using PATE, however the model only achieved an ϵ-value of around 125 while still suffering a considerable performance penalty. More importantly, PATE requires a separate large, public unlabelled dataset representing the identical task as the one used to train the teacher models. In medical imaging, this is problematic, as the existence of such datasets cannot be assumed. DP-SGD allows *transfer-learning* on arbitrary datasets, as seen above with ImageNET and ChestX-ray8 pretraining, both different from the three-class classification problem of our case study.

9 Secure Multi-Party Computation overview

As mentioned in the main manuscript, specific SMPC operations (e.g. matrix multiplication) are conducted in two phases, an *offline phase* and an *online phase.* The term *preprocessing phase* is also used for the offline phase. We use the terminology from SPDZ ([28]) for these phases, as our protocol uses a similar implementation [29]. For more details, we refer to [30].

The offline phase is based on the generation of *cryptographic primitives.* In practice, these consist of *multiplication triples* and *message authentication codes.* The former are required to perform multiplications [31], the latter are utilised to ensure computational correctness. The generation of primitives is computationally expensive, which is why it is performed in the offline phase. After their production, the primitives are distributed to the parties for use in designated steps of the protocol. The entity generating the primitives is referred to in PriMIA as the *cryptographic provider*. In plain terms, the cryptographic provider is a cryptographically secure random number generator which never comes in contact with any party's data. The parties are only required to trust the cryptographic primitives themselves. Therefore, a cryptographic provider disjoint from the federation can be chosen, whose interest of preserving their reputation prevent them from engaging in dishonest behaviour. In theory, primitives can also be generated by secure hardware devices at each participating institution of the federation, provided they are coordinated, e.g. by a shared initialisation procedure.

During the online phase, the computations are performed between parties. Primitives are consumed at every step of the procedure. The cryptographic provider does not participate in this phase, which occurs between the individual parties. As noted in the main manuscript, a "stockpile" cryptographic primitives can also be generated ahead of time and distributed to parties to be gradually used up. Lastly, it should be noted that the utilisation of the cryptographic provider greatly reduces the time complexity of the protocol. Protocols for *two-party-computation* without cryptographic providers also exist (e.g. *circuit garbling*, compare [32]) but suffer from unacceptably high latency for neural network inference, which requires a large number of operations. An overview of the SMPC process in PriMIA is presented in Figure 5.

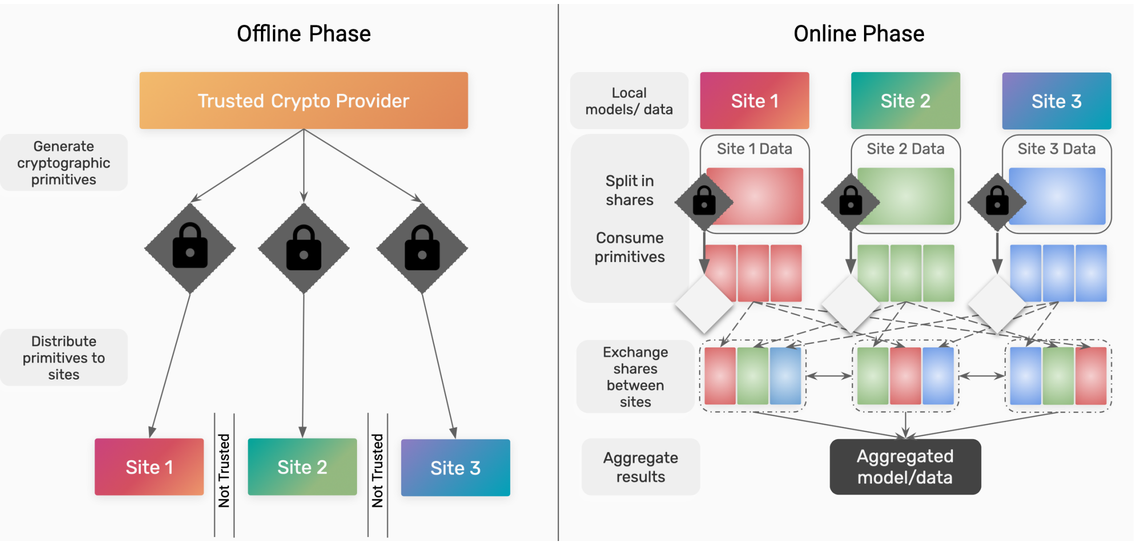


Figure 5: **Simplified conceptual overview of the offline/preprocessing and online phases of a SMPC protocol.** In the offline phase, the cryptographic (*Crypto*) provider generates cryptographic primitives and distributes them to the sites. Sites don't trust each-other but trust the cryptographic provider, also called a *trusted third party*. In the online phase, each site (party) splits their corresponding data or model weights into *shares* and exchanges them with other parties to perform computations. Cryptographic primitives are consumed during this process. The cryptographic provider is not involved in the online phase, which occurs between individual parties. At the end, the processed shares are united to reveal the result.

10 Synchronisation rate hyperparameter

We here describe empirical results of varying the synchronisation rate σ for otherwise identical model hyperparameters. We recall that σ represents the number of batches trained on each node before a(-n) (secure) aggregation pass is performed. Smaller values of σ therefore represent a more frequent synchronisation, expected to yield the following effects:

1. Increased input/output burden on the network. Since the models have to be synchronised more often, network traffic is increased and a net increase in data transmission occurs

2. Increased training time. Since the sending and receiving of models as well as the (secure) aggregation process require non-trivial time, net training time is increased.

3. Finer weight updates. Since fewer batches elapse before synchronisation takes place, lower σ values result in reduced gradient accumulation and thus, an overall less smooth gradient profile.

We expect point (3) to have an effect similar to batch size variation. At a constant batch size (in our case, 200 samples), an increase in σ is thus expected to result in a smoother gradient profile and reduced affinity for loss surface minima, risking reduced convergence. This effect is discussed in [33], notably however, experimentation is only performed on small images (MNIST) and no deep convolutional neural networks are included. To determine the effect of σ on model performance, we performed the following experiment: Under otherwise identical deterministic circumstances and using an identical dataset of 150 randomly selected images per training node on three nodes and the full validation set, we trained 20 ResNet18 models as described in the main manuscript with identical hyperparameters except σ, which was varied between 1 (i.e. synchronisation every batch) and 7. Total training time and validation set *Matthew's Correlation Coefficient* were noted for each of the 20 runs. Results are visualised in Figure 6. Corroborating previous findings ([33]), we conclude that the σ parameter represents both a strong regulariser as well as a central determinant of training time and should be selected judiciously. In the example shown, for instance, an increase of σ from a value of 1 to a value of 2 would have netted a ca. 15% decrease in training time against a ca. 2% decrease in validation set performance.

