Beyond the Privacy-Utility Tradeoff: Fairness in Differentialy Private Stochastic Gradient Descent

***Abstract*—As the use of machine learning models for decision-making that impacts personal, societal, and economic outcomes increases, ensuring fairness in these models has become a crucial issue. The addition of privacy techniques to machine learning models can have unintended consequences for fairness, particularly for certain groups. This paper evaluates the impact of differential privacy techniques on fairness at both the group and subgroup levels. While the overall accuracy of the model decreased, as expected with the addition of privacy techniques, fairness was not necessarily equal for all groups. Our sub-group analysis revealed the exacerbation of model bias when trained with differentially private stochastic gradient descent. Our results highlight the importance of evaluating fairness before and after applying privacy mechanisms to a dataset to ensure fair outcomes for all individuals and groups.**

Introduction

Institutions, companies, and governments increasingly rely on datasets collected from individuals to make important decisions. These datasets often leverage machine learning (ML) and other artificial intelligence (AI) techniques that influence critical policy and decision-making processes. For instance, hiring, legal decisions, and financial lending all rely on data-driven algorithms that can have significant social and economic impacts [1]. The US Government also uses rich datasets and methods to decide on funding for schools, the number of national and state legislative seats, and the distribution of welfare assistance. Given the large social, economic, and political consequences, it is essential that these decisions are made accurately. However, this requirement conflicts directly with the need to protect individuals from potential privacy threats, which can introduce some error into the properties of groups to make it more difficult to distinguish individuals from one another. For the sake of privacy, error could distort important decisions, such as schools losing funding, legislative seats being lost, or loans being denied, with far-reaching consequences for the affected individuals and communities [2].

Specifically, research in recent years has cast doubts and concerns about ML fairness, where researchers discovered significant accuracy disparities across demographic groups in healthcare systems[3], among many others. Often it is the most vulnerable populations that shoulder the burden of inaccurate and unfair models.

The balance of accuracy, fairness, and privacy is an important subject as automated decision making becomes interwoven more and more in public, private, and legal institutions. This project aims to measure how the paradigm for privacy models, differential privacy, impacts the fairness of groups using healthcare data.

Background

A. Definitions of fairness

While there are clear benefits for algorithmic decision making, algorithms, like the people that designed them, are vulnerable to biases that result in unfair decisions [4]. In the context of machine learning, outcomes are not influenced, skewed, or biased towards sensitive attributes. There are several nuanced definitions and interpretations of fairness in machine learning that Mehrabi et. al. explain in detail [4], including Equalized Odds, Treatment / Test Equality, Fairness through Awareness / Unawareness, and Conditional Statistical Parity, that cover fairness throughout the development, application, and interpretation of algorithms. For the purposes of this project, fairness is measured by Equal Opportunity. While other definitions of fairness can and should be applied and evaluated in privacy models, this project's scope defines fairness through the lens that members of protected and unprotected groups should have equal true positive rates in the outcome of the model [4].

Broadly, there are three categories of fairness: individual, sub-group, and group fairness. While there are notable studies in fairness that have explored differential privacy at individual [5] and subgroup levels [6], this project focuses on the impact of differential privacy at the group level. Group fairness means that the model does not favor one group over another based on their membership in a protected class, such as race or gender, and that any statistical property should be similar to that of the entire population. Group fairness criteria are typically defined based on statistical measures such as demographic parity or equal opportunity.

B. Primary fairness issues with differential privacy

Differential privacy aims to protect the privacy of individuals by making it difficult for an observer to identify an individual in a dataset, while fairness seeks to ensure that individuals are treated equally and without bias [1]. There are arguments in literature that differential privacy is aligned with fairness. For example, Dwork et. al argue that fairness is a generalization of differential privacy [5].

However, while differential privacy provides robust privacy guarantees on the released data, recent studies have shown that it can introduce biases and fairness issues in downstream decision-making processes [7]. Differential privacy alters the

statistical properties of the data and this can ultimately affect an algorithm's accuracy and impact. Differentially private stochastic gradient descent (DP-SGD) is often used in deep-learning, however it can increase bias in outcomes towards the most popular elements of the distribution being learned. Bagdasaryan demonstrated that applying DP-SGD can lead to bias in image and text analysis compared to models without DP-SGD, resulting in decreased accuracy. DP-SGD can add additional noise and alter statistical properties so that the output of a model is more inaccurate [6]. Pujol et. al also demonstrated that when stricter privacy constraints (i.e. a smaller ϵ) are applied, the noise added to achieve privacy may disproportionately impact some groups over others [2].

