

Towards Standardized Evaluation in Differentially Private Image Classification: A Critical Approach

by

Sara Kodeiri

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2024

© Sara Kodeiri 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis is based on collaborative research I co-authored with Sabrina Mokhtari, Shubhankar Mohapatra, Gautam Kamath, and Florian Tramèr. While the ideas and overall research direction were a collective effort, the content presented in this thesis represents my original writing.

The foundational concepts and introductory framework, initially drafted by Gautam Kamath for our joint paper, have been adapted and expanded upon in this thesis. My primary contributions to the research include the design and execution of experiments involving Wide-ResNets [ZMH21]. Sabrina Mokhtari was responsible for the ScatterNet [TB21] experiments, while Shubhankar Mohapatra conducted the CLIP [RKH⁺21] experiments. The ablation study presented in Section 5.4 was a collaborative effort between Mohapatra and Mokhtari. Mokhtari also led the experiment findings.

Abstract

This thesis critically examines differentially private machine learning (DPML) in image classification, addressing recent critiques about the effectiveness of current techniques and the validity of existing benchmarks. We focus on three key questions: the accuracy of current benchmarks in measuring DPML progress, the impact of public pre-training datasets on DPML model performance, and strategies for unifying future research efforts.

Our study introduces standardized benchmark datasets and evaluation settings using two medical image datasets as private data sources. We assess various DPML methods across different scenarios, including those with no public data, training from scratch, and fine-tuning approaches.

The main contributions include: proposing standardized benchmark datasets and evaluation settings, conducting a validation study of previous DPML techniques, and introducing a moderated public leaderboard to track progress in DPML. This research aims to provide a comprehensive assessment of DPML in image classification, offering insights into existing methods and suggesting future research directions in machine learning and privacy.

Acknowledgements

I would like to thank my advisor, Gautam Kamath, for the opportunity of a lifetime. None of this would have been possible without your insights and guidance, and I am forever grateful to have had the chance to work with you.

I am thankful to Florian Tramèr for his collaborative efforts and his technical insights.

Sabrina and Shubhankar, even though I have credited you as co-authors, I would like to acknowledge you as friends. I had the best time learning alongside and from you.

I would also like to thank Florian Kerschbaum and Pascal Poupart for making time to read and for serving on the thesis committee.

I extend my sincere gratitude to the administrative and technical staff members at the University of Waterloo: Kimberley Beckman, Tom Caudro, Lucas Gamez, Carly McLeod, Lori Paniak, and Denise Shantz. Thank you for your patience with my emails!

Last but not least, I would like to thank my family for their unwavering support from oceans away.

If I could, I would acknowledge every person I have met in the past two years. It takes a village to finish a research-based degree, and I truly had the best one.

Dedication

To my parents, Elham and Keivan. Only we know what it took to get here.

Table of Contents

Author’s Declaration	ii
Statement of Contributions	iii
Abstract	iv
Acknowledgements	v
Dedication	vi
List of Figures	x
List of Tables	xi
1 Introduction	1
1.1 The Story	1
1.2 Thesis Outline	3
2 Related Work	4
2.1 Differential Privacy	4
2.2 Differentially Private Machine Learning	4
2.3 Current DPML Research	5

3	Preliminaries	7
3.1	Classification	7
3.2	Differential Privacy	8
3.3	DP-SGD	10
3.3.1	Accounting Methods	12
4	Benchmark Design	14
4.1	Private Datasets	15
4.2	Public Datasets	16
4.2.1	“Anything goes” zero-shot	17
4.3	Privacy Settings	17
4.4	Leaderboard	18
5	Experiments and Findings	20
5.1	Experiment Settings	20
5.2	Models	21
5.3	Main Findings	23
5.3.1	Improvements from pre-training are not uniform across datasets and depend on the private data.	23
5.3.2	Increased privacy budget enhances pre-training benefits	24
5.3.3	CIFAR-10 success is not a proxy for benchmark progress	26
5.4	Revisiting Ablations	26
6	Conclusions and Future Work	31
6.1	Concluding Remarks	31
6.2	Future Work	32
	References	34

APPENDICES	46
A All Experiment Results	47

List of Figures

5.1	Improvements from pre-training depend on the private data: Public data consistently benefits EyePACS, while ViT-G/14 without full fine-tuning yields the best results on CheXpert.	24
5.2	Pre-training public data is more beneficial with higher ε values. For CheXpert, ScatterNet performs better at smaller ε values, while pretrained models show marginal improvements at larger ε values. Similarly, for EyePACS, CLIP ViT-G/14 linear performs better as ε value increases.	25
5.3	Normalization generally improves the final performance of all models. The yellow bars correspond to groups of 8 and 9, while the gray bars correspond to groups of 16 and 27 for CLIP-ViTs and ScatterNet, respectively. (The choice of 27 over 81 is due to its superior performance.)	27
5.4	Evaluating augmentation multiplicity by adding 4 and 8 augmentations of each image in the training data. Augmentation multiplicity helps in general for CheXpert but not for EyePACS.	28

List of Tables

A.1	Test AUC evaluation for EyePACS at different privacy regimes. Baselines include ScatterNet (SN), WideResNet (WRN) and CLIP models on datasets with different public data pre-training. The SOTA is due to [VMB19].	47
A.2	Test AUC evaluation for CheXpert at different privacy regimes. Baselines include ScatterNet (SN), WideResNet (WRN) and CLIP models on datasets with different public data pre-training. The SOTA is due to [BDS ⁺ 23](Private) and [YYSY21](Non-Private).	48
A.5	Studying the impact of augmentation multiplicity, we find that it consistently improves performance for CIFAR-10. However, looking at EyePACS and CheXpert, we observe inconsistent behavior, except that it generally seems to reduce performance with EyePACS. "Best so far" refers to the baseline plus all previously effective ablations listed in Table A.6	48
A.3	Studying the impact of normalization for CLIP ViT models, we find that normalization generally improves performance, but the effectiveness depends on the architecture and dataset. Normalizations marked in red indicate a drop in performance compared to the baseline.	49
A.4	Studying the impact of normalization for ScatterNet + CNN, we observe that normalization consistently improves performance. Data normalization tends to outperform group normalization for EyePACS and CIFAR-10, particularly due to the large gradients of their Scatter features.	49
A.6	Studying all ablations together, we observe an almost consistent improvement in performance for CIFAR-10, whereas this pattern is not observed with the other datasets.	50

Chapter 1

Introduction

1.1 The Story

In recent years, deep learning and artificial intelligence have become ubiquitous, impacting various aspects of society to such an extent that they have evolved into prevalent buzzwords across industries and academia alike. The efficacy of these AI systems and algorithms is heavily data-dependent, requiring vast amounts of information to train models and perform a wide range of tasks accurately. However, this reliance on data poses significant challenges when dealing with privacy-sensitive information, e.g. in fields such as healthcare, finance, and personal communications [POG⁺24].

The challenge of utilizing data while protecting individual privacy predates the rise of artificial intelligence. As a solution, Differential Privacy (DP) [DMNS06] provides a standard framework for quantifying how much privacy has been spent on a specific task and guarantees an upper bound. In the context of learning theory, a model is considered privately trained if it meets a specific privacy criterion: Even if an adversary has full access to the trained model and external statistical information, they should not be able to determine whether any particular data point was used in the model’s training process. In other words, it guarantees that the model will not heavily rely on any individual’s sensitive data. We formally define DP in Chapter 2.

Privately training models has entered various subdomains of machine learning and deep learning within the past decade. Despite these advancements, the field is not without its critics. Specifically, researchers Florian Tramèr, Gautam Kamath, and Nicholas Carlini

have raised concerns about certain developments in the DP ML landscape in a thought-provoking position paper [TKC24]. Their critique primarily challenges whether using large web-scraped datasets can be considered truly privacy-preserving under differential privacy. The authors also question the overall utility of the current DP paradigm. Most related to this work, they suggest that existing benchmarks of privacy-preserving learning techniques may not be *truly* measuring progress in the field, and question how *meaningful* the current benchmarks are. We thoroughly examine this paper in Chapter 2.

While this issue affects all areas of differentially private machine learning (DPML), and Tramèr, Kamath, and Carlini’s critique encompasses the entire field, we focus on privacy-preserving image classification. Taking inspiration from their work, in this thesis, we aim to answer the following questions:

1. To what extent are current benchmarks in differentially private machine learning accurately measuring progress in the field, particularly in DP image classification?
2. How does using public pre-training datasets impact the performance and validity of differentially private machine learning models?
3. How can we unify future research efforts to address the inconsistencies in result reporting, especially regarding disparities in public pre-training datasets, and better isolate truly effective techniques in privacy-critical settings?

To address our research questions, we designed a series of experiments across various settings to evaluate the DPML methods. Our study focuses on two medical image datasets as private data sources: EyePACS [Kag] and CheXpert [IRK⁺19]. To establish our benchmark, we drew upon existing literature on state-of-the-art privacy methods and non-private image classification tasks to select appropriate models for evaluation. Our choices aim to cover a range of scenarios in privacy-preserving machine learning. We opted for ScatterNets [TB21] for when where no public data is available, wide residual networks (Wide-ResNets) [ZK16] for training from scratch and full fine-tuning, and CLIP-based models [RKH⁺21] for full fine-tuning approaches. These selections allow us to produce baseline numbers across diverse model architectures and training paradigms, providing a comprehensive starting point for assessing differentially private machine learning techniques. To facilitate a broader comparison and evaluation of methods, we also introduce standardized convex settings by linear probe experiments. A comprehensive discussion of the experiments’ details and the rationale behind our choices will be presented in Chapter 5.

The three main contributions are listed below:

1. Two proposed standardized benchmark datasets and their evaluation settings, focusing on healthcare as a privacy-sensitive domain.
2. A validation study of previous techniques for differentially private image classification across various settings, utilizing the newly proposed benchmarks.
3. A moderated public leaderboard for differentially private machine learning to track progress and recent work on a unified platform.

1.2 Thesis Outline

We begin in Chapter 2 by examining previous work in the field, with a focus on DPML methods on vision data. Chapter 3 lays the groundwork by introducing and defining relevant concepts essential for understanding our research. The core of our work is presented in Chapter 4, where we detail the design of our benchmark, including its objectives, dataset selection, privacy settings, and the reasoning behind our choices. Chapter 5 focuses on the experimental process, presenting our findings and offering a thorough evaluation analysis. The thesis concludes in Chapter 6, where we summarize our key contributions and discuss potential directions for future research in this domain.

Chapter 2

Related Work

2.1 Differential Privacy

When dealing with data, a fundamental question arises: How can we analyze or learn from sensitive data while preserving individual privacy? Dwork et al. [DMNS06] introduced ε -differential privacy as a rigorous answer to this question. The framework provides a formal, mathematical approach for quantifying privacy loss and offers provable privacy guarantees for individuals within a dataset. Differential privacy (DP) specifically addresses the challenge of preventing information leakage about individuals in a database, rather than protecting the database as a whole. It is crucial to note that DP does not entirely prevent information leakage; instead, it allows for the extraction of aggregate statistics or patterns from the data while limiting the amount of information that can be inferred about any specific individual from these aggregates. A formal definition of differential privacy is provided in Section 3.2.