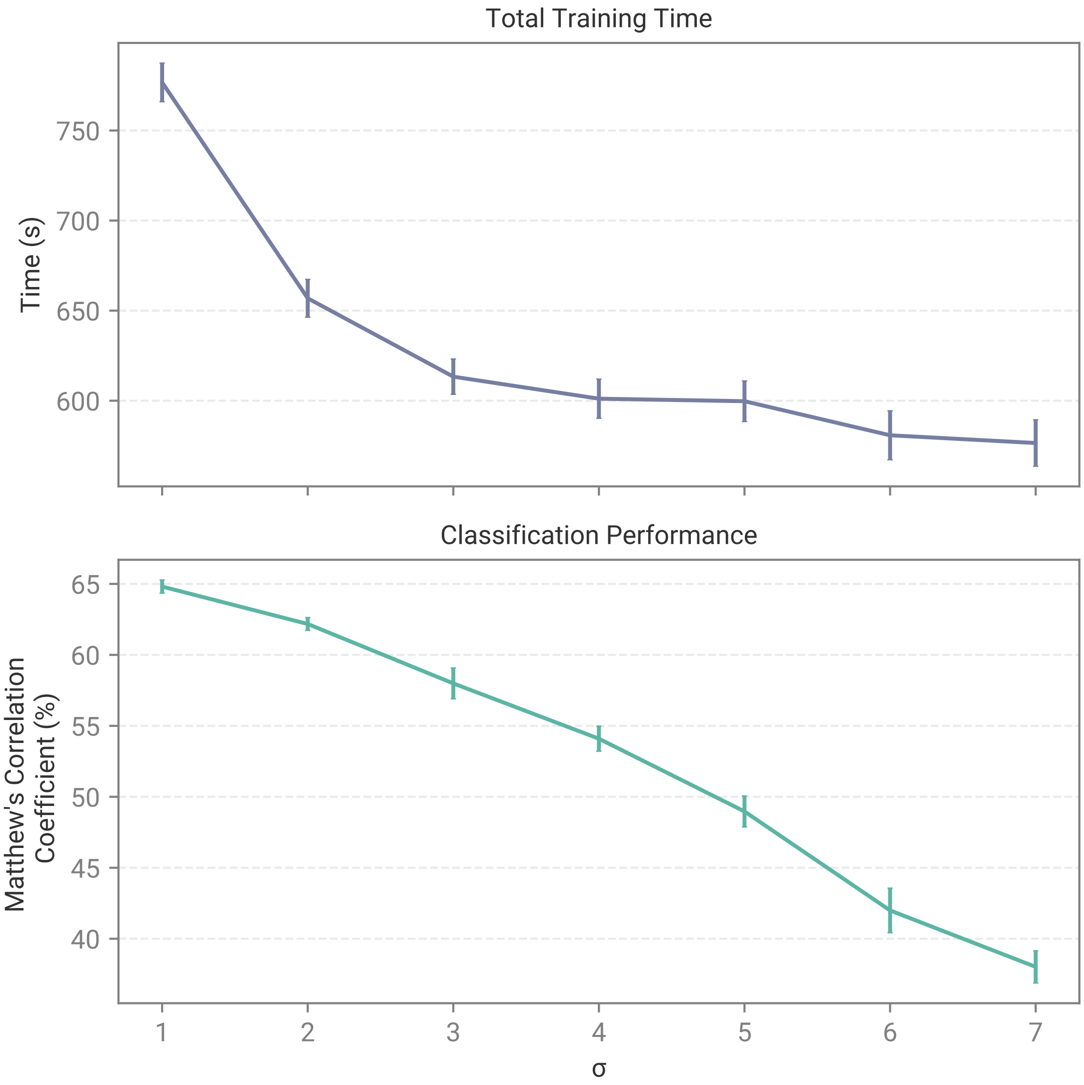


Figure 6: **Influence of the σ parameter on model performance and training times.** The top panel shows the total training time, the bottom panel the classification performance measured with the Matthew's Correlation Coefficient, both as a function of the synchronisation hyperparameter σ.

References

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

[2] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, (Red Hook, NY, USA), p. 2546–2554, Curran Associates Inc., 2011.

[3] A. Inselberg, *Parallel Coordinates - Visual Multidimensional Geometry and Its Applications*. Berlin Heidelberg: Springer Science & Business Media, 2009.

[4] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica: Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[5] I. Mironov, K. Talwar, and L. Zhang, "Rényi differential privacy of the sampled gaussian mechanism," *arXiv preprint arXiv:1908.10530*, 2019.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[7] P. Yakubovskiy, "Segmentation models pytorch." `https://github.com/qubvel/segmentation_models.pytorch`, 2020.

[8] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 240–248, Springer International Publishing, 2017.

[9] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

[10] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016.

[11] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, pp. 1310–1318, 2013.

[12] J. Le Ny and G. J. Pappas, "Differentially private filtering," *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 341–354, 2013.

[13] I. Goodfellow, "Efficient per-example gradient computations," *arXiv preprint arXiv:1510.01799*, 2015.

[14] G. Rochette, A. Manoel, and E. W. Tramel, "Efficient per-example gradient computations in convolutional neural networks," 2019.

[15] P. Subramani, N. Vadivelu, and G. Kamath, "Enabling fast differentially private sgd via just-in-time compilation and vectorization," *arXiv preprint arXiv:2010.09063*, 2020.

[16] H. B. McMahan, G. Andrew, U. Erlingsson, S. Chien, I. Mironov, N. Papernot, and P. Kairouz, "A general approach to adding differential privacy to iterative training procedures," *arXiv preprint arXiv:1812.06210*, 2018.

[17] L. Fan, K. W. Ng, C. Ju, T. Zhang, C. Liu, C. S. Chan, and Q. Yang, "Rethinking privacy preserving deep learning: How to evaluate and thwart privacy attacks," in *Federated Learning*, pp. 32–50, Springer, 2020.

[18] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *Advances in Neural Information Processing Systems*, pp. 6277–6287, 2018.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[20] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[21] A. Davody, D. I. Adelani, T. Kleinbauer, and D. Klakow, "Robust differentially private training of deep neural networks," *arXiv preprint arXiv:2006.10919*, 2020.

[22] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[23] C. Sabater, A. Bellet, and J. Ramon, "Distributed differentially private averaging with improved utility and robustness to malicious parties," 2020.

[24] J. Lee and C. Clifton, "How much is enough? choosing ϵ for differential privacy," in *Information Security* (X. Lai, J. Zhou, and H. Li, eds.), (Berlin, Heidelberg), pp. 325–340, Springer Berlin Heidelberg, 2011.

[25] N. Papernot, M. Abadi, Úlfar Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," 2017.

[26] 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, IEEE, 2017.

[27] D. Fay, J. Sjölund, and T. J. Oechtering, "Decentralized differentially private segmentation with pate," 2020.

[28] I. Damgard, V. Pastro, N. Smart, and S. Zakarias, "Multiparty computation from somewhat homomorphic encryption." Cryptology ePrint Archive, Report 2011/535, 2011. `https://eprint.iacr.org/2011/535`.

[29] E. Boyle, N. Gilboa, and Y. Ishai, "Function secret sharing: Improvements and extensions." Cryptology ePrint Archive, Report 2018/707, 2018. `https://eprint.iacr.org/2018/707`.

[30] T. Ryffel, D. Pointcheval, and F. Bach, "Ariann: Low-interaction privacy-preserving deep learning via function secret sharing," *arXiv preprint arXiv:2006.04593*, 2020.

[31] D. Beaver, "Efficient multiparty protocols using circuit randomization," in *Advances in Cryptology — CRYPTO '91* (J. Feigenbaum, ed.), (Berlin, Heidelberg), pp. 420–432, Springer Berlin Heidelberg, 1992.

[32] O. Goldreich, "Cryptography and cryptographic protocols," *Distributed Computing*, vol. 16, no. 2-3, pp. 177–199, 2003.

[33] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

B Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging - Supplementary Information

Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging - Supplementary Information

Soroosh Tayebi Arasteh[1,+,*], Alexander Ziller[2,3,+,*], Christiane Kuhl[1], Marcus Makowski[2], Sven Nebelung[1], Rickmer Braren[2], Daniel Rueckert[3], Daniel Truhn[1,x,*], and Georgios Kaissis[2,3,4,5,x,*]

[1]Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany.
[2]Institute of Diagnostic and Interventional Radiology, Technical University of Munich, Munich, Germany.
[3]Artificial Intelligence in Healthcare and Medicine, Technical University of Munich, Munich, Germany.
[4]Department of Computing, Imperial College London, London, United Kingdom.
[5]Institute for Machine Learning in Biomedical Imaging, Helmholtz-Zentrum Munich, Neuherberg, Germany.

[*]{sarasteh, dtruhn}@ukaachen.de, {alex.ziller, g.kaissis}@tum.de
[+]These authors contributed equally to this work
[x]These authors jointly supervised this work

Supplementary Note 1: Additional remarks on privacy-utility trade-off

Varying model architectures

In addition to the ResNet9-architecture reported in the main manuscript, we additionally used three more architectures: An EfficientNet B0, with 4 017 796 parameters, adhering to the original implementation proposed by Tan et al. [1], with the sole exception of replacing all batch normalization layers with group normalization; DenseNet121, with 6 962 056 parameters, following the original design put forth by Huang et al. [2], again with the exclusive modification of substituting batch normalization layers with group normalization; and ResNet18, with 11 180 616 parameters, following the original blueprint developed by He et al. [3], with the unique alteration of replacing batch normalization layers with group normalization. All three models displayed a trend consistent with the utility penalties we observed for ResNet9 in both DP and non-DP training. Compare also Supplementary Figure 4.

Further datasets

To prevent domain-specific bias in our results, we employed the Artificial Intelligence for Robust Glaucoma Screening (AIROGS) dataset [4]. This dataset comprises 101 354 RGB ocular fundus images from approximately 60 000 patients of diverse ethnicities, aimed at detecting the presence of referable glaucoma. We allocated 80% of the patients—both with and without glaucoma—to the training set, reserving the remaining 20% for the test set. Image pre-processing involved cropping and other schemes as detailed in [5] and [6]. The images were resized to a dimension of $3 \times 224 \times 224$, with 3 representing the number of channels. We adopted the same EfficientNet B0 network architecture, with identical DP and non-DP training parameters as described earlier, with the same $\delta = 6 \cdot 10^{-6}$. The network was pre-trained on the ImageNet [7] dataset.

Supplementary Figure 10 shows a similar trend as our observations on chest radiographs regarding the privacy-utility trade-off.