Additionally, bias is particularly more pronounced using complex datasets when applying privacy techniques. As mentioned, Bagdasaryan showed how natural-language processing and image datasets can become more unfair with the application of differential privacy [6]. With the growing integration of image and voice recognition systems in everyday use, privacy models could increase and amplify existing bias towards communities disproportionately. As algorithms are used more commonly in high stakes situations, such as judicial and banking institutions, it is critical that accuracy, privacy, and fairness are equally considered [4].

C. The privacy-fairness trade-off

Fairness is an increasingly important consideration in machine learning and privacy-preserving models, and it is a growing topic in the literature on these subjects. Differentially private stochastic gradient descent (DP-SDG) is often used in deep-learning, however it can increase bias in outcomes towards the most popular elements of the distribution being learned.

Farrand et. al. demonstrated that when stricter privacy guarantees are applied to imbalance datasets, demographic parity and equality of opportunity metrics of fairness are reduced [6]. Bagdasaryan et. al. also show that there is a disproportionately large reduction in accuracy for subgroups when deep learning models are trained with DP-SGD [6].

This project builds on these works as well as the work conducted by Suriyakumar et. al., who evaluated the effects of privacy techniques in predicting outcomes in healthcare data [8]. The authors of this paper used the MIMIC-III dataset for binary and multi-class predictions. The authors found that using DP-SGD is not suited for healthcare because the model reduces the utility of underrepresented classes and can create unequal health outcomes for different population groups. This project evaluates the difference in fairness rather than utility in health outcomes on a different dataset.

METHOD

D. Data Source

This study used the CDC's COVID-19 Case Surveillance Public Use Data dataset accessed in April 2023. The dataset includes demographic information, exposure history, disease severity, and health outcomes of 97 million patients in the US [9].

E. Data Preparation

To get the data ready for a machine-learning model, we cleaned up the formats, imputed missing data, cleaned and encoded categorical values as one-hot vectors, and standardized the numerical features. It was then split into training, validation, and test set using a 70-15-15 ratio. The sets were then passed to a custom dataloader to make them compatible with PyTorch.

F. Data Modeling

DP-SGD is a commonly used privacy-preserving technique in machine learning. Training with DP-SGD limits the amount of information that can be extracted about any individual data point. This ensures that a model trained on a particular dataset is virtually indistinguishable from a model trained on the same dataset with a single point change. To achieve this, machine learning models are often trained using gradient descent, which involves taking small steps to minimize an error function. To prevent a single data point from significantly altering the model, two operations can be added to each training step. These include gradient clipping, which involves limiting the size of the gradient, and adding random noise to the gradient. These two techniques have a disproportionate effect on underrepresented classes when used with DP-SGD [10].

In order to evaluate the impact of DP-SGD on fairness, we compared the fairness measurements and accuracy metrics of a logistic regression model that predicted the likelihood of death with DG-SGD applied and without. We developed two versions of our logistic regression: a non-private logistic regression model and a private logistic regression model. Both models included an SGD optimizer and cross entropy loss parameters. The private logistic regression model also included privacy parameters inside a differential privacy engine from Opacus, which was inspired by the work by Yousefpour et al. [3]

For the private model, we additionally trained on different values of ϵ to measure the difference in impact for varying levels of privacy. All the other experiments were evaluated for worst-case with the lowest degree of privacy.

G. Fairness Evaluation

In our evaluation of the private and non-private models, we used Equal Opportunity metrics to assess fairness. Specifically, we measured the range of True Positive Rate (TPR), False Positive Rate (FPR), and Accuracy across groups defined by race and sex of patients in our dataset. We compared these metrics between the private and non-private models in order to determine whether the addition of DP-SGD helped to improve fairness in the model's predictions.

Our subgroup analysis included sex (male and female), age, and race.

RESULTS

H. Effects on Overall Accuracy

Overall, the addition of DP-SGD slightly improved the model's accuracy in predicting the likelihood of a patient's death (Figure 1). This was expected from the privacy-utility trade-off that differential privacy techniques affect the accuracy of the model.

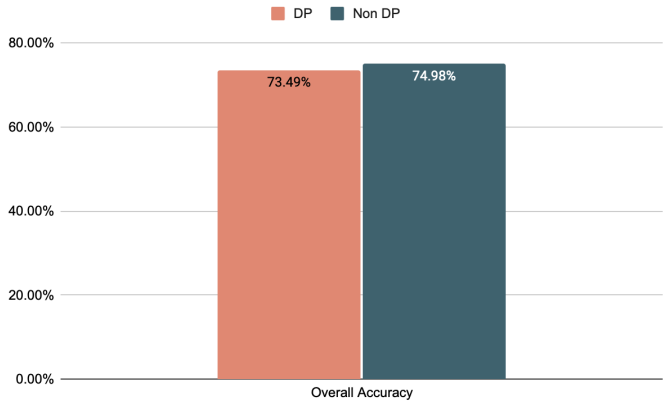