2.2 Differentially Private Machine Learning

Building upon the foundational concepts of differential privacy, Abadi et al. [ACG⁺16a] introduced Differentially Private Stochastic Gradient Descent (DP-SGD) as a practical approach to apply privacy guarantees to machine learning models, particularly deep neural networks. DP-SGD modifies the standard stochastic gradient descent algorithm by adding

carefully calibrated noise to the gradients during the training process. This noise injection mechanism ensures that the resulting model satisfies differential privacy guarantees, limiting the model’s dependence on any individual data point and protecting against unintended memorization of training data. However, the application of DP-SGD introduces a trade-off between privacy and model utility; as the privacy guarantees become stronger (i.e., lower ϵ values), the model’s performance may decrease. The noise added to gradients significantly disrupts optimization, often leading to much worse performance compared to non-private training methods. [DFAP21, KZR⁺22, KSC⁺22]. Some researchers believe that highly over-parameterized models, which typically excel in non-private settings, struggle with DP-SGD due to the increased norm of the noise in higher-dimensional gradients [KSC⁺22, SWSS21, TB21, YZCL21a]. We will go into theoretical details in Section 3.3

Since its introduction, DP-SGD has inspired numerous variations and improvements aimed at enhancing this privacy-utility trade-off, making differentially private machine learning more practical for real-world applications. In healthcare, works such as [TFR22, AKC⁺22] have integrated differential privacy into medical data analysis, effectively safeguarding the privacy of original medical data and mitigating malicious attacks on sensitive information. Government agencies have also adopted these techniques; for instance, the United States Census Bureau incorporates calibrated perturbation into government data to reduce privacy leakage risks in statistics [CDMS22, BS22]. The tech industry has embraced differential privacy as well, with major companies such as Google and Facebook incorporating it into deep learning frameworks like TensorFlow and PyTorch [PGM⁺19b, YSS⁺22]. In the financial sector, differential privacy has been crucial in enabling cross-industry data sharing for more effective fraud detection while protecting sensitive financial information [YQAH22]. More recently, the application of differential privacy has expanded to large language models [AGG⁺21, YNB⁺22, LTLH22], image classifiers [ACG⁺16a, TB21, PTS⁺21, DBH⁺22], and other generative models [XLW⁺18, BWW⁺19, CBV⁺21, BKZ23, DCVK23, HJSP23]. For a thorough report on DP ML, see [POG⁺24].

2.3 Current DPML Research

Rather than the applications of DPML, our work focuses on the methodologies used to bridge the gap between privacy and utility in image classification and focus on work that

try to bring the private utility number closest to their non-private equivalent. Many works have approached this problem by reducing the dimensionality of the model during training [GAW⁺22, YZCL21b, ZMH21, ZWB21]. On the other hand, several studies [PTS⁺21, TB21] have rethought the architectures used for private training altogether. In particular, Tramèr and Boneh introduce a way of feature extraction on image data that does not need any learning, referred to as “handcrafted features”. After numerous experiments, they conclude that models trained with privacy constraints require “about an order of magnitude more private training data” to surpass the performance of the baselines that use their handcrafted features, and that needing such an amount of data challenges in domains where vast amounts of data are not readily available, such as many medical applications.

In [DBH⁺22], De and Berrada argue that the same over-parameterized architectures used in non-private settings can achieve state-of-the-art performance in private contexts when coupled with a set of novel techniques. To support their claim, they conduct comprehensive experiments implementing several modifications: using larger batch sizes [BFGT20, AGG⁺21], replacing batch normalization layers with group normalization [WH18], incorporating weight standardization in convolutional layers [QWL⁺19], augmenting private data [FBP⁺22], and applying parameter-averaging techniques [PJ92]. Furthermore, they demonstrate that pre-training on public data followed by fine-tuning with DP-SGD on private data significantly enhances model performance. While their experiments utilized ImageNet [DDS⁺09] as public data and CIFAR-10 [Kri09] as private data - a combination we believe differs qualitatively from typical private downstream tasks in differential privacy machine learning (DPML) - the approach has nonetheless achieved state-of-the-art results in private machine learning. Consequently, we evaluate their architectures and techniques using our proposed benchmarks to further validate their effectiveness in more representative DPML scenarios. It is worth noting that De and Berrada address this criticism in a follow-up paper [BDS⁺23]. They expand their approach by utilizing two medical datasets as private data, testing their methods in a more relevant context for DPML. Additionally, they modify their baseline models from the previous study. For pretraining, they use ImageNet-21K [RBBNZM21], which is not accessible enough to be used in benchmark design.

Chapter 3

Preliminaries

In this chapter, we formally define the concepts of classification, differential privacy and differentially private stochastic gradient descent (DP-SGD).

3.1 Classification

Machine learning, at its core, is about finding a function that maps inputs to outputs based on example input-output pairs [Mit97]. This process is formalized through the concept of loss functions and optimization. A loss function $L(y, f(x))$ measures the discrepancy between the true output y and the predicted output $f(x)$ for an input x . The goal is to find a function f that minimizes the expected loss over the entire data distribution [FHT01]: $\mathbb{E}[L(y, f(x))]$. Common loss functions include Mean Squared Error for regression: $L(y, f(x)) = (y - f(x))^2$, and Cross-Entropy for classification: $L(y, f(x)) = -y \log(f(x)) - (1 - y) \log(1 - f(x))$ [GBC16].

Optimization is the process of finding the function $\mathcal{L}$ that minimizes this expected loss. Complex networks typically have loss functions that are non-convex and challenging to optimize. To address this, practitioners often employ the mini-batch stochastic gradient descent (SGD) [Bot10] algorithm. This iterative method works by randomly selecting a small random subset (batch $\mathcal{B}$) of training examples at each step. It then computes $\mathbf{g}_B = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \nabla_{\theta} \mathcal{L}(\theta, x)$ as an estimation to the gradient $\nabla_{\theta} \mathcal{L}(\theta)$. This process gradually guides the model parameters θ following the gradient direction $-\mathbf{g}_B$ towards a local minimum of

the loss function with η as the learning rate, $\theta_{t+1} = \theta_t - \eta \nabla \mathbf{g}_B$ effectively optimizing the network.

Image classification is the task of assigning a label or category to an input image. In the context of machine learning, we treat an image as a 3D tensor $I \in \mathbb{R}^{H \times W \times C}$, where H is height, W is width, and C is the number of color channels [KSH12]. The goal is to learn a function $f : \mathbb{R}^{H \times W \times C} \rightarrow Y$, where Y is the set of possible labels.

The network is trained end-to-end using backpropagation to minimize a chosen loss function [RHW86].

The two different downstream tasks used in this paper are multi-class classification and multi-label classification. In the former, each image belongs to exactly one of K mutually exclusive classes. The output layer has K neurons with a softmax activation. We use categorical cross-entropy loss [DDS⁺09]: $L = -\sum_i y_i \log(f(x)_i)$ where y is a one-hot encoded vector of true labels. In multi-label classification, each image can belong to multiple classes simultaneously. The output layer has K neurons with sigmoid activations. We use binary cross-entropy loss for each label [ZZ13]: $L = -\sum_i [y_i \log(f(x)_i) + (1 - y_i) \log(1 - f(x)_i)]$ where y is a binary vector indicating the presence/absence of each label. The key difference lies in how we model the problem: multi-class assumes one true class per image, enforced by the softmax function, while multi-label allows for multiple true labels per image, treating each label as an independent binary classification problem [TK07]. This difference affects not only the model architecture and loss function but also how we evaluate performance (e.g., accuracy for multi-class vs. precision/recall for multi-label) and how we interpret the model’s outputs [Sor10].

3.2 Differential Privacy

Differential privacy (DP) [DMNS06, DR14] is a mathematical framework that provides a provable and quantifiable guarantee of privacy when the results of an algorithm run on private data are made public [GLW21]. Its formal definition is as follows:

Definition 1 (Differential Privacy) *A randomized algorithm $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ϵ, δ) -differential privacy (DP) if for any two adjacent inputs $D, D' \in \mathcal{D}$ that differ in an entry and for any subset of outputs $\mathcal{S} \subseteq \mathcal{R}$ it holds that :*

$$\Pr[\mathcal{M}(D) \in \mathcal{S}] \leq e^\epsilon \Pr[\mathcal{M}(D') \in \mathcal{S}] + \delta$$

Pure differential privacy, denoted as ϵ -DP, provides a strict privacy guarantee with $\delta = 0$. Approximate differential privacy, denoted as (ϵ, δ) -DP, relaxes this guarantee slightly by allowing a small probability of violation, where δ is typically a very small positive value ($\delta > 0$). Some of the most popular mechanisms in DP include the Laplace mechanism [DMNS06], the Gaussian mechanism [DR14] and the Exponential mechanism [MT07]. These mechanisms are foundational in the DP literature and are widely implemented in both academic research and practical applications. DP also has many useful properties like *post-processing* and *composition* which we define below:

Lemma 3.2.1 (Post-Processing [DMNS06]) *If a mechanism $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ is (ϵ, δ) -differentially private, then for any function $f : \mathcal{R} \rightarrow \mathcal{R}'$, we have that $f \circ \mathcal{M}$ is also (ϵ, δ) -differentially private.*

Lemma 3.2.2 (Basic Composition [DMNS06]) *For each $i \in [k]$, let algorithm $\mathcal{M}_i : \mathcal{D} \times \prod_{j=1}^{i-1} \mathcal{R}_j \rightarrow \mathcal{R}_i$ be (ϵ_i, δ_i) -DP in its first argument. If algorithm $\mathcal{M}_{[k]} : \mathcal{D} \times \prod_{j=1}^k \mathcal{R}_j$ is defined such that :*

$$\mathcal{M}_{[k]} = (\mathcal{M}_1(\mathcal{D}), \mathcal{M}_2(\mathcal{D}, \mathcal{M}_1(\mathcal{D})), \dots, \mathcal{M}_k(\mathcal{D}, \mathcal{M}_1(\mathcal{D})), \dots, \mathcal{M}_{k-1}(\mathcal{D}))$$

then $\mathcal{M}_{[k]}$ is $(\sum_{i=1}^k \epsilon_i, \sum_{i=1}^k \delta_i)$ -differentially private.