Supplementary Figures and Tables

	Training Set		Test Set		All	
	N	percentage	N	percentage	N	percentage
Total	153,502		39,809		193,311	
Female	52,843	(34.42%)	14,449	(36.30%)	67,292	(34.81%)
Male	100,659	(65.58%)	25,360	(63.70%)	126,019	(65.19%)
Aged [0, 30)	4,279	(2.79%)	1,165	(2.93%)	5,444	(2.82%)
Aged [30, 60)	42,340	(27.58%)	10,291	(25.85%)	52,631	(27.23%)
Aged [60, 70)	36,882	(24.03%)	10,025	(25.18%)	46,907	(24.27%)
Aged [70, 80)	48,864	(31.83%)	12,958	(32.55%)	61,822	(31.98%)
Aged [80, 100)	21,137	(13.77%)	5,370	(13.49%)	26,507	(13.71%)
Cardiomegaly	71,732	(46.72%)	18,616	(46.75%)	90,348	(46.74%)
Congestion	13,096	(8.53%)	3,275	(8.22%)	16,371	(8.47%)
Pleural effusion right	12,334	(8.03%)	3,275	(8.22%)	15,609	(8.07%)
Pleural effusion left	9,969	(6.49%)	2,602	(6.53%)	12,571	(6.50%)
Pneumonic infiltration right	17,666	(11.51%)	4,847	(12.17%)	22,513	(11.64%)
Pneumonic infiltration left	12,431	(8.10%)	3,562	(8.94%)	15,993	(8.27%)
Atelectasis right	14,841	(9.67%)	3,920	(9.84%)	18,761	(9.71%)
Atelectasis left	11,916	(7.76%)	3,166	(7.95%)	15,082	(7.80%)
	Age Training Set		Age Test Set		Age All	
	Mean	StD	Mean	StD	Mean	StD
Total	66	15	66	15	66	15
Female	66	15	66	16	66	15
Male	65	14	66	14	65	14
Aged [0, 30)	21	8	21	8	21	8
Aged [30, 60)	50	8	51	8	51	8
Aged [60, 70)	65	3	65	3	65	3
Aged [70, 80)	75	3	75	3	75	3
Aged [80, 100)	84	3	84	3	84	3

Supplementary Table 1: Statistics over subgroups of the UKA-CXR dataset used in this study. The upper part of the table shows the number of samples in each group and their relative share in training and test set, as well as the complete dataset. The lower part shows the mean and standard deviation of the age in the subgroups again over training and test sets as well as the complete dataset.

	AUROC	Accuracy	Specificity	Sensitivity
Cardiomegaly	0.84 ± 0.00	0.75 ± 0.00	0.71 ± 0.02	0.79 ± 0.02
Congestion	0.85 ± 0.00	0.75 ± 0.02	0.75 ± 0.02	0.79 ± 0.02
Pleural Effusion Right	0.94 ± 0.00	0.83 ± 0.01	0.83 ± 0.02	0.91 ± 0.02
Pleural Effusion Left	0.92 ± 0.00	0.83 ± 0.02	0.83 ± 0.02	0.86 ± 0.02
Pneumonic Infiltration Right	0.93 ± 0.00	0.85 ± 0.02	0.85 ± 0.02	0.86 ± 0.02
Pneumonic Infiltration Left	0.94 ± 0.00	0.86 ± 0.01	0.86 ± 0.02	0.87 ± 0.02
Atelectasis Right	0.89 ± 0.00	0.78 ± 0.01	0.78 ± 0.01	0.84 ± 0.02
Atelectasis Left	0.87 ± 0.00	0.78 ± 0.01	0.78 ± 0.02	0.81 ± 0.02
Average	0.90 ± 0.04	0.81 ± 0.04	0.80 ± 0.05	0.84 ± 0.04

Supplementary Table 2: Detailed evaluation results of training without DP. The results show the average and individual area under the receiver-operator-characteristic curve (AUROC), accuracy, specificity, and sensitivity values for each label tested on $N = 39,809$ test images. The training dataset includes $N = 153,502$ images.

	AUROC	Accuracy	Specificity	Sensitivity
Cardiomegaly	0.82 ± 0.00	0.73 ± 0.00	0.71 ± 0.02	0.76 ± 0.02
Congestion	0.81 ± 0.00	0.72 ± 0.02	0.71 ± 0.03	0.76 ± 0.03
Pleural Effusion Right	0.92 ± 0.00	0.82 ± 0.01	0.82 ± 0.01	0.88 ± 0.01
Pleural Effusion Left	0.89 ± 0.00	0.79 ± 0.02	0.79 ± 0.02	0.84 ± 0.02
Pneumonic Infiltration Right	0.91 ± 0.00	0.84 ± 0.01	0.83 ± 0.02	0.81 ± 0.02
Pneumonic Infiltration Left	0.91 ± 0.00	0.84 ± 0.01	0.84 ± 0.01	0.83 ± 0.01
Atelectasis Right	0.87 ± 0.00	0.78 ± 0.01	0.77 ± 0.01	0.81 ± 0.01
Atelectasis Left	0.85 ± 0.00	0.76 ± 0.02	0.76 ± 0.02	0.79 ± 0.02
Average	0.87 ± 0.04	0.79 ± 0.04	0.78 ± 0.05	0.81 ± 0.04

Supplementary Table 3: Detailed evaluation results of DP training with $\varepsilon = 7.89$, $\delta = 6 \cdot 10^{-6}$. The results show the average and individual AUROC, accuracy, specificity, and sensitivity values for each label tested on $N = 39,809$ test images. The training dataset includes $N = 153,502$ images.

	AUROC	Accuracy	Specificity	Sensitivity
Cardiomegaly	0.81 ± 0.00	0.73 ± 0.00	0.70 ± 0.01	0.77 ± 0.01
Congestion	0.81 ± 0.00	0.71 ± 0.02	0.70 ± 0.02	0.77 ± 0.02
Pleural Effusion Right	0.92 ± 0.00	0.82 ± 0.01	0.81 ± 0.01	0.87 ± 0.01
Pleural Effusion Left	0.89 ± 0.00	0.80 ± 0.01	0.80 ± 0.02	0.81 ± 0.02
Pneumonic Infiltration Right	0.90 ± 0.00	0.81 ± 0.01	0.81 ± 0.01	0.82 ± 0.01
Pneumonic Infiltration Left	0.91 ± 0.00	0.82 ± 0.01	0.82 ± 0.01	0.85 ± 0.02
Atelectasis Right	0.86 ± 0.00	0.76 ± 0.01	0.75 ± 0.02	0.83 ± 0.02
Atelectasis Left	0.85 ± 0.00	0.78 ± 0.02	0.78 ± 0.03	0.76 ± 0.03
Average	0.87 ± 0.04	0.78 ± 0.04	0.77 ± 0.05	0.81 ± 0.04

Supplementary Table 4: Detailed evaluation results of DP training with $\varepsilon = 4.71$, $\delta = 6 \cdot 10^{-6}$. The results show the average and individual AUROC, accuracy, specificity, and sensitivity values for each label tested on $N = 39,809$ test images. The training dataset includes $N = 153,502$ images.

	AUROC	Accuracy	Specificity	Sensitivity
Cardiomegaly	0.81 ± 0.00	0.73 ± 0.00	0.68 ± 0.02	0.78 ± 0.02
Congestion	0.80 ± 0.00	0.70 ± 0.02	0.69 ± 0.03	0.76 ± 0.03
Pleural Effusion Right	0.90 ± 0.00	0.80 ± 0.01	0.79 ± 0.01	0.86 ± 0.01
Pleural Effusion Left	0.87 ± 0.00	0.75 ± 0.02	0.74 ± 0.02	0.84 ± 0.02
Pneumonic Infiltration Right	0.90 ± 0.00	0.80 ± 0.01	0.80 ± 0.02	0.83 ± 0.02
Pneumonic Infiltration Left	0.90 ± 0.00	0.83 ± 0.01	0.83 ± 0.02	0.81 ± 0.02
Atelectasis Right	0.85 ± 0.00	0.74 ± 0.02	0.73 ± 0.02	0.82 ± 0.02
Atelectasis Left	0.83 ± 0.00	0.73 ± 0.03	0.73 ± 0.03	0.77 ± 0.03
Average	0.86 ± 0.04	0.76 ± 0.05	0.75 ± 0.05	0.81 ± 0.04

Supplementary Table 5: Detailed evaluation results of DP training with $\varepsilon = 2.04$, $\delta = 6 \cdot 10^{-6}$. The results show the average and individual AUROC, accuracy, specificity, and sensitivity values for each label tested on $N = 39,809$ test images. The training dataset includes $N = 153,502$ images.

	AUROC	Accuracy	Specificity	Sensitivity
Cardiomegaly	0.80 ± 0.00	0.72 ± 0.00	0.69 ± 0.02	0.76 ± 0.02
Congestion	0.80 ± 0.00	0.70 ± 0.02	0.69 ± 0.02	0.75 ± 0.02
Pleural Effusion Right	0.90 ± 0.00	0.80 ± 0.01	0.79 ± 0.02	0.86 ± 0.02
Pleural Effusion Left	0.86 ± 0.00	0.73 ± 0.02	0.72 ± 0.02	0.83 ± 0.02
Pneumonic Infiltration Right	0.89 ± 0.00	0.80 ± 0.02	0.80 ± 0.03	0.81 ± 0.03
Pneumonic Infiltration Left	0.89 ± 0.00	0.79 ± 0.01	0.79 ± 0.02	0.83 ± 0.02
Atelectasis Right	0.84 ± 0.00	0.74 ± 0.02	0.74 ± 0.02	0.80 ± 0.02
Atelectasis Left	0.82 ± 0.00	0.70 ± 0.01	0.69 ± 0.02	0.79 ± 0.02
Average	0.85 ± 0.04	0.75 ± 0.04	0.74 ± 0.05	0.80 ± 0.04

Supplementary Table 6: Detailed evaluation results of DP training with $\varepsilon = 1.06$, $\delta = 6 \cdot 10^{-6}$. The results show the average and individual AUROC, accuracy, specificity, and sensitivity values for each label tested on $N = 39,809$ test images. The training dataset includes $N = 153,502$ images.

	AUROC	Accuracy	Specificity	Sensitivity
Cardiomegaly	0.79 ± 0.00	0.72 ± 0.00	0.69 ± 0.01	0.74 ± 0.01
Congestion	0.79 ± 0.00	0.67 ± 0.02	0.66 ± 0.02	0.78 ± 0.02
Pleural Effusion Right	0.89 ± 0.00	0.77 ± 0.01	0.76 ± 0.02	0.86 ± 0.02
Pleural Effusion Left	0.84 ± 0.00	0.71 ± 0.02	0.70 ± 0.03	0.84 ± 0.03
Pneumonic Infiltration Right	0.88 ± 0.00	0.80 ± 0.01	0.80 ± 0.02	0.79 ± 0.02
Pneumonic Infiltration Left	0.88 ± 0.00	0.77 ± 0.02	0.77 ± 0.03	0.83 ± 0.03
Atelectasis Right	0.83 ± 0.00	0.74 ± 0.01	0.73 ± 0.01	0.79 ± 0.01
Atelectasis Left	0.81 ± 0.00	0.70 ± 0.03	0.70 ± 0.03	0.77 ± 0.03
Average	0.84 ± 0.04	0.73 ± 0.04	0.73 ± 0.05	0.80 ± 0.04

Supplementary Table 7: Detailed evaluation results of DP training with $\varepsilon = 0.54$, $\delta = 6 \cdot 10^{-6}$. The results show the average and individual AUROC, accuracy, specificity, and sensitivity values for each label tested on $N = 39,809$ test images. The training dataset includes $N = 153,502$ images.

	AUROC	Accuracy	Specificity	Sensitivity
Cardiomegaly	0.79 ± 0.00	0.71 ± 0.00	0.67 ± 0.01	0.75 ± 0.01
Congestion	0.78 ± 0.00	0.68 ± 0.02	0.68 ± 0.02	0.74 ± 0.02
Pleural Effusion Right	0.88 ± 0.00	0.77 ± 0.01	0.77 ± 0.02	0.83 ± 0.02
Pleural Effusion Left	0.84 ± 0.00	0.73 ± 0.01	0.72 ± 0.02	0.80 ± 0.02
Pneumonic Infiltration Right	0.87 ± 0.00	0.79 ± 0.01	0.79 ± 0.02	0.79 ± 0.02
Pneumonic Infiltration Left	0.88 ± 0.00	0.79 ± 0.01	0.79 ± 0.01	0.81 ± 0.01
Atelectasis Right	0.82 ± 0.00	0.73 ± 0.02	0.73 ± 0.02	0.77 ± 0.02
Atelectasis Left	0.80 ± 0.00	0.71 ± 0.02	0.71 ± 0.02	0.75 ± 0.02
Average	0.83 ± 0.04	0.74 ± 0.04	0.73 ± 0.05	0.78 ± 0.04

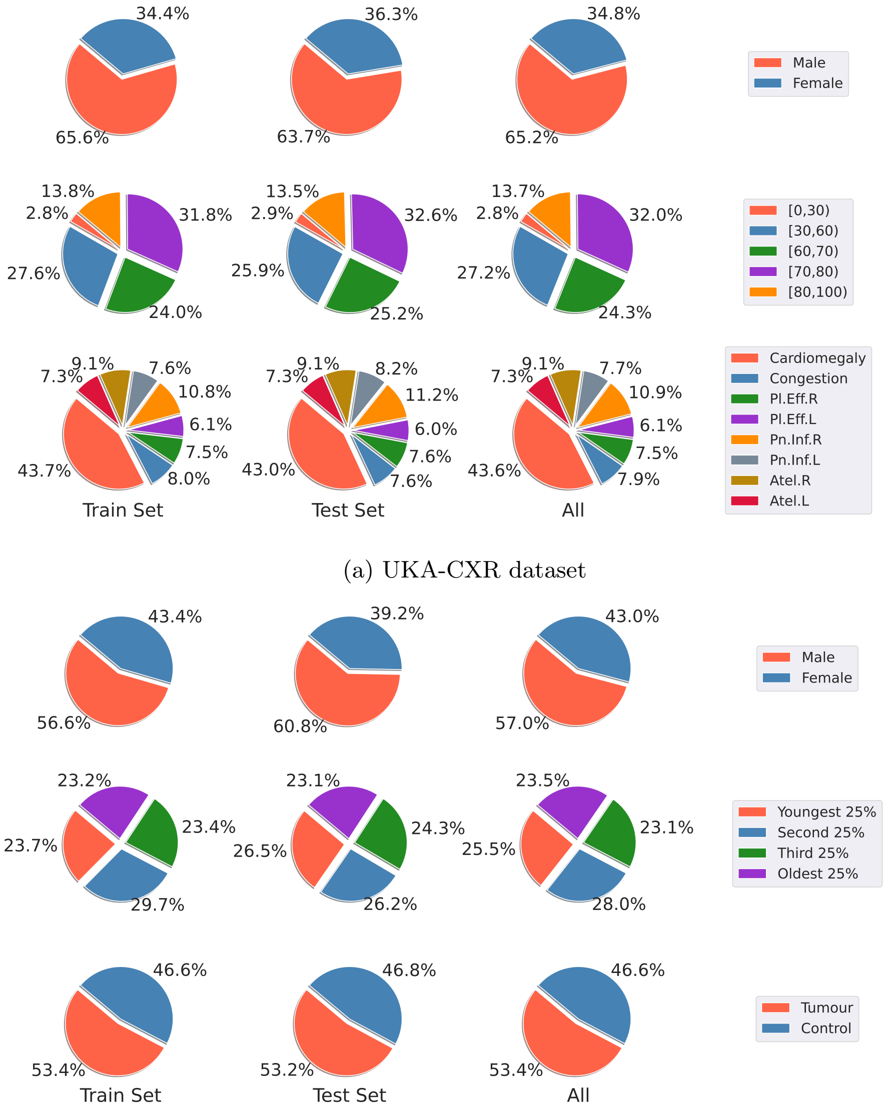
Supplementary Table 8: Detailed evaluation results of DP training with $\varepsilon = 0.29$, $\delta = 6 \cdot 10^{-6}$. The results show the average and individual AUROC, accuracy, specificity, and sensitivity values for each label tested on $N = 39,809$ test images. The training dataset includes $N = 153,502$ images.

PDAC														
	Total		Male		Female		Youngest 25%		Second 25%		Third 25%		Oldest 25%	
ε	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
0.29	24.86	10.7	23.86	9.6	25.54	14.0	20.29	23.9	15.97	7.9	32.10	8.8	27.78	12.0
0.54	11.37	3.2	11.23	3.4	10.82	4.2	8.70	8.7	4.86	2.4	19.14	7.0	10.42	2.1
1.06	5.97	1.7	5.96	1.6	6.06	2.0	2.90	2.5	1.39	2.4	11.11	3.7	6.25	2.1
2.04	2.70	0.9	2.46	0.6	3.03	1.5	1.45	2.5	1.39	1.2	3.09	1.1	4.17	3.6
4.71	1.73	1.0	1.40	0.6	2.16	1.5	1.45	2.5	0.69	1.2	1.85	0.0	2.78	1.2
5.0	2.31	2.0	1.75	1.2	3.03	3.0	1.45	2.5	1.39	2.4	2.47	1.1	3.47	2.4
6.0	3.08	2.3	2.46	2.2	3.90	2.6	1.45	2.5	2.08	2.1	3.70	3.2	4.17	2.1
7.0	1.54	1.2	1.40	1.6	1.73	1.5	0.00	0.0	0.69	1.2	2.47	2.8	2.08	2.1
8.0	0.58	0.6	0.00	0.0	1.30	1.3	0.00	0.0	1.39	2.4	0.00	0.0	0.69	1.2
Non-private	0.77	0.7	0.00	0.0	1.73	1.5	0.00	0.0	2.08	2.1	0.62	1.1	0.00	0.0

Supplementary Table 9: Underdiagnosis rates of subgroups. Underdiagnosis rate is the false positive rate of non-tumor cases. μ denotes the mean underdiagnosis rate for a certain subgroup, while σ denotes the standard deviation.

	Tumor		Control		PtD	
N Test	173		152			
ε	μ	σ	μ	σ	μ	σ
0.29	75.14	10.7	85.09	2.3	−9.94	13.0
0.54	88.63	3.2	86.40	2.5	2.23	5.4
1.06	94.03	1.7	85.53	3.5	8.50	4.7
2.04	97.30	0.9	87.94	1.0	9.36	0.4
4.71	98.27	1.0	90.57	1.9	7.70	2.9
5.0	97.69	2.0	91.01	2.1	6.68	4.1
6.0	96.92	2.3	91.89	1.7	5.03	4.0
7.0	98.46	1.2	90.79	1.7	7.67	2.8
8.0	99.42	0.6	95.39	3.7	4.03	3.5
∞	99.23	0.7	97.81	1.5	1.42	1.3

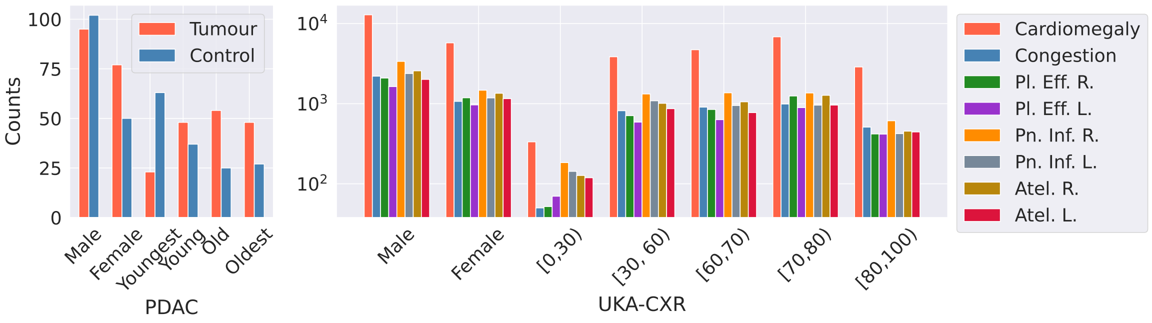
Supplementary Table 10: Per Diagnosis Accuracy on the PDAC dataset. PtD is the statistical parity difference between the tumor and control group. μ denotes the mean, σ the standard deviation over three runs.