Although this composition is useful in many scenarios, the bound it provides weakens when numerous consecutive privacy mechanisms are applied. This limitation motivates the development of advanced composition techniques:

Lemma 3.2.3 (Advanced Composition [KOV17, DRV10]) *For each $i \in [k]$, let algorithm $\mathcal{M}_i : \mathcal{D} \times \prod_{j=1}^{i-1} \mathcal{R}_j \rightarrow \mathcal{R}_i$ be (ϵ_i, δ_i) -DP in its first argument. If algorithm $\mathcal{M}_{[k]} : \mathcal{D} \times \prod_{j=1}^k \mathcal{R}_j$ is defined such that :*

$$\mathcal{M}_{[k]} = (\mathcal{M}_1(\mathcal{D}), \mathcal{M}_2(\mathcal{D}, \mathcal{M}_1(\mathcal{D})), \dots, \mathcal{M}_k(\mathcal{D}, \mathcal{M}_1(\mathcal{D})), \dots, \mathcal{M}_{k-1}(\mathcal{D}))$$

then for every $\delta' > 0$, algorithm $\mathcal{M}_{[k]}$ is $(\varepsilon', 1 - (1 - \delta') \prod_{i=1}^k (1 - \delta_i))$ - DP, where

$$\varepsilon' = \min \left(\sum_{i=1}^k \varepsilon_i, \sum_{i=1}^k \frac{\varepsilon_i (e^{\varepsilon_i} - 1)}{e^{\varepsilon_i} + 1} + \sqrt{\min \left(\sum_{i=1}^k 2\varepsilon_i^2 \log \left(e + \frac{\sqrt{\sum_{i=1}^k 2\varepsilon_i^2}}{\delta'} \right), \sum_{i=1}^k 2\varepsilon_i^2 \log \frac{1}{\delta'} \right)} \right)$$

Both of these properties are powerful tools in proving privacy guarantees for complex algorithms. Essentially, if we ensure that each part of a larger system is differentially private, we can be confident that the system as a whole is also private.

3.3 DP-SGD

There are different approaches to safeguarding privacy when using data to train machine learning models. One potential method involves treating the training process as a black box and focusing solely on the final model parameters. However, this approach has limitations, because it tends to add too much noise and completely jeopardizes the model's utility [ACG⁺16b]. Recognizing these shortcomings, Abadi et al. propose a more nuanced strategy. Their method aims to control the influence of training data throughout the learning process, with a particular focus on stochastic gradient descent (SGD) computation. This approach offers a more balanced way to protect privacy while maintaining the model's utility.

We briefly explained the basis of stochastic gradient descent (SGD) in section 3.1. Now we will do a deeper analysis of its privatized version, differentially private SGD (DP-SGD) [ACG⁺16b]. As seen in Algorithm 1, comparing to non-private SGD, there are two extra steps: per-example gradient clipping and Gaussian noise addition. While gradient clipping might be seen in traditional SGD as a form of regularization, it is not a requirement and the average gradient of the *entire minibatch* is typically clipped. In contrast, in DP-SGD, the gradient for each *individual training example* is clipped before averaging. The addition of Gaussian noise to the clipped gradients further ensures privacy by masking individual contributions to model updates. Another difference to note is subsampling. Randomly

Algorithm 1 DP-SGD Outline

Input: Training set $\{x_1, \dots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N} \sum_i \mathcal{L}(\theta, x_i)$.

Parameters: learning rate η_t , noise scale σ , group size L , gradient norm bound C .

- 1: Initialize θ_0 randomly
- 2: **for** $t \in [T]$ **do**
- 3: Randomly sample L_t from training set with sampling probability L/N
- 4: **Compute gradient**
- 5: **for** each $i \in L_t$ **do**
- 6: compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$
- 7: **end for**
- 8: **Clip gradient**
- 9: $\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$
- 10: **Add noise**
- 11: $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L} (\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}))$
- 12: **Descent**
- 13: $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$
- 14: **end for**

Output: θ_T and the overall privacy cost (ε, δ)

subsampling the minibatch guarantees differential privacy in DP-SGD. Because of this, it is crucial that the uniformity in the sampling step is of cryptographic strength [PT21]. This is not necessary in regular SGD, and subsampling is just a way to reduce training time [RCR16] or again as a form of regularization.

It is important to note that Algorithm 1 treats all parameters as a single input θ , holding for one layer in a neural network. For multi-layer neural networks which are the more common cases, each layer is treated separately, allowing for layer-specific clipping thresholds C and noise scales σ . This is where the composition property of DP (discussed in lemma 3.1) is used to prove the overall privacy guarantee of a neural network. Also, the privacy parameters can be adjusted dynamically if needed based on the number of training steps, providing a flexible balance between model performance and privacy protection.

3.3.1 Accounting Methods

A critical aspect of differential privacy is *measuring* the privacy loss for a specific DP algorithm using a particular noise distribution. The way of measurement provides a formal proof of privacy by establishing upper bounds on ϵ and δ . Before the DP-SGD paper [ACG⁺16a], the best overall bound was derived from the strong composition theorem. Using the composability of DP, an "accountant" method could keep track of the privacy cost at each access to the sensitive training data. However, Abadi et al. found this bound to be loose as it failed to account for the specific noise distribution used. To address this limitation, they introduced the **moments accountant** method.

First, we need to define the privacy loss random variable (PRV), and the privacy curve: Consider running an algorithm $\mathcal{M}$ on a pair of databases x, y . For an outcome o , the privacy loss on o is the log-ratio of its probability when $\mathcal{M}$ is run on each database:

$$\mathcal{L}_{(\mathcal{M}(x) \parallel \mathcal{M}(y))}^{(o)} = \ln \frac{\Pr[\mathcal{M}(x) = o]}{\Pr[\mathcal{M}(y) = o]} \quad (3.1)$$

Definition 2 (Privacy Curve) A DP algorithm M is said to have privacy curve $\delta : \mathbb{R} \rightarrow [0, 1]$, if for every $\epsilon \in \mathbb{R}$, M is $(\epsilon, \delta(\epsilon))$ -DP.

The Rényi-DP-based [Mir17] moments accountant method tracks the upper bounds of the privacy random variable's *moments*. Abadi et al. demonstrate that this approach not only provides an upper bound on privacy cost by monitoring (ϵ, δ) and applying composition theorems to Gaussian mechanisms, but it also extends to Gaussian mechanisms with random sampling. Importantly, this method offers a substantially tighter estimate of privacy loss compared to conventional techniques. They define privacy loss as a random variable dependent on the random noise added in algorithm 1. More specifically, according to the PRV definition, a mechanism being (ϵ, δ) -differentially private is equivalent to its PRV having a certain tail bound. While the tail bound is very useful information on a distribution, composing directly from it can result in quite loose bounds. Their innovation is computing the log moments of the privacy loss random variable, which composes linearly. Analyzing the privacy loss of DP-SGD involves composing the privacy curve of each iteration of training k times, where k is the total number of training iterations [GLW21]. These accounting methods have a computational advantage in that their execution time is independent of k . However, they only provide an upper bound on privacy loss rather than

an exact measure. As a result, they cannot achieve exact precision in approximating the privacy curve.

Shortly after the moments accountant paper, [MV16] showed that if $M_1, M_2, \dots, M_k$ are DP algorithms such that M_i is $(\varepsilon_i, \delta_i)$ -DP, computing the *exact* DP guarantees for $M = M_1 \circ M_2 \circ \dots \circ M_k$ is #P-complete. Regardless, researchers did not stop developing algorithms that provide tighter bounds on the privacy cost, or reduced run-times. An important example is the work of Gopi et al. [GLW21] in which they introduce the **PRV accountant** method. This new algorithm provides a tighter and faster analysis for computing the privacy curve of the composition of many DP algorithms, including DP-SGD. It achieves this by employing *adaptive* composition techniques. The main approach involves approximating PRVs through truncation and discretization. Notably, this approach significantly enhances performance, reducing both computational time and memory usage by a factor of k compared to previous state-of-the-art methods.

Chapter 4

Benchmark Design

Designing a benchmark in the context of machine learning and privacy research involves creating a comprehensive, standardized framework for evaluating and comparing different methods or algorithms. As [TKC24] puts it, the main purpose is to measure progress on a particular task of interest. This section outlines the components of our proposed benchmarks, including the selection of datasets, parameters, and best practices. It is worth emphasizing that we have deliberately not curated any new datasets for this purpose. Instead, by utilizing existing datasets, we ensure that our proposed benchmark remains comparable to non-private settings, facilitating a more direct evaluation of the impact of privacy-preserving techniques on model performance.

We emphasize again that the goal of this work, is not pushing the state-of-the-art numbers on the chosen datasets or methods. In designing this benchmark, our primary goal was to cover a wide range of settings and provide the research community with baseline numbers required to improve techniques used in private image classification. There is more to private machine learning than merely achieving the highest performing numbers, for example ethical, legal, safety or computational considerations, which makes this line of work more challenging. Importantly, there is no hierarchy among the introduced approaches. The multifaceted nature of private machine learning underscores the importance of diverse research directions, as different approaches may offer unique insights or advantages in specific contexts.

4.1 Private Datasets

EyePACS. The EyePACS [Kag] dataset on Kaggle, supported by the California Healthcare Foundation, features a collection of colored images for diabetic retinopathy screening. EyePACS, a free platform for retinopathy screening, provided these images. The dataset consists of a training set with 35,126 images and a test set with 53,576 eye fundus images. Each image is labelled according to the severity of Diabetic Retinopathy (DR) on a scale from 0 to 4, where 0 indicates a normal retina without DR, and 4 indicates proliferative DR. These labels conform to the standards of the International Clinical Diabetic Retinopathy severity scale. The original image resolution varies, averaging around 3000×2000 pixels, but we resize them all to 224×224 pixels for our experiments.

CheXpert. The CheXpert (**C**hest **e**Xpert) [IRK⁺19] dataset is a comprehensive dataset for chest radiograph interpretation, comprising 224,316 chest radiographs images of size 390×390 from 65,240 patients. The downstream task on this data is multilabel classification. Each label corresponds to one of five pathology observations: ‘Cardiomegaly’, ‘Edema’, ‘Consolidation’, ‘Atelectasis’, and ‘Pleural Effusion’. In this work, we implement state-of-the-art training practices by re-scaling all images to 224×224 pixels and augmenting the dataset with random affine transformations [YYSY21, YZQ⁺23].

We have previously discussed why this work is focusing on medical imaging data; simply because it is privacy-critical, and differentially private methods are likely to be implemented in this setting. We anticipate that advancements made using these benchmarks will correlate with improvements in truly private tasks within similar domains. We selected two of the most common areas within medical imaging: chest X-rays and fundus photography. Chest X-rays are one of the most frequently performed diagnostic imaging procedures, vital for detecting various lung and heart conditions. Fundus photography, which captures images of the retina, is crucial for diagnosing eye diseases. We chose the CheXpert dataset for chest X-rays and the EyePACS dataset for fundus images. These selections were based on their popularity and widespread use in the research community, their sufficient size to meaningfully apply DP techniques, and their relevance to real-world medical applications. Their difference in size, how their respective classes are balanced, and downstream tasks adds to the diversity of the benchmarks.

Our decision was also influenced by the work of Raghu et al. [RZKB19], which explored

medical image classification using CheXpert and a proprietary diabetic retinopathy dataset from Google. This study provided valuable insights into the application of machine learning in these domains. It is worth noting that while several public fundus photography datasets exist, many are too small (containing fewer than 100 images) to be suitable for effective DP implementation and be presented as a benchmark. EyePACS stands out as the largest and most comprehensive public dataset in this field, making it an ideal choice for our study.