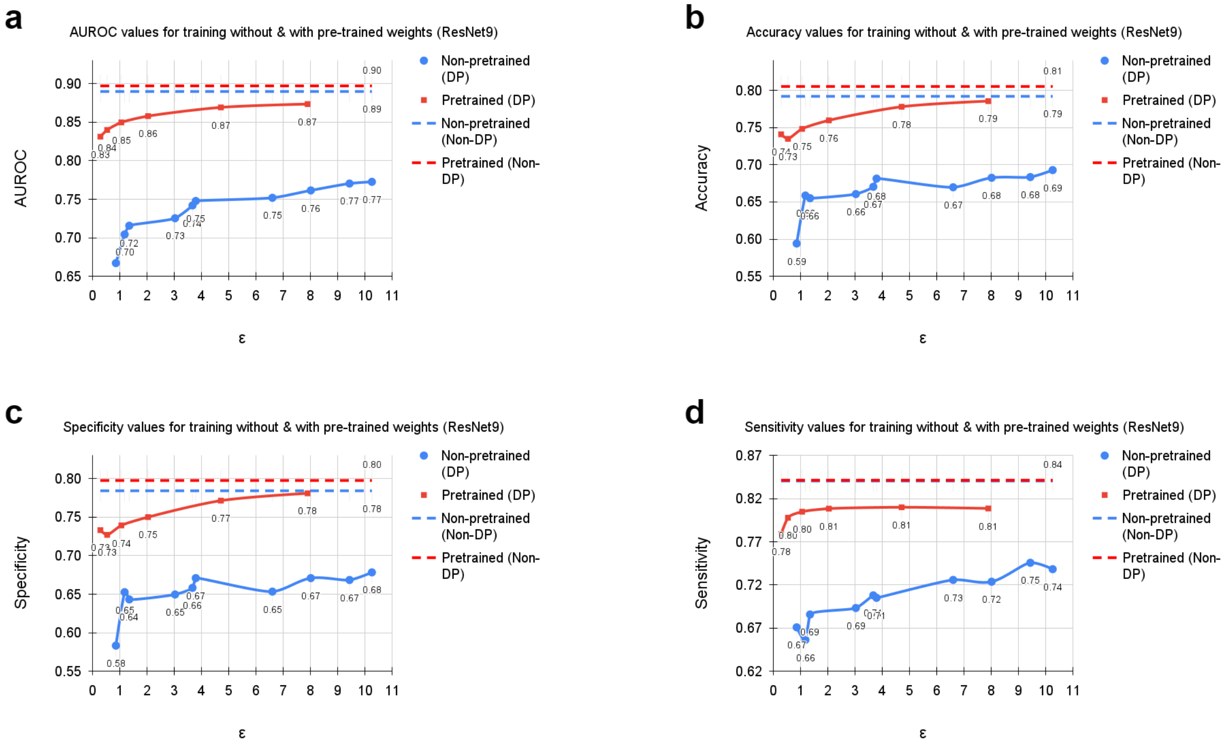
(a) UKA-CXR dataset

(b) PDAC dataset

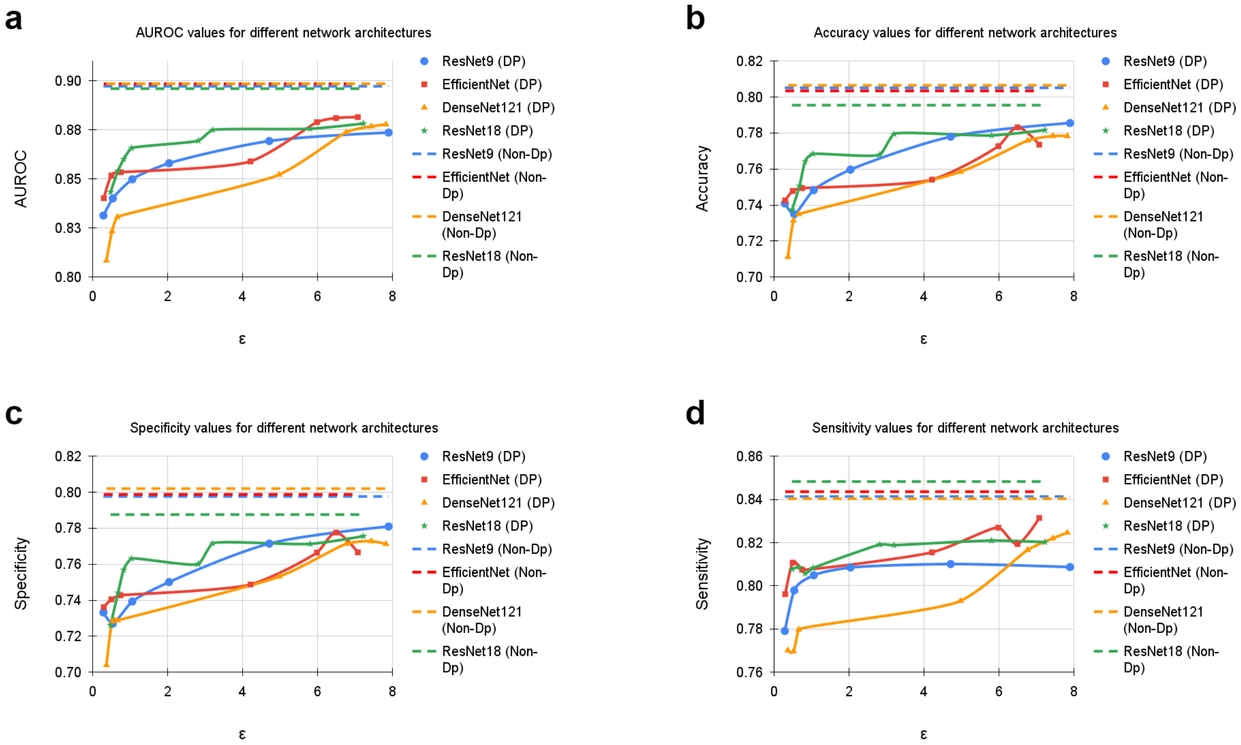
Supplementary Figure 1: Visual overview of the distribution over subgroups



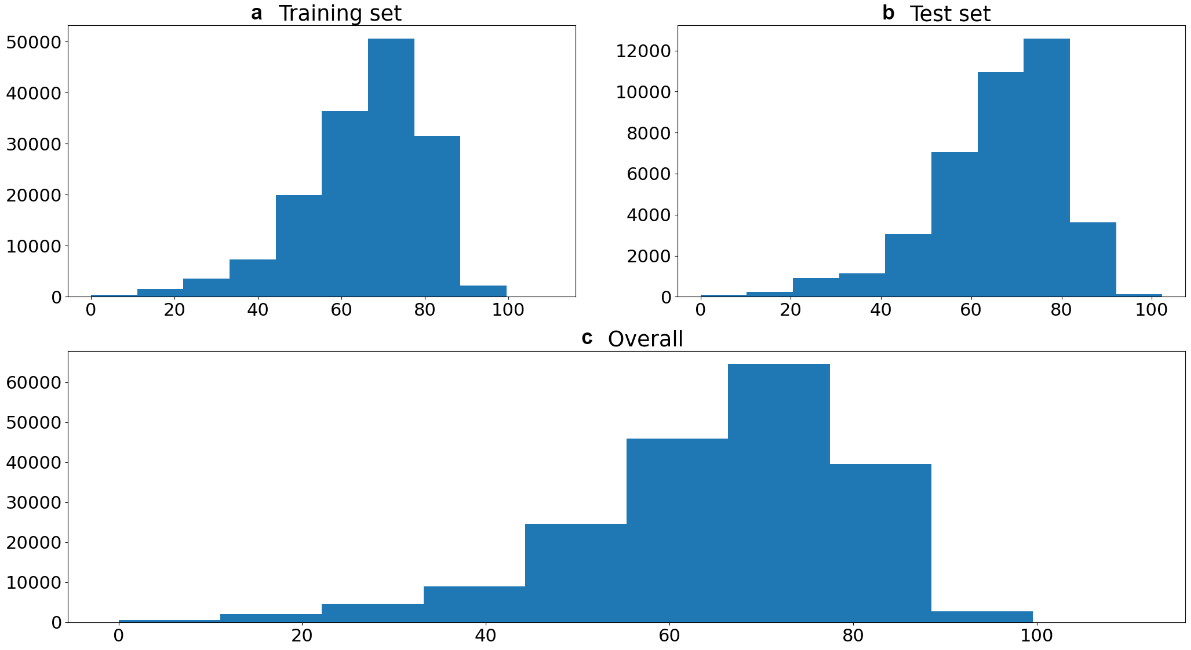
Supplementary Figure 2: Distribution of labels within subgroups



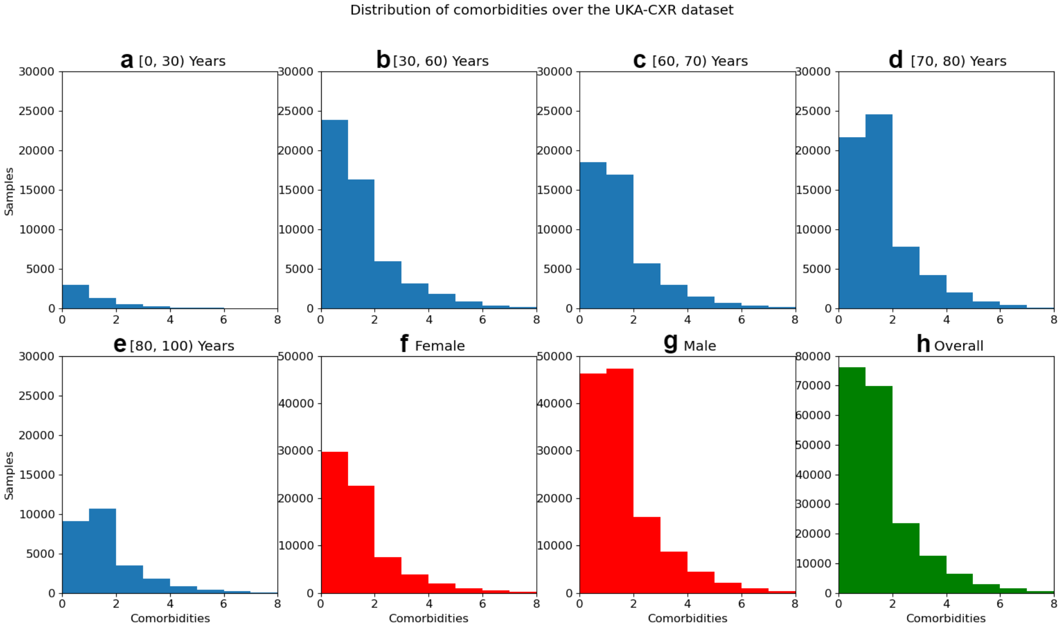
Supplementary Figure 3: Average results of DP training with different ε values for $\delta = 6 \cdot 10^{-6}$ using pre-trained weights versus training from scratch. The curves show the average **a** AUROC, **b** accuracy, **c** specificity, and **d** sensitivity values over all labels, including cardiomegaly, congestion, pleural effusion right, pleural effusion left, pneumonic infiltration right, pneumonic infiltration left, atelectasis right, and atelectasis left tested on $N = 39\,809$ test images. The training dataset includes $N = 153\,502$ images. Note, that the AUROC is monotonically increasing, while sensitivity, specificity, and accuracy exhibit more variation. This is due to the fact that all training processes were optimized for the AUROC. Dashed lines correspond to the non-private training results depicted as upper bounds. The pre-training was done using the MIMIC-CXR dataset with $N = 210\,652$ images.



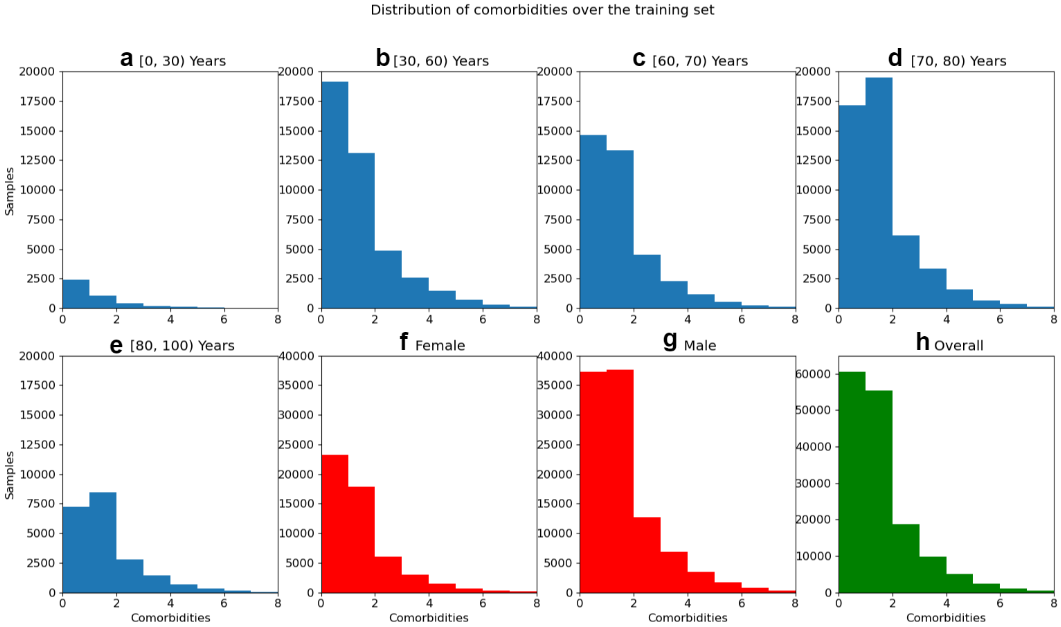
Supplementary Figure 4: Average results of training with DP with different ε values for $\delta = 6 \cdot 10^{-6}$ using different network architectures. The curves show the average **a** AUROC, **b** accuracy, **c** specificity, and **d** sensitivity values over all labels, including cardiomegaly, congestion, pleural effusion right, pleural effusion left, pneumonic infiltration right, pneumonic infiltration left, atelectasis right, and atelectasis left tested on $N = 39\,809$ test images. The training dataset includes $N = 153\,502$ images. Note, that the AUROC is monotonically increasing, while sensitivity, specificity, and accuracy exhibit more variation. This is due to the fact that all training processes were optimized for the AUROC. Dashed lines correspond to the non-private training results depicted as upper bounds.



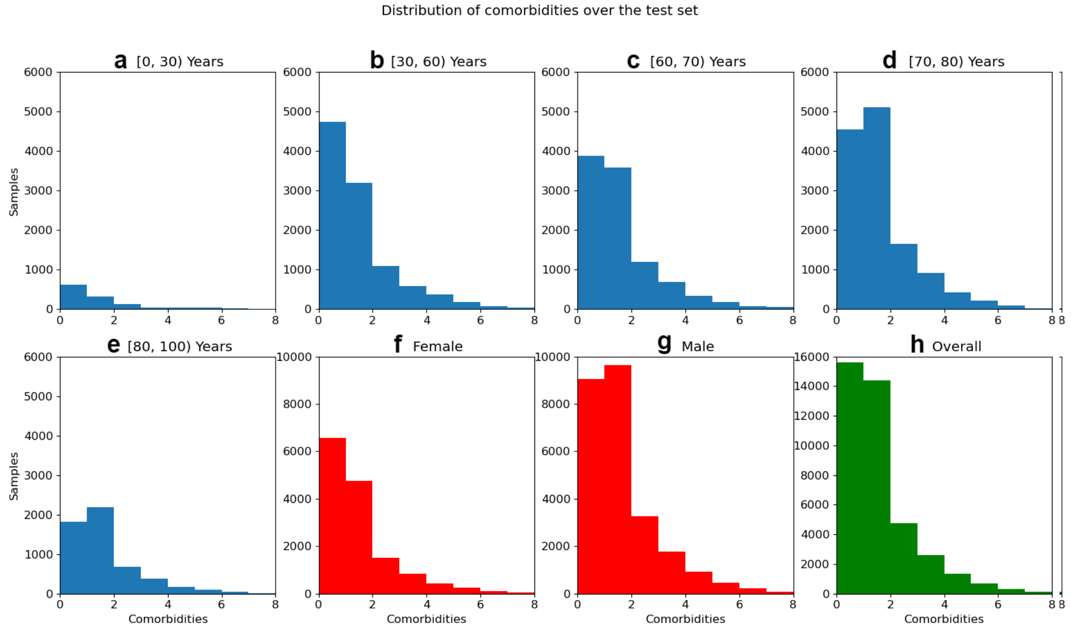
Supplementary Figure 5: Age histogram of the UKA-CXR dataset. **a** Training set. **b** Test set. **c** Overall.