While it might seem that we recommend the sole use of EyePACS and CheXpert in private settings, this is not the case. It is crucial to remember a large part of advancements in DP ML often come from being able to directly compare private and non-private approaches. With this in mind, we advocate for the continued use of CIFAR-10 and ImageNet as benchmarks for training DP ML models from scratch, i.e. without any pre-training data. The widespread use of these datasets in the broader machine learning community provides a valuable reference point, allowing researchers to bridge the current performance gap between DP ML and traditional, non-private methods. This comparative approach helps track how far behind DP ML methods are compared to their non-private counterparts.

4.2 Public Datasets

Recent advancements in DPML have shown that pre-training models on public data can significantly enhance utility, as demonstrated in [DBH⁺22]. Building on this insight, our experimental design categorizes public datasets into four distinct groups, each serving a unique purpose in our analysis: We begin with a baseline of no public data, which provides the purest measure of progress in DP ML. We then move on to utilizing ImageNet-1K [CLH17], a widely used, large-scale image classification dataset that represents a common benchmark in the field. Because privately fine-tuning a pre-trained CLIP-based model is a common use case, our third category utilizes LAION-2B [SBV⁺22], the dataset used for pre-training OpenCLIP’s ViT-G/14 model. Finally, we include an “anything goes” category, allowing unrestricted public pre-training data, with the condition that the private dataset has not contaminated the public one to prevent issues similar to data leakage. While the diverse nature of the “anything goes” category may limit direct comparisons between individual results, these experiments still provide valuable insights into overall advancements in the field, measuring the absolute progress on a specific benchmark.

Our methodology draws on insights from previous research, such as [TKC24], which highlights how the size and nature of pre-training data can dramatically affect the downstream utility of privately fine-tuned models. By tracking progress across these four categories, we aim to provide a comprehensive analysis of how different public datasets impact the performance of DP ML models.

4.2.1 “Anything goes” zero-shot

One of the subsets of the “anything goes” setting which only needs public data, would be to use zero-shot classification. In non-private image classification, the goal of zero-shot learning is to correctly classify an image it has never “seen”. By seeing, we mean the model has had training examples in its class during training. This method needs large-scale public pre-training to yield acceptable results. In the private world, this could come in handy as there would be no need to expose the private data to the model, corresponding to $\varepsilon = 0$. We recommend keeping track of the current state-of-the-art for these zero-shot methods. They provide a crucial benchmark for overall progress in the field. Additionally, without this context, it’s possible to falsely claim SOTA status for a differentially private method using public data and $\varepsilon > 0$, when there could be a better existing zero-shot result.

4.3 Privacy Settings

Techniques. As discussed thoroughly in Chapter 3, DPSGD is the most popular algorithm for DPML, primarily because of its flexibility, as it can be applied to any differentiable model, including non-convex models. This versatility makes DPSGD particularly valuable in the context of modern deep learning, where non-convex optimization is commonplace. In contrast, other methods in the DP ML toolkit (such as objective perturbation [CMS11, KST12, INS⁺19, RKW23]) often face limitations. These approaches typically work only with convex models, which restricts their applicability in many contemporary ML scenarios. We propose establishing standardized convex settings to address this disparity and facilitate more comprehensive comparisons across different DP ML methods. We recommend two specific configurations: a linear probe (essentially logistic regression) on a Wide ResNet-28-10 pre-trained on ImageNet-1K, and a similar setup using OpenCLIP’s ViT-G/14 pre-trained on LAION-2B.

Privacy Accounting. To summarize the explanations given in Chapter 3, privacy accountants take various hyperparameters and δ as inputs, and output the upper-bound of ε . In an ideal world, the exact privacy cost would suffice as all estimations would be precise, eliminating the need of upper bounds. However, recent work ([PHK⁺23, CGK⁺24, LRKS24]) has shown that some approaches, while purporting to offer improved results, may in fact be creating a misleading impression of better results. In reality, these methods often employ identical training procedures, calling into question the validity of their claimed enhancements. To maintain the comparability of results, we highly recommended to report the privacy proof or the methods used for privacy accounting.

Privacy Parameters. A key motivation of benchmarking is to establish a common basis for comparison. However, in the context of differential privacy (DP), this comparison becomes more complex due to the introduction of two hyperparameters in the evaluation of image classification tasks. Consider a scenario where one method achieves 85% accuracy at $\varepsilon = 1$, while another reaches 90% at $\varepsilon = 2$ on the same private dataset. Determining which method is "better" is not straightforward. To address this challenge, we standardize our approach by fixing the ε values to 1, 3, 5, and 8. This does not invalidate other values of ε , but covers low and high privacy regimes and provides an acceptable range. For δ , we adopt a consistent method across experiments: we set it to the largest power of 10 that does not exceed the inverse of the training set size. It is worth noting that in many settings, substantial increases or decreases in δ often have only a minor impact on the ε value. This standardization allows for more meaningful comparisons across different methods and datasets, providing a clearer picture of the trade-offs between privacy and accuracy in DP image classification.

4.4 Leaderboard

Leaderboards have long been a cornerstone of progress in non-private machine learning, serving as benchmarks for performance and catalysts for innovation¹. These platforms provide researchers with standardized comparisons and historical records of advancement. However, in the realm of Differentially Private Machine Learning (DPML), such a unifying platform has been notably absent.

¹See e.g. paperswithcode.com/sota

Creating a leaderboard for DPML presents unique challenges not encountered in non-private machine learning. The primary difficulty lies in result verification. Unlike non-private models where performance can be easily validated, DPML requires rigorous mathematical proofs to ensure that submitted methods truly operate within their claimed privacy budgets. This verification process must account for various factors including privacy parameters, accounting mechanisms, and the nature of public data utilized. To address these complexities, the leaderboard² implements a community-driven verification system. New submissions are initially marked as unverified, and the broader research community is encouraged to scrutinize and validate claims through a transparent process facilitated by GitHub pull requests.

Transparency and openness are core principles of this DPML leaderboard. Contributors are encouraged to open-source as much of their work as possible, including code, pretrained model weights, and mathematical proofs. This approach fosters trust and reproducibility while providing valuable resources to the community, especially benefiting newcomers to the field.

The current focus on private image classification provides a starting point, with potential for future expansion. Subsequent iterations of the leaderboard could incorporate a wider array of DPML tasks, such as natural language processing, time series analysis, or federated learning scenarios. This growth would provide a more comprehensive view of DPML’s capabilities across diverse domains. Additionally, the leaderboard could evolve to track multiple metrics beyond accuracy, including privacy budget efficiency, model size, and training time, offering a nuanced view of the trade-offs inherent in privacy-preserving techniques.

As research in DPML progresses, this leaderboard will likely adapt, potentially incorporating new privacy definitions, automated verification tools, and expanding its scope to cover a broader spectrum of DPML tasks. While the challenges in implementing and maintaining such a platform are substantial, it has the potential to serve as a valuable tool for benchmarking and advancing the field of DPML.

²See private-machinelearning.github.io/

Chapter 5

Experiments and Findings

This section presents our experimental work, exploring various techniques and their efficacy. We seek to understand the extent to which insights gained from DP image classification on common datasets like CIFAR-10 and ImageNet apply to the more sensitive domain of medical images. Additionally, we examine the role of public data in enhancing the performance on medical datasets. Finally, we aim to determine the level of performance achievable using DP on these medical datasets across different evaluation scenarios.

We will begin by outlining our experimental setup, including the models employed and our chosen evaluation metrics. Following this, we will present our baseline results and analyze their implications. Finally, we will examine whether certain ablations commonly used in DP research remain relevant and effective in this context.¹

5.1 Experiment Settings

Our experimental setup, as outlined in Section 4.3, covers a range of privacy constraints with $\epsilon \in \{1, 3, 5, 8\}$. We fix δ based on the private dataset size: 10^{-5} for EyePACS and 10^{-6} for CheXpert. Using PyTorch [PGM⁺19a] and its DPML library, Opacus [YSS⁺21], we employ the Adam [KB15] optimizer to avoid hyperparameter tuning on the learning rate, fixing it on 0.001. However, we still tune the clipping norm, considering values

¹All codes for this project can be found at github.com/mshubhankar/DP-Benchmarks/

in $\{0.001, 0.01, 0.1, 1, 10\}$. All experiments maintain a consistent batch size of 1024 and run for 20 epochs, except for non-private experiments where we implement early stopping due to overfitting. Notably, private experiments did not exhibit overfitting, likely due to the regularizing effect of DP-Adam. To ensure robustness, we conduct each experiment independently three times and report the mean result.²

Evaluation Metrics The Area Under the Curve (AUC) is a widely used performance metric in machine learning, particularly for binary classification problems [Bra97]. It represents the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. The AUC ranges from 0 to 1, where a value of 1 indicates perfect classification, 0.5 represents random guessing, and values below 0.5 suggest worse-than-random performance. One of the key advantages of AUC is its insensitivity to class imbalance, making it particularly useful when dealing with datasets where one class significantly outnumbers the other. The AUC can be interpreted as the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance, providing a measure of the model’s ability to discriminate between classes across all possible classification thresholds [Faw06]. We use the AUC metric for assessing results on the CheXpert and EyePACS datasets. For the CIFAR-10 dataset, we use the accuracy metric. These choices align with the commonly accepted evaluation practices for each respective dataset in the machine learning community.

5.2 Models

ScatterNets ScatterNets [OM14] are convolutional neural networks (CNNs) that utilize pre-defined wavelets for their architecture and filters, as opposed to learning them. In other words, the features are “hand-crafted” rather than learned from data, and so use neither public nor private data. Tramèr and Boneh [TB21] employ this architecture for DP image classification. Once handcrafted features are extracted, the classifier uses DP-SGD and normalization techniques to train the final model. They demonstrate compelling results on MNIST and CIFAR-10, particularly for small values of ϵ . What sets this concept apart is

²Experiments were run on NVIDIA A100 80GB PCIe GPUs.

its shallowness and complete data independence, making it compatible with differentially private learning. We exclusively use ScatterNets without any public data.

Wide-ResNets Wide-ResNets [ZK16] are powerful Residual Network (ResNet) [HZRS15] models that minimize gradient vanishing and exploding problems by “widening” the models instead of deepening them. By adding a widening factor, [ZK16] multiply the number of features in the convolutional layers of the original ResNet architecture. De et al. [DBH⁺22] reach the best results on CIFAR-10 by taking a Wide-ResNet with 28 convolutional layers and widening factor of 10 (WRN-28-10). For pre-training, they use ImageNet1k [RDS⁺15] downsampled to 32×32 and augmented with random flipping, mirror padding, and random cropping. Then, they privately fine-tune it on the CIFAR-10 dataset. To have a fair comparison, we keep all the settings they used, namely using weight standardization [BDS21a] and replacing all batch normalization layers with group normalization ones. Their models along with their pre-trained weights are available in JAX³. By utilizing their identical pre-trained model, we aim to study whether their proposed state-of-the-art holds for our proposed datasets.