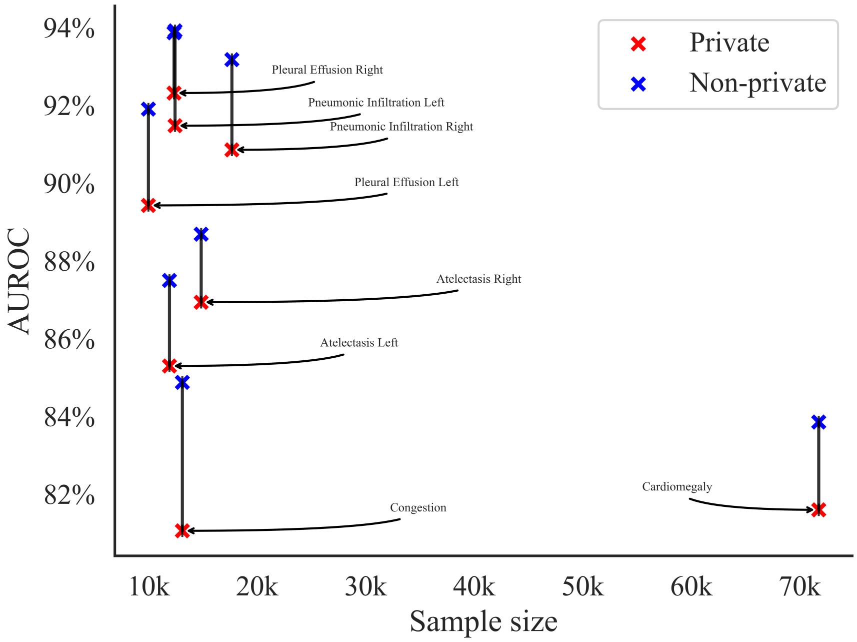
Supplementary Figure 6: Distribution of comorbidities over the UKA-CXR dataset. Histograms of comorbidities are given for different subsets of the dataset including subjects aging in the range of **a** $[0, 30)$ years old with a mean of 0.8 ± 1.2 comorbidities, **b** $[30, 60)$ years old with a mean of 1.0 ± 1.3 comorbidities, **c** $[60, 70)$ years old with a mean of 1.1 ± 1.3 comorbidities, **d** $[70, 80)$ years old with a mean of 1.1 ± 1.2 comorbidities, **e** $[80, 100)$ years old with a mean of 1.1 ± 1.3 comorbidities, as well as **f** females with a mean of 1.0 ± 1.2 comorbidities, **g** males with a mean of 1.1 ± 1.3 comorbidities, and **h** overall with a mean of 1.1 ± 1.3 comorbidities.



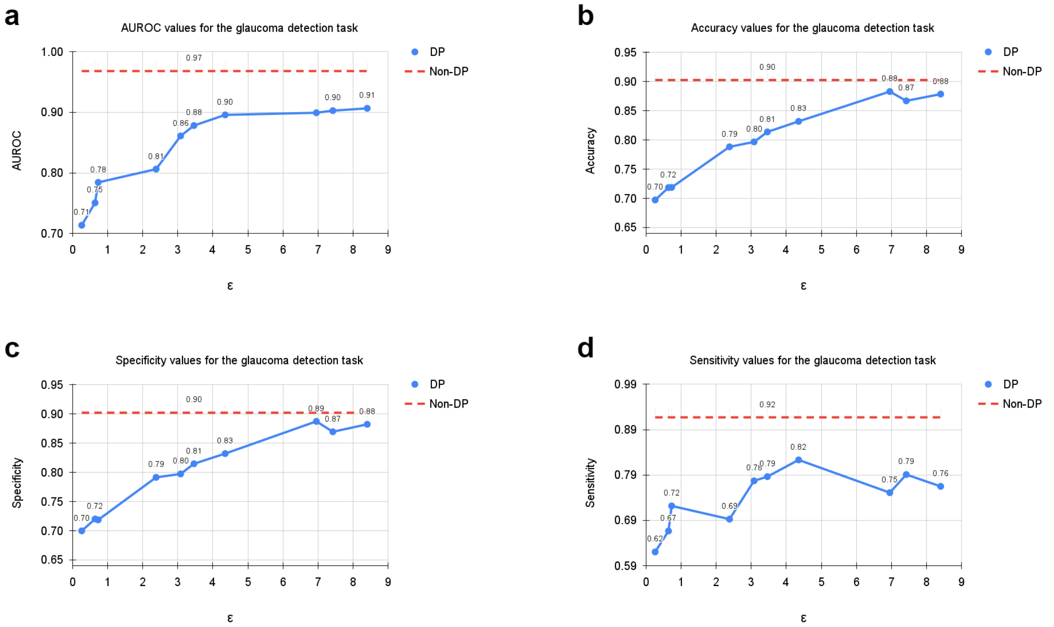
Supplementary Figure 7: Distribution of comorbidities over the training set. Histograms of comorbidities are given for different subsets of the training set including subjects aging in the range of **a** $[0, 30)$ years old with a mean of 0.8 ± 1.2 comorbidities, **b** $[30, 60)$ years old with a mean of 1.0 ± 1.3 comorbidities, **c** $[60, 70)$ years old with a mean of 1.1 ± 1.3 comorbidities, **d** $[70, 80)$ years old with a mean of 1.1 ± 1.2 comorbidities, **e** $[80, 100)$ years old with a mean of 1.1 ± 1.3 comorbidities, as well as **f** females with a mean of 1.0 ± 1.2 comorbidities, **g** males with a mean of 1.1 ± 1.3 comorbidities, and **h** overall training set with a mean of 1.1 ± 1.3 comorbidities.



Supplementary Figure 8: Distribution of comorbidities over the test set. Histograms of comorbidities are given for different subsets of the test set including subjects aging in the range of **a** $[0, 30)$ years old with a mean of 0.9 ± 1.4 comorbidities, **b** $[30, 60)$ years old with a mean of 1.0 ± 1.3 comorbidities, **c** $[60, 70)$ years old with a mean of 1.1 ± 1.3 comorbidities, **d** $[70, 80)$ years old with a mean of 1.1 ± 1.2 comorbidities, **e** $[80, 100)$ years old with a mean of 1.1 ± 1.3 comorbidities, as well as **f** females with a mean of 1.0 ± 1.3 comorbidities, **g** males with a mean of 1.1 ± 1.3 comorbidities, and **h** overall test set with a mean of 1.1 ± 1.3 comorbidities.



Supplementary Figure 9: Relation of sample size to training performance for private and performance loss compared to non private training. Each dot marks the performance on the test set on one diagnosis of the private model at $\varepsilon = 7.89$. Colors indicate the performance loss compared to the non private model.



Supplementary Figure 10: Evaluation results of the Glaucoma detection task [4] for training with DP with different ε values for $\delta = 6 \cdot 10^{-6}$. The curves show the **a** AUROC, **b** accuracy, **c** specificity, and **d** sensitivity values tested on $N = 20\,268$ test images. The training dataset includes $N = 81\,086$ images. Note, that the AUROC is monotonically increasing, while sensitivity, specificity, and accuracy exhibit more variation. This is due to the fact that all training processes were optimized for the AUROC. Dashed lines correspond to the non-private training results depicted as upper bounds.

Supplementary References

[1] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114, 2019.

[2] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, Adrian Galdran, Miguel Ángel González Ballester, Gustavo Carneiro, Devika R G, Hrishikesh P S, Densen Puthussery, Hong Liu, Zekang Yang, Satoshi Kondo, Satoshi Kasai, Edward Wang, Ashritha Durvasula, Jónathan Heras, Miguel Ángel Zapata, Teresa Araújo, Guilherme Aresta, Hrvoje Bogunović, Mustafa Arikan, Yeong Chan Lee, Hyun Bin Cho, Yoon Ho Choi, Abdul Qayyum, Imran Razzak, Bram van Ginneken, Hans G. Lemij, and Clara I. Sánchez. Airogs: Artificial intelligence for robust glaucoma screening challenge. *arXiv preprint arXiv:2302.01738*, 2023.

[5] Firas Khader, Christoph Haarburger, Jörg-Christian Kirr, Marcel Menke, Jakob Nikolas Kather, Johannes Stegmaier, Christiane Kuhl, Sven Nebelung, and Daniel Truhn. Elevating fundoscopic evaluation to expert level - automatic glaucoma detection using data from the airogs challenge. In *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*, pages 1–4, 2022.

[6] Ahmed Al-Mahrooqi, Dmitrii Medvedev, Rand Muhtaseb, and Mohammad Yaqub. Gardnet: Robust multi-view network for glaucoma classification in color fundus images. In Bhavna Antony, Huazhu Fu, Cecilia S. Lee, Tom MacGillivray, Yanwu Xu, and Yalin Zheng, editors, *Ophthalmic Medical Image Analysis*, pages 152–161, Cham, 2022. Springer International Publishing.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

C Reconciling Privacy and Accuracy in AI for Medical Imaging - Supplementary Information

Reconciling Privacy and Accuracy in AI for Medical Imaging - Supplementary Material

A Threat models

Here, we provide a more concrete explanation of how the risks of re-identification through a data reconstruction attack are moderated by Differential Privacy (DP). Concretely, we avail ourselves of the framework of reconstruction robustness (ReRo), which will allow us to formulate an upper bound on the success rate of data reconstruction attacks against Artificial Intelligence (AI) models trained with DP under the specific threat models discussed below.

ReRo was introduced by [1]. It is a guarantee pertaining to an algorithm which processes sensitive data, e.g. an AI model trained with DP. Intuitively, if at most a proportion $0 \leq \gamma \leq 1$ of the total samples used to train the model can be successfully reconstructed by an adversary with a reconstruction error lower than $\eta \geq 0$, then the model satisfies (η, γ)-ReRo. Recent works have proven that all models trained with DP automatically satisfy ReRo and that for certain settings, it is possible to directly quantify the upper bound for γ [1, 2, 3]. In other words, DP automatically provides strong and quantifiable protection against data reconstruction attacks.

We study the ReRo guarantees of models trained with DP under three distinct sets of assumptions about the capabilities of the adversary, i.e. three distinct threat models:

1. The **worst-case** threat model: This corresponds to the adversary usually considered in DP, namely one who has unbounded computational abilities, can deeply manipulate the model's (hyper-)parameters and has access to the target image itself, which they can use to attack the model. Evidently, this threat model is not realistic (as an adversary who has access to the target point would not need to attack the model), but is used to provide guarantees when "all bets are off", i.e. in a so-called *privacy auditing* scenario when one is interested in the absolute worst-case behaviour of a system.

2. The **relaxed** threat model [4]: This threat model is still quite pessimistic, as it still assumes unbounded computational ability and access to model (hyper-)parameters. However, this adversary only has restricted access to the dataset, notably, they cannot use the target image itself to attack the model. Although it renders this threat model more appropriate for scenarios where the dataset can be safely assumed to be kept secure, e.g. in a hospital's database, it still makes assumptions, which are not encountered in any practical scenario. Most importantly, the adversary has a black-box reconstruction algorithm, which yields either a perfect reconstruction or fails, and the only decision the adversary has to make is whether the reconstruction was indeed the target data. The term *relaxed* stems from security research, where a *relaxation* denotes a weakening of a security assumption (as can be seen in e.g. [1, 2, 5, 6, 7, 8]).

3. The **realistic** threat model: The final threat model considers an adversary with unbounded computational ability and the power to manipulate model (hyper-)parameters but only very limited access to information about the dataset. For example, the adversary can know the dimensions of the images to be reconstructed but not any of their contents. We note that even this threat model is relatively pessimistic, as it assumes an active adversary who is trusted and therefore the actions are not reviewed by other participants. Such adversaries could manipulate the model to their advantage in order to reconstruct training data. We term this threat model as realistic as it stems from federated learning research, although in many cases, this could be detected simply by inspecting the model architecture. Nonetheless, we use this threat model as it is conceivable that such adversaries can exist in, e.g. federated learning settings in untrustworthy consortia,

which are common in real-world settings. Protection against this threat model generalizes to all weaker threat models, such as black-box attacks after training and adversaries lacking the ability to manipulate the deep learning model or its hyperparameters.

Of note, it is possible to provide theoretical bounds on the reconstruction attack success rate in both the worst-case and the relaxed threat models using the techniques presented in [2, 3]. For the realistic threat model, we assess the attack success rate empirically. A concise overview of the aforementioned threat models is provided in Table 1.

In summary, while conservative (i.e. worst-case or relaxed) threat models are important tools in security research because they allow one to derive closed-form bounds on the attack success rate of very powerful adversaries, such threat models are in most reasonable scenarios too pessimistic.

B Setup

B.1 Dataset Description

Here we outline our rationale for choosing specific datasets for our experiments. We identified four characteristics of medical imaging datasets, which reoccur frequently: (1) Datasets are often **small** compared to non-medical datasets. For example, most medical AI algorithms, which are currently approved by the US Food and Drug Administration (FDA), are trained on less than 1 000 data samples [9]. (2) Diagnoses occur with very different frequencies, leading to often **imbalanced** datasets skewed toward more common diagnoses. In segmentation tasks, this may happen due to different spatial extensions of objects. (3) While natural images are all captured with standard cameras as RGB images, medical images are from **multi-modal** imaging devices such as computed tomography (CT), magnetic resonance imaging (MRI), or ultrasound.

In this study, we aim to give a broad discussion of settings in medical AI. Hence, we have chosen three datasets, which encompass the above-discussed scenarios (c.f. Table 2).

1. The RadImageNet dataset [10] contains over 1.3 million 2D images with CT, MRI, and ultrasound scans representing three imaging modalities with 165 classification targets, which are highly imbalanced.

2. The HAM10000 dataset [11] is a collection of 10 000 skin lesion RGB images spread across seven categories. We intentionally amplified the class imbalance to a strong but not untypical 80 : 20 class ratio by merging classes based on the need for immediate treatment (see Section 4.1).

3. Lastly, we use the MSD Liver dataset [12, 13], a demanding image-to-image task involving just 131 CT scans annotated at voxel level. Given the small number of available training samples as well as a segmentation task (i.e., per-pixel classification) with tumours only encompassing a tiny fraction of each scan, it represents a very challenging medically relevant task.

To the best of our knowledge, no prior work shows the performance of AI models trained under formal privacy guarantees on such a comprehensive and large dataset as RadImageNet or a 3D image-to-image task as MSD Liver represents.

B.2 Metrics

To measure the performance of the models on classification tasks, we use Matthews' Correlation Coefficient (MCC) [14]. Opposed to more frequently used metrics such as accuracy or F_1-score, it incorporates the entire confusion matrix and, by that, is extremely robust against any class imbalance [15]. It is also better interpretable as for random predictions it is 0 and for perfect predictions 1, whereas the accuracy depends on the class distribution. For the segmentation task, we measure the class-wise Dice score of the 3D volumes and report the average over all volumes for the liver and tumour, as they are the targets of interest in our task. A perfect prediction yields a 100% Dice score.

References

[1] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1138–1156. IEEE, 2022.

[2] Jamie Hayes, Saeed Mahloujifar, and Borja Balle. Bounding training data reconstruction in dp-sgd. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[3] Georgios Kaissis, Jamie Hayes, Alexander Ziller, and Daniel Rueckert. Bounding data reconstruction attacks with the hypothesis testing interpretation of differential privacy. *Theory and Practice of Differential Privacy*, 2023.

[4] Georgios Kaissis, Alexander Ziller, Stefan Kolek, Anneliese Riess, and Daniel Rueckert. Optimal privacy guarantees for a relaxed threat model: Addressing sub-optimal adversaries in differentially private machine learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[5] Christina Brzuska, Marc Fischlin, Nigel P Smart, Bogdan Warinschi, and Stephen C Williams. Less is more: Relaxed yet composable security notions for key exchange. *International Journal of Information Security*, 12:267–297, 2013.

[6] Boaz Barak, Ran Canetti, Jesper Buus Nielsen, and Rafael Pass. Universally composable protocols with relaxed set-up assumptions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 186–195. IEEE, 2004.

[7] Ran Canetti, Hugo Krawczyk, and Jesper B Nielsen. Relaxing chosen-ciphertext security. In *Advances in Cryptology-CRYPTO 2003: 23rd Annual International Cryptology Conference, Santa Barbara, California, USA, August 17-21, 2003. Proceedings 23*, pages 565–582. Springer, 2003.

[8] Peng Li and Steve Zdancewic. Downgrading policies and relaxed noninterference. In *Proceedings of the 32nd ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 158–170, 2005.

[9] U.S. Food and Drug Administration. Artificial intelligence and machine learning (ai/ml)-enabled medical devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

[10] Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 0(ja):e210315, 0.

[11] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[12] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[13] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[14] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[15] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

D Licences

RightsLink

Sign in/Register

SPRINGER NATURE

End-to-end privacy preserving deep learning on multi-institutional medical imaging

Author: Georgios Kaissis et al

Publication: Nature Machine Intelligence

Publisher: Springer Nature

Date: May 24, 2021

Author Request

If you are the author of this content (or his/her designated agent) please read the following. If you are not the author of this content, please click the Back button and select no to the question "Are you the Author of this Springer Nature content?".
Ownership of copyright in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights:
To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).
The author and any academic institution, where they work, at the time may reproduce the contribution for the purpose of course teaching.
To reuse figures or tables created by the Author and contained in the Contribution in oral presentations and other works created by them.
To post a copy of the contribution as accepted for publication after peer review (in locked Word processing file, of a PDF version thereof) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the contribution on the publisher's website.
Authors wishing to use the published version of their article for promotional use or on a web site must request in the normal way.

If you require further assistance please read Springer Nature's online author reuse guidelines.

For full paper portion: Authors of original research papers published by Springer Nature are encouraged to submit the author's version of the accepted, peer-reviewed manuscript to their relevant funding body's archive, for release six months after publication. In addition, authors are encouraged to archive their version of the manuscript in their institution's repositories (as well as their personal Web sites), also six months after original publication.

v1.0

BACK CLOSE WINDOW

Privacy - Terms

CCC RightsLink

SPRINGER NATURE

Medical imaging deep learning with differential privacy

Author: Alexander Ziller et al

Publication: Scientific Reports

Publisher: Springer Nature

Date: Jun 29, 2021

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.
To request permission for a type of use not listed, please contact Springer Nature

RightsLink

SPRINGER NATURE

Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging

Author: Soroosh Tayebi Arasteh et al

Publication: Communications Medicine

Publisher: Springer Nature

Date: Mar 14, 2024

Creative Commons

You are not required to obtain permission to reuse this article.
To request permission for a type of use not listed, please contact Springer Nature

RightsLink

SPRINGER NATURE

Reconciling privacy and accuracy in AI for medical imaging

Author: Alexander Ziller et al

Publication: Nature Machine Intelligence

Publisher: Springer Nature

Date: Jun 21, 2024

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.
To request permission for a type of use not listed, please contact Springer Nature