CLIP-based models. CLIP [RKH⁺21] is a popular contrastive learning pre-training technique, which allows one to jointly train a language and image encoder, bridging the gap between natural language and visual understanding. CLIP has been observed to enable robust zero-shot image classification when pre-training on very large Internet datasets. It does so by training on a very large corpus of image-text pairs gathered from the internet. We use two ViT [DBK⁺21] models pre-trained using CLIP: OpenAI’s ViT-B/16, which is pre-trained on the proprietary WebImageText (WIT) dataset, and OpenCLIP’s ViT-G/14 model (pre-trained on LAION-2B [SBV⁺22]).⁴ Both these models consist of multiple vision transformer layers that contain a multi-head attention mechanism followed by a feed-forward neural network. The two models differ extensively on the size of the model and image patch size that is used for dividing the image into sections during processing. The B/16 model has 12 transformer layers with 16 patch size and the G/14 model consists of 48 layers and 14 patch size. For zero-shot experiments, we use these models as-is. For other experiments, we first pass the datasets into the CLIP model and store its image

³https://github.com/google-deepmind/jax_privacy/

⁴https://github.com/mlfoundations/open_clip

representations. Then we train two models, logistic regression and a two-layer linear neural network (TLNN) on top of these features using DPSGD. We exclusively use CLIP-based models with their respective public pre-training datasets.

5.3 Main Findings

The following section outlines some findings from our empirical study. All results can be found in Appendix [A](#)

5.3.1 Improvements from pre-training are not uniform across datasets and depend on the private data.

To investigate the impact of different public datasets on downstream tasks, we conducted experiments across various settings, as illustrated in Figure [5.1](#). For simplicity, we present only the Scatternet+Linear results for scenarios without public data, as this consistently outperforms training WideResnets from scratch.

Our EyePACS experiments reveal a clear trend: fine-tuning always improves performance. In the zero-shot setting, where ViT models are unexposed to private data, the AUC is 50.73%, indicating near-random classification. However, relaxing privacy constraints by increasing ε and adding a tunable final layer significantly boosts performance, with AUC reaching 63.42% and 63.9% for ViT-G/14 and ViT-B/16, respectively.

Notably, outside the zero-shot setting, the Scatternet model consistently underperforms, underscoring the benefits of pre-training for the EyePACS dataset. The relative performance ranking of different models remains largely consistent across ε values, with one exception: Moving towards less private settings, the advantages of the pre-trained ViT-G/14 become more pronounced, particularly between $\varepsilon = 1$ and $\varepsilon = 3$, where its performance approaches that of the Wide-ResNet+Linear model.

Building on our analysis of EyePACS, we now turn our attention to the CheXpert dataset. The zero-shot settings for CheXpert similarly yield results approximating random classification. As expected, the addition of a tunable final layer significantly enhances model performance. Interestingly, this improvement is more pronounced than what we observed with EyePACS, suggesting that the LAION-2B and WIT datasets may be more

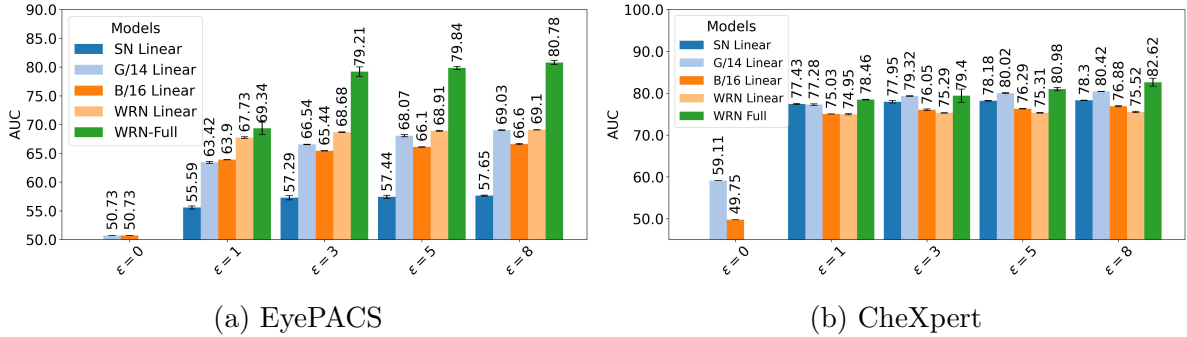


Figure 5.1: Improvements from pre-training depend on the private data: Public data consistently benefits EyePACS, while ViT-G/14 without full fine-tuning yields the best results on CheXpert.

suitable as public datasets for CheXpert classification compared to EyePACS. When comparing models that extract features and tune only a single layer, ViT-G/14 demonstrates superior feature extraction capabilities for CheXpert. However, the fully fine-tuned Wide-ResNet model outperforms ViT-G/14 in overall performance. It is important to note that this comparison isn't entirely fair, given that Wide-ResNet undergoes complete fine-tuning while ViT-G/14 does not. Therefore, this performance difference does not necessarily indicate that Wide-ResNet is inherently better suited for CheXpert tasks. Notably, for smaller ϵ values, which represent stricter privacy constraints, ViT-G/14 emerges as the superior model when considering only linear fine-tuning. This observation highlights the model's efficiency in extracting meaningful features while maintaining strong privacy guarantees.

5.3.2 Increased privacy budget enhances pre-training benefits

Our examination of the CheXpert experiments reveals interesting patterns across different privacy budgets. At $\epsilon = 1$, the strictest privacy setting, the Scatternet linear model outperforms other approaches, as shown in 5.2. This result aligns with Tramèr and Boneh's research, demonstrating that ScatterNets excel under stringent privacy conditions, although the absolute performance here is lower than their CIFAR-10 findings.

However, this advantage diminishes as we relax privacy restrictions. At $\epsilon=3$, we observe a substantial performance boost in the ViT-G/14 model, surpassing the Scatternet results. This suggests that the benefits of pre-training on public datasets become more pronounced

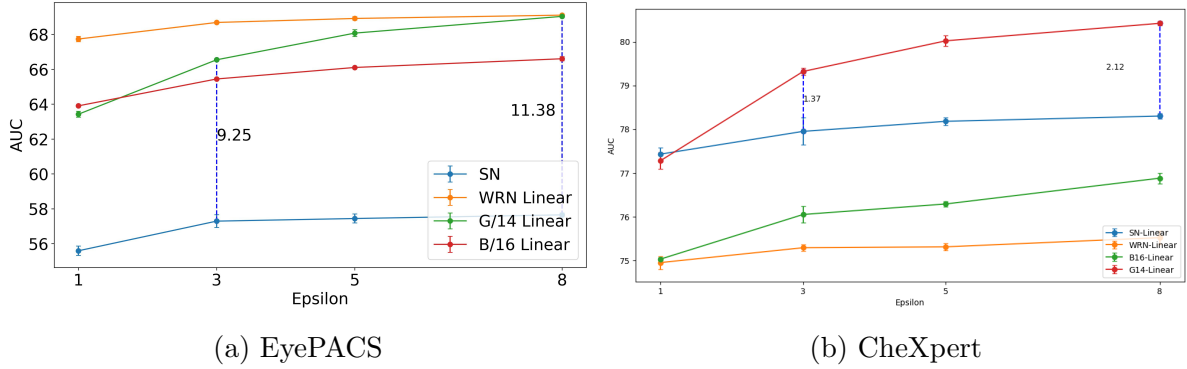


Figure 5.2: Pre-training public data is more beneficial with higher ϵ values. For CheXpert, ScatterNet performs better at smaller ϵ values, while pretrained models show marginal improvements at larger ϵ values. Similarly, for EyePACS, CLIP ViT-G/14 linear performs better as ϵ value increases.

as privacy constraints are relaxed. The ViT-B/16 linear model, while initially underperforming, gradually narrows the gap with increased ϵ values, though it never quite matches the Scatternet model’s performance.

The EyePACS task presents a different scenario. Here, the Scatternet model’s performance is less impressive, likely due to EyePACS’ size being approximately one-sixth of CheXpert’s, and the high dimensionality of Scatter features. This combination makes it challenging to balance complexity with private training requirements.

For EyePACS, the Wide-ResNet+Linear model consistently outperforms others across all ϵ values. Its performance remains high and stable, with limited improvement as ϵ increases, suggesting it achieves near-optimal results even under strict privacy constraints.

The strong performance of Wide-ResNet from the outset aligns with our earlier observation that pre-trained models seem particularly beneficial for EyePACS. Notably, while Wide-ResNet maintains its advantage, ViT-G/14 shows more substantial improvements as the privacy budget increases, indicating its potential in scenarios with relaxed privacy constraints.

5.3.3 CIFAR-10 success is not a proxy for benchmark progress

The advancements in differentially private machine learning on CIFAR-10 paint an overly optimistic picture when contrasted with performance on more complex, real-world datasets. Take, for instance, the ViT-G/14 model: Despite achieving a near-perfect 99.75% AUC on CIFAR-10, it falls short on EyePACS with a 50.73% AUC and on CheXpert with a 59.11% AUC. Similarly, while Wide-ResNet demonstrates strong performance on CIFAR-10 with a 94.7% AUC at $\varepsilon = 1$, its effectiveness diminishes considerably on CheXpert (78.46% AUC) and EyePACS (69.34% AUC) under the same privacy constraints.

This performance gap underscores a critical insight: techniques and models that excel on CIFAR-10 may not have the same advantages when applied to more diverse and challenging datasets like EyePACS and CheXpert. As we analyze some ablations in the next section, it will become evident that the successful strategies on CIFAR-10 do not consistently produce similar benefits across different medical image datasets. Additionally, the different outcomes between EyePACS and CheXpert emphasize the need for caution when generalizing DPML methods, reminding us that celebrated approaches in private image classification are not *universally* applicable across all settings and datasets.

5.4 Revisiting Ablations

Our investigation builds upon the groundbreaking work of [DBH⁺22], which involves one of the most extensive ablation studies for differentially private image classification. Their research on CIFAR-10 demonstrated significant improvements, elevating accuracy from 50.8% to 79.7% through a combination of techniques including using group normalization, large batch sizes, weight standardization, augmentation multiplicity, and parameter averaging.

We sought to extend this analysis to our own architectures, namely CLIP-based ViTs and ScatterNets, while maintaining a fixed privacy budget of $\varepsilon = 3$. Our methodology involves replicating [DBH⁺22]’s sequence of ablations, with a key distinction: we only retain techniques that demonstrate clear benefits in our context.

Our findings reveal a more nuanced picture than the uniformly positive results observed in the original study on CIFAR-10. The effectiveness of these techniques varied significantly across different parameters and scenarios in our experiments. A breakdown of the results

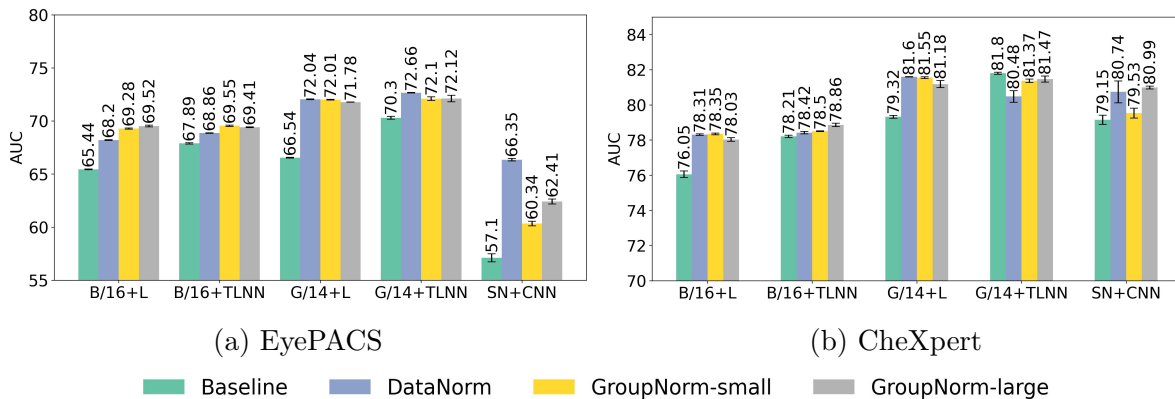


Figure 5.3: Normalization generally improves the final performance of all models. The yellow bars correspond to groups of 8 and 9, while the gray bars correspond to groups of 16 and 27 for CLIP-ViTs and ScatterNet, respectively. (The choice of 27 over 81 is due to its superior performance.)

is available in the Appendix A.

Normalization. Normalization techniques play a crucial role in deep learning, particularly in computer vision architectures where batch normalization (BatchNorm) [IS15] layers are often seen. However, BatchNorm poses challenges in DP-SGD due to its mixing of representations across samples, which complicates privacy loss analysis [YSS+22]. To address this incompatibility, alternative normalization methods have been proposed for differentially private learning [TB21, DBH+22, BDS+23, DFAP21]. Group normalization (GroupNorm) [WH18] divides channels into groups and normalizes within each group, operating independently of batch size. This independence makes it suitable for privacy-preserving scenarios, as it does not need to access the mean and variance of the private data. Another approach is data normalization (DataNorm) [SS20], which normalizes activations using pre-computed statistics from the training data, maintaining sample independence crucial for differential privacy. It is important to note that the per-channel means and variances must be privately estimated, and this poses a computational overhead. We use Gaussian noise with ($\sigma = 8$) to calculate the estimations for all runs. These techniques enable effective normalization in DP-SGD without compromising privacy guarantees.

For the CLIP ViT models, we experiment with 8 and 16 groups. For Scatter features,

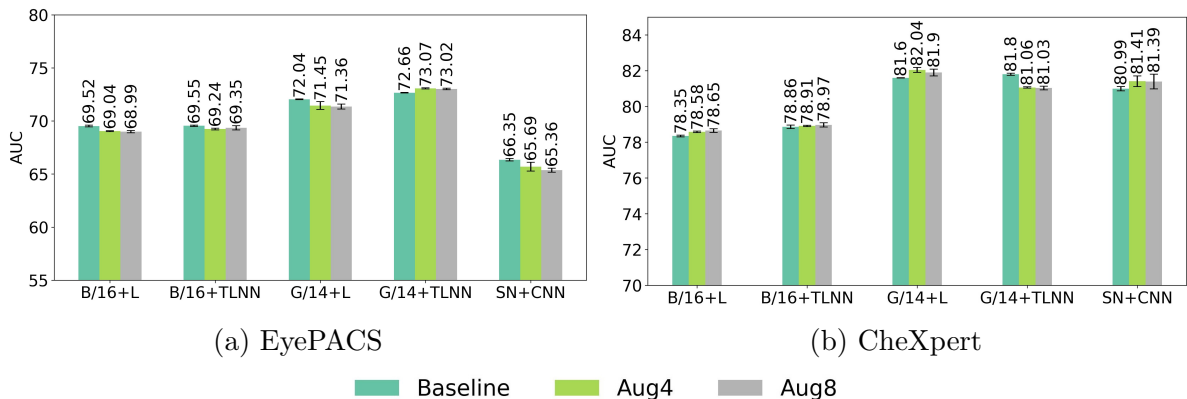


Figure 5.4: Evaluating augmentation multiplicity by adding 4 and 8 augmentations of each image in the training data. Augmentation multiplicity helps in general for CheXpert but not for EyePACS.

we use 9, 27, and 81 groups, following [TB21]. In Figure 5.3, we see that normalization demonstrates benefits across various techniques and datasets, though its impact varies depending on the specific dataset and architecture. The EyePACS dataset shows particularly notable improvements with normalization, especially when using Scatter features. This is likely due to the large gradients of its features, as evidenced by the highest performance achieved with the largest clipping norm in experiments using this dataset. For CLIP features, normalization remains beneficial, again more on EyePACS, with group normalization yielding the best results. The impact on CLIP features, while significant, is not as pronounced as with Scatter features. CheXpert demonstrates only minor improvements with normalization. For Scatter features on CheXpert, group normalization produces the best results, suggesting that CheXpert’s Scatter features have smaller gradients compared to EyePACS and CIFAR-10. This moderation in gradient size likely explains the reduced impact of data normalization on CheXpert. An interesting observation is that normalization appears most effective in cases where models exhibit larger unclipped gradients. This suggests that data normalization can effectively manage large gradient magnitudes, particularly when gradient clipping underestimates the actual gradient norms.

Larger Batch Size. The effect of larger batch sizes in differentially private training has been established through both theoretical research [BFGT20, TTZ14] and empirical studies [AGG⁺21, DBH⁺22]. One thing to note is [DBH⁺22]’s warning about larger batch

sizes: It can lead to overfitting, and increased overfitting can indicate a rise in privacy leakage [YSS⁺21]. While DP inherently provides generalization bounds that prevent overfitting for small ε values [DSS⁺15, BNS⁺16], these bounds degrade rapidly as ε increases. Our experiments revealed varying impacts across different datasets when scaling the batch size from 1024 to 4096. CheXpert showed improvements in four out of five experiments, while EyePACS demonstrated no significant benefits. Getting different results could be due to the difference in dataset sizes since CheXpert’s training set is approximately six times larger than EyePACS. Consequently, with a fixed number of epochs, EyePACS experienced fewer model update steps, potentially leading to underfitting. Interestingly, when we extended the number of epochs for EyePACS, we observed positive effects from larger batch sizes, particularly when using the ScatterNet model. These findings highlight the complex relationship between batch size, dataset characteristics, and model performance in differentially private training scenarios.

Weight Standardization Weight Standardization (WS) is a technique designed to improve the training of convolutional neural networks by normalizing the weights of convolutional layers [QWL⁺19]. It operates by standardizing the weight tensor to have zero mean and unit variance before the convolution operation. WS aims to reduce internal covariate shift and stabilize the optimization process, potentially leading to faster convergence and improved generalization [HMT⁺17]. When combined with other normalization techniques like Batch Normalization, WS has shown particular effectiveness in tasks with limited data and in training very deep networks [QWL⁺19]. We apply this ablation specifically to the ScatterNet+CNN model, as it is the only configuration in this study that incorporates CNNs. We observe that WS does not help with EyePACS but helps with CheXpert and CIFAR-10. In line with observations from previous studies [BDS21b, DBH⁺22], our results suggest a positive correlation between group normalization and Weight Standardization. However, due to the limited scope of our experiments, we cannot draw definitive conclusions.

Augmentation Multiplicity. Data augmentation is a crucial technique in computer vision that artificially expands the training dataset by applying various transformations to existing images [SK19]. These techniques especially affect over-parameterized models. They help improve model generalization by exposing the network to a wider range of image variations, thereby reducing overfitting [PW17]. The effectiveness of data augmentation

varies depending on the specific task, dataset, and model architecture.

In their DP-SGD experiments, [DBH⁺22] observed that using only one augmentation per example reduced both training and validation accuracy. Their analysis is that this happens because data augmentation increases gradient variance, necessitating more training iterations. They take inspiration from non-private work in the field, where multiple augmentations per example have been shown to enhance performance. However, using multiple augmentations in DP-SGD typically raises the privacy cost, as each augmented image would require a separate gradient calculation. To counter this, they average the gradients of different augmentations for the same example before clipping. This approach keeps the privacy cost unchanged.

Following their work, we apply a sequence of augmentations to our benchmark datasets: reflect padding, random cropping, and random horizontal flipping. They find that 16 augmentations yield the best result on CIFAR-10. Since our benchmark datasets are larger and don't have as much computational resources, we only evaluate 4 and 8 augmentations per example. Our results are again in contradiction with theirs, as augmentations generally benefit CheXpert but not EyePACS. It is possible that creating data augmentations specific to each dataset (e.g. instead of "random cropping", do something that would make more sense to the nature of the dataset) could yield better results.

Parameter Averaging The differential privacy guarantees of DP-SGD are based on the assumption that all interim parameter values generated during the training process can be safely disclosed. As a result, techniques that involve averaging these parameters do not introduce any additional privacy concerns or costs. Because of this, [DBH⁺22] suggests using exponential moving average (EMA) [PJ92] of all the parameters in the model. EMA is a type of moving average that gives more weight to recent data points, improving model stability and performance. Our findings show this ablation, similar to previous ones, does not universally help and only occasionally improves performance.

Chapter 6

Conclusions and Future Work

6.1 Concluding Remarks

This thesis has explored differentially private machine learning techniques for image classification, with a particular emphasis on privacy-critical domains such as medical imaging. Our research has yielded significant findings and contributions to the field.

We have proposed a set of standardized settings for benchmarking DP image classification, providing a consistent framework for evaluating and comparing different techniques, with a focus on medical imaging as a privacy-sensitive domain. To facilitate ongoing research and development, we have established a leaderboard for tracking progress on image classification benchmarks under differential privacy constraints.

Our experimental investigation revealed important insights into the effectiveness of DP ML techniques. We found that methods which have shown great success in certain contexts are not *universally* effective across all datasets and architectures. Notably, progress on standard benchmarks like CIFAR-10 does not necessarily translate to improved performance on benchmark datasets. This finding underscores the importance of developing different approaches when it comes to private data.

While it is challenging to design universally effective techniques, our work highlights that significant opportunities remain for new ideas and methods to advance the field of private image classification. However, it is crucial to note that our experiments were conducted on a limited number of datasets and architectures. Therefore, while our findings

provide valuable insights, they should not be interpreted as definitive conclusions for the entire field of DP image classification.

6.2 Future Work

Based on our findings and the current state of the field, we propose several directions for future research. While our work focused on image classification, future research should explore benchmarks in other areas such as Natural Language Understanding and Generation. This expansion would provide a more comprehensive understanding of DPML across diverse applications.

To ensure fair comparisons and push the boundaries of performance, future work should investigate the use of more advanced model architectures. For instance, experiments using the NFNet-F7 [BDSS21] model pre-trained on ImageNet-1k could be compared with our WideResnet experiments. We encourage the privacy research community to build upon our baseline experiments by exploring different models and working to improve performance further. This collaborative effort will be crucial in advancing the field of differentially private machine learning.

Future research should also extend to a wider range of datasets, both within and beyond the medical domain. This exploration will help in understanding the generalizability of DP ML techniques and identifying domain-specific challenges. Given our finding that progress on standard benchmarks doesn't always translate to medical images, future work should be mindful of the privacy guarantees they provide, and what it really means. This could involve developing techniques that are more adaptable across different domains or creating specialized approaches for privacy-critical areas.

The continued maintenance and updating of the leaderboard we've established will be crucial for tracking long-term progress in the field and identifying emerging trends or breakthroughs. This ongoing effort will provide valuable insights into the evolution of DP ML techniques over time.

In conclusion, while our work has provided valuable insights and benchmarks for DP image classification, it also highlights the need for continued research and innovation in this critical area of privacy-preserving machine learning. The field remains ripe for new ideas and approaches that can push the boundaries of what's possible in private image

classification across various domains and architectures. As we move forward, the collective efforts of the research community will be essential in addressing the challenges and opportunities identified in this thesis, ultimately advancing the state of the art in differentially private machine learning.

References

- [ACG⁺16a] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM Conference on Computer and Communications Security, CCS '16*, pages 308–318, New York, NY, USA, 2016. ACM.
- [ACG⁺16b] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.
- [AGG⁺21] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private BERT. *arXiv preprint arXiv:2108.01624*, 2021.
- [AKC⁺22] Mohammed Adnan, Shivam Kalra, Jesse C Cresswell, Graham W Taylor, and Hamid R Tizhoosh. Federated learning and differential privacy for medical image analysis. *Sci Rep*, 12(1):1953, February 2022.
- [BDS21a] Andrew Brock, Soham De, and Samuel L. Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets. *CoRR*, abs/2101.08692, 2021.
- [BDS21b] Andrew Brock, Soham De, and Samuel L. Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets. *CoRR*, abs/2101.08692, 2021.

- [BDS⁺23] Leonard Berrada, Soham De, Judy Hanwen Shen, Jamie Hayes, Robert Stanforth, David Stutz, Pushmeet Kohli, Samuel L Smith, and Borja Balle. Unlocking accuracy and fairness in differentially private image classification. *arXiv preprint arXiv:2308.10888*, 2023.
- [BDSS21] Andrew Brock, Soham De, Samuel L. Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization, 2021.
- [BFGT20] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.
- [BKZ23] Alex Bie, Gautam Kamath, and Guojun Zhang. Private GANs, revisited. *Transactions on Machine Learning Research*, 2023.
- [BNS⁺16] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, STOC ’16, pages 1046–1059, New York, NY, USA, 2016. ACM.
- [Bot10] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010*, pages 177–186. Physica-Verlag HD, 2010.
- [Bra97] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.
- [BS22] Danah Boyd and Jayshree Sarathy. Differential perspectives: Epistemic disconnects surrounding the us census bureau’s use of differential privacy. *Harvard Data Science Review (Forthcoming)*, 2022.
- [BWW⁺19] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.
- [CBV⁺21] Tianshi Cao, Alex Bie, Arash Vahdat, Sanja Fidler, and Karsten Kreis. Don’t generate me: Training differentially private generative models with

- sinkhorn divergence. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21, pages 12480–12492. Curran Associates, Inc., 2021.
- [CDMS22] Aloni Cohen, Moon Duchin, JN Matthews, and Bhushan Suwal. Private numbers in public policy: Census, differential privacy, and redistricting. *Harvard Data Science Review*, Jun 2022.
- [CGK⁺24] Lynn Chua, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. How private are DP-SGD implementations? In *Proceedings of the 41st International Conference on Machine Learning*, ICML '24. JMLR, Inc., 2024.
- [CLH17] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the CIFAR datasets. *CoRR*, abs/1707.08819, 2017.
- [CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011.
- [DBH⁺22] Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.
- [DBK⁺21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.
- [DCVK23] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *Transactions on Machine Learning Research*, 2023.
- [DDS⁺09] 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. Ieee, 2009.

- [DFAP21] Friedrich Dörmann, Osvald Frisk, Lars Nørvang Andersen, and Christian Fischer Pedersen. Not all noise is accounted equally: How differentially private learning benefits from large sampling rates, 2021.
- [DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.
- [DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Machine Learning*, 9(3–4):211–407, 2014.
- [DRV10] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.
- [DSS⁺15] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 2015.
- [Faw06] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
- [FBP⁺22] Stanislav Fort, Andrew Brock, Razvan Pascanu, Soham De, and Samuel L. Smith. Drawing multiple augmentation samples per image during training efficiently decreases test error, 2022.
- [FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer, 2001.
- [GAW⁺22] Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision, 2022.
- [GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

- [GLW21] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21, pages 11631–11642. Curran Associates, Inc., 2021.
- [HJSP23] Frederik Harder, Milad Jalali, Danica J Sutherland, and Mijung Park. Pre-trained perceptual features improve differentially private image generation. *Transactions on Machine Learning Research*, 2023.
- [HMT⁺17] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1558–1567. JMLR. org, 2017.
- [HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [INS⁺19] Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *Proceedings of the 40th IEEE Symposium on Security and Privacy*, SP '19, pages 299–316, Washington, DC, USA, 2019. IEEE Computer Society.
- [IRK⁺19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Illcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI '19, pages 590–597, 2019.
- [IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, ICML '15, pages 448–456. JMLR, Inc., 2015.

- [Kag] Diabetic retinopathy grading on kaggle eye-pacs. <https://paperswithcode.com/sota/diabetic-retinopathy-grading-on-kaggle>. Accessed: 2024-06-01.
- [KB15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15, 2015.
- [KOV17] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.
- [Kri09] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.
- [KSC⁺22] Alexey Kurakin, Shuang Song, Steve Chien, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy, 2022.
- [KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
- [KST12] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, COLT '12, pages 25.1–25.40, 2012.
- [KZR⁺22] Helena Klaue, Alexander Ziller, Daniel Rueckert, Kerstin Hammernik, and Georgios Kaissis. Differentially private training of residual networks with scale normalisation, 2022.
- [LRKS24] Christian Janos Lebeda, Matthew Regehr, Gautam Kamath, and Thomas Steinke. Avoiding pitfalls for privacy accounting of subsampled mechanisms under composition. *arXiv preprint arXiv:2405.20769*, 2024.
- [LTLH22] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *Proceedings*

of the 10th International Conference on Learning Representations, ICLR '22, 2022.

- [Mir17] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.
- [Mit97] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
- [MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science, FOCS '07*, pages 94–103, Washington, DC, USA, 2007. IEEE Computer Society.
- [MV16] Jack Murtagh and Salil Vadhan. The complexity of computing the optimal composition of differential privacy, 2016.
- [OM14] Edouard Oyallon and Stéphane Mallat. Deep roto-translation scattering for object classification. *CoRR*, abs/1412.8659, 2014.
- [PGM⁺19a] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 8026–8037. Curran Associates, Inc., 2019.
- [PGM⁺19b] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.
- [PHK⁺23] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and

- Abhradeep Guha Thakurta. How to DP-fy ML: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023.
- [PJ92] Boris Polyak and Anatoli Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 07 1992.
- [POG⁺24] Ke Pan, Yew Ong, Maoguo Gong, Hui Li, A.K. Qin, and Yuan Gao. Differential privacy in deep learning: A literature survey. *Neurocomputing*, 589:127663, 04 2024.
- [PT21] Nicolas Papernot and Abhradeep Thakurta. How to deploy machine learning with differential privacy? DifferentialPrivacy.org, 10 2021. <https://differentialprivacy.org/how-to-deploy-ml-with-dp/>.
- [PTS⁺21] Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Úlfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI ’21, pages 9312–9321, 2021.
- [PW17] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.
- [QWL⁺19] Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan L. Yuille. Weight standardization. *CoRR*, abs/1903.10520, 2019.
- [RBBNZM21] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021.
- [RCR16] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization, 2016.
- [RDS⁺15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015.

- [RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.
- [RKH⁺21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [RKW23] Rachel Redberg, Antti Koskela, and Yu-Xiang Wang. Improving the privacy and practicality of objective perturbation for differentially private linear learners. In *Advances in Neural Information Processing Systems 36*, NeurIPS ’23, pages 13819–13853. Curran Associates, Inc., 2023.
- [RZKB19] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging, 2019.
- [SBV⁺22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [SK19] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.
- [Sor10] Mohammad S Sorower. A literature survey on algorithms for multi-label learning. Technical report, Oregon State University, 2010.
- [SS20] Dalwinder Singh and Birmohan Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524, 2020.

- [SWSS21] Yinchun Shen, Zhiguo Wang, Ruoyu Sun, and Xiaojing Shen. Towards understanding the impact of model size on differential private classification, 2021.
- [TB21] Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *Proceedings of the 9th International Conference on Learning Representations*, ICLR '21, 2021.
- [TFR22] Amirsina Torfi, Edward A. Fox, and Chandan K. Reddy. Differentially private synthetic medical data generation using convolutional gans. *Information Sciences*, 586:485–500, March 2022.
- [TK07] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.
- [TKC24] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. In *Proceedings of the 41st International Conference on Machine Learning*, ICML '24. JMLR, Inc., 2024.
- [TTZ14] Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2014.
- [VMB19] Mike Voets, Kajsa Møllersen, and Lars Ailo Bongo. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLOS ONE*, 14(6):e0217541, June 2019.
- [WH18] Yuxin Wu and Kaiming He. Group normalization. *CoRR*, abs/1803.08494, 2018.
- [XLW⁺18] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

- [YNB⁺22] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *Proceedings of the 10th International Conference on Learning Representations, ICLR '22*, 2022.
- [YQAH22] Fan Yang, Yanan Qiao, Mohammad Zoynul Abedin, and Cheng Huang. Privacy-preserved credit data sharing integrating blockchain and federated learning for industrial 4.0. *IEEE Transactions on Industrial Informatics*, 18(12):8755–8764, 2022.
- [YSS⁺21] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Gosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.
- [YSS⁺22] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in pytorch, 2022.
- [YYSY21] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049, 2021.
- [YZCL21a] Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning, 2021.
- [YZCL21b] Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning, 2021.
- [YZQ⁺23] Zhuoning Yuan, Dixian Zhu, Zi-Hao Qiu, Gang Li, Xuanhui Wang, and Tianbao Yang. Libauc: A deep learning library for x-risk optimization. In *29th SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

- [ZK16] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016.
- [ZMH21] Huanyu Zhang, Ilya Mironov, and Meisam Hejazinia. Wide network learning with differential privacy, 2021.
- [ZWB21] Yingxue Zhou, Zhiwei Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private SGD with gradient subspace identification. In *Proceedings of the 9th International Conference on Learning Representations*, ICLR '21, 2021.
- [ZZ13] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2013.

APPENDICES

Appendix A

All Experiment Results

Table A.1: Test AUC evaluation for EyePACS at different privacy regimes. Baselines include ScatterNet (SN), WideResNet (WRN) and CLIP models on datasets with different public data pre-training. The SOTA is due to [VMB19].

Public data	Model	Test AUC (%)				
		$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 5$	$\varepsilon = 8$	$\varepsilon = \infty$
None	SN + CNN	55.02 ± 0.12	57.1 ± 0.38	57.13 ± 0.1	57.34 ± 0.34	68.81 ± 0.5
None	SN + Linear	55.59 ± 0.27	57.29 ± 0.37	57.44 ± 0.26	57.65 ± 0.10	69.70 ± 0.1
None	WRN (Scratch)	55.51	56.49	56.83	57.79	62.15
IN-32	WRN + Linear	67.73 ± 0.15	68.68 ± 0.05	68.91 ± 0.10	69.10 ± 0.03	73.25
IN-32	WRN (Full)	69.34 ± 1.09	79.21 ± 0.83	79.84 ± 0.27	80.78 ± 0.38	83.66
WIT	B/16 + Linear	63.9 ± 0.04	65.44 ± 0.04	66.1 ± 0.05	66.6 ± 0.12	69.93 ± 0.0
WIT	B/16 + TLNN	65.12 ± 0.04	67.89 ± 0.06	69.22 ± 0.1	69.84 ± 0.02	70.54 ± 0.0
LAION	G/14 + Linear	63.42 ± 0.17	66.54 ± 0.04	68.07 ± 0.2	69.03 ± 0.06	69.88 ± 0.2
LAION	G/14 + TLNN	65.47 ± 0.02	70.30 ± 0.13	71.74 ± 0.3	72.3 ± 0.19	73.36 ± 0.1
IN-1K	SOTA	-	-	-	-	95.1

¹The authors use an ensemble of five models.

Table A.2: Test AUC evaluation for CheXpert at different privacy regimes. Baselines include ScatterNet (SN), WideResNet (WRN) and CLIP models on datasets with different public data pre-training. The SOTA is due to [BDS⁺23](Private) and [YYSY21](Non-Private).

Public data	Model	Test AUC (%)				
		$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 8$	$\epsilon = \infty$
None	SN + CNN	78.16 ± 0.22	79.15 ± 0.26	79.16 ± 0.18	79.68 ± 0.04	$80.65 \pm 0.$
None	SN + Linear	77.43 ± 0.15	77.95 ± 0.31	78.18 ± 0.09	78.30 ± 0.06	$78.94 \pm 0.$
None	WRN (Scratch)	76.87	77.79	77.92	78.75	87.28
IN-32	WRN + Linear	74.95 ± 0.15	75.29 ± 0.07	75.31 ± 0.08	75.52 ± 0.16	75.84
IN-32	WRN (Full)	78.46 ± 0.07	79.40 ± 1.57	80.98 ± 0.42	82.62 ± 0.94	87.57
WIT	B/16 + Linear	75.03 ± 0.03	76.05 ± 0.19	76.29 ± 0.06	76.88 ± 0.12	$76.89 \pm 0.$
WIT	B/16 + TLNN	77.28 ± 0.19	78.21 ± 0.07	78.33 ± 0.04	78.54 ± 0.06	$78.56 \pm 0.$
LAION	G/14 + Linear	77.28 ± 0.19	79.32 ± 0.08	80.02 ± 0.12	80.42 ± 0.05	$80.48 \pm 0.$
LAION	G/14 + TLNN	80.63 ± 0.4	81.80 ± 0.06	82.25 ± 0.02	$82.27 \pm 0.0.4$	$82.28 \pm 0.$
IN-21K	SOTA	86.3	-	-	89.2	-
IN-1K	SOTA	-	-	-	-	93 ¹

Table A.5: Studying the impact of augmentation multiplicity, we find that it consistently improves performance for CIFAR-10. However, looking at EyePACS and CheXpert, we observe inconsistent behavior, except that it generally seems to reduce performance with EyePACS. "Best so far" refers to the baseline plus all previously effective ablations listed in Table A.6

Dataset	Model	Best So Far	+Augmult(4)	+Augmult(8)
EyePACS	SN + CNN	66.35 $\pm$ 0.12	65.69 $\pm$ 0.42	65.36 $\pm$ 0.19
EyePACS	B/16 + L	69.52 $\pm$ 0.08	69.04 $\pm$ 0.05	68.99 $\pm$ 0.11
EyePACS	B/16 + TLNN	69.55 $\pm$ 0.06	69.24 $\pm$ 0.08	69.35 $\pm$ 0.2
EyePACS	G/14 + L	72.04 $\pm$ 0.03	71.45 $\pm$ 0.37	71.36 $\pm$ 0.24
EyePACS	G/14 + TLNN	73.07 $\pm$ 0.05	73.07 $\pm$ 0.05	73.02 $\pm$ 0.07
CheXpert	SN + CNN	82.11 $\pm$ 0.29	81.41 $\pm$ 0.3	81.39 $\pm$ 0.41
CheXpert	B/16 + L	78.42 $\pm$ 0.07	78.58 $\pm$ 0.04	78.65 $\pm$ 0.1
CheXpert	B/16 + TLNN	78.86 $\pm$ 0.1	78.91 $\pm$ 0.04	78.97 $\pm$ 0.12
CheXpert	G/14 + L	81.76 $\pm$ 0.05	82.04 $\pm$ 0.14	81.9 $\pm$ 0.19
CheXpert	G/14 + TLNN	81.8 $\pm$ 0.06	81.06 $\pm$ 0.04	81.03 $\pm$ 0.1
CIFAR-10 (ACC)	SN + CNN	68.96 $\pm$ 0.26	69.07 $\pm$ 0.2	69.16 $\pm$ 0.08
CIFAR-10 (ACC)	B/16 + L	99.75(94.57)	99.76(94.68)	99.77(94.76)
CIFAR-10 (ACC)	B/16 + TLNN	99.76(94.55)	99.78(94.75)	99.79(94.81)

Table A.3: Studying the impact of normalization for CLIP ViT models, we find that normalization generally improves performance, but the effectiveness depends on the architecture and dataset. Normalizations marked in red indicate a drop in performance compared to the baseline.

Dataset	Model	Baseline	DataNorm	GroupNorm8	GroupNorm16
EyePACS	B/16 + Linear	65.44 ± 0.04	68.2 ± 0.04	69.28 ± 0.06	69.52 ± 0.08
EyePACS	B/16 + TLNN	67.89 ± 0.06	68.86 ± 0.03	69.55 ± 0.06	69.41 ± 0.04
EyePACS	G/14 + Linear	66.54 ± 0.04	72.04 ± 0.03	72.01 ± 0.04	71.78 ± 0.01
EyePACS	G/14 + TLNN	70.30 ± 0.13	72.66 ± 0.03	72.1 ± 0.17	72.12 ± 0.31
CheXpert	B/16 + Linear	76.05 ± 0.19	78.31 ± 0.04	78.35 ± 0.05	78.03 ± 0.1
CheXpert	B/16 + TLNN	78.21 ± 0.07	78.42 ± 0.07	78.5 ± 0.02	78.86 ± 0.1
CheXpert	G/14 + Linear	79.32 ± 0.08	81.6 ± 0.01	81.55 ± 0.05	81.18 ± 0.2
CheXpert	G/14 + TLNN	81.80 ± 0.06	80.48 ± 0.33	81.37 ± 0.1	81.47 ± 0.17
CIFAR10	CLIP + Linear	99.64(93.91)	99.76(94.41)	99.75(94.43)	99.75(94.57)
CIFAR10	CLIP + TLNN	99.69(94.10)	99.74(94.15)	99.74(94.36)	99.75(94.51)

Table A.4: Studying the impact of normalization for ScatterNet + CNN, we observe that normalization consistently improves performance. Data normalization tends to outperform group normalization for EyePACS and CIFAR-10, particularly due to the large gradients of their Scatter features.

Dataset	Baseline	DataNorm	GroupNorm9	GroupNorm27	GroupNorm81
EyePACS (AUC)	57.1 ± 0.38	66.35 ± 0.12	60.34 ± 0.22	62.41 ± 0.24	63.68 ± 0.15
CheXpert (AUC)	79.15 ± 0.26	80.74 ± 0.62	79.53 ± 0.28	80.99 ± 0.08	80.72 ± 0.35
CIFAR-10 (Acc)	55.18 ± 0.28	68.29 ± 0.17	65.97 ± 0.13	66.26 ± 0.11	66.45 ± 0.45

Table A.6: Studying all ablations together, we observe an almost consistent improvement in performance for CIFAR-10, whereas this pattern is not observed with the other datasets.

Dataset	Model	Best Norm	+Larger Batch	+WS	+Best Augmult	+EMA
EyePACS	SN + CNN	66.35	66.24 $\pm$ 0.23	65.81	65.69 $\pm$ 0.42	66.65 $\pm$ 0.16
EyePACS	B/16 + L	69.52	68.95 $\pm$ 0.24	-	69.04 $\pm$ 0.05	69.09 $\pm$ 0.12
EyePACS	B/16 + TLNN	69.55	69.16 $\pm$ 0.28	-	69.24 $\pm$ 0.08	69.3 $\pm$ 0.21
EyePACS	G/14 + L	72.04	71.59 $\pm$ 0.24	-	71.45 $\pm$ 0.37	71.29 $\pm$ 0.17
EyePACS	G/14 + TLNN	72.66	73.07 $\pm$ 0.06	-	73.07 $\pm$ 0.05	73 $\pm$ 0.03
CheXpert	SN + CNN	80.99	81.54 $\pm$ 0.34	82.11	81.41 $\pm$ 0.3	82.24 $\pm$ 0.22
CheXpert	B/16 + L	78.35	78.42 $\pm$ 0.07	-	78.65 $\pm$ 0.1	78.65 $\pm$ 0.04
CheXpert	B/16 + TLNN	78.86	78.86 $\pm$ 0.1	-	78.97 $\pm$ 0.12	79.01 $\pm$ 0.1
CheXpert	G/14 + L	81.6	81.76 $\pm$ 0.05	-	82.04 $\pm$ 0.14	82.01 $\pm$ 0.05
CheXpert	G/14 + TLNN	81.8	81.04 $\pm$ 0.16	-	81.06 $\pm$ 0.34	81.14 $\pm$ 0.26
CIFAR10	B/16 + L	94.57	99.74(94.49)	-	99.77(94.76)	99.76(94.67)
CIFAR10	B/16 + TLNN	94.51	99.76(94.55)	-	99.79(94.81)	99.78(94.76)
CIFAR-10	SN + CNN	68.29	66.47 $\pm$ 0.31	68.96	69.16 $\pm$ 0.08	68.07 $\pm$ 0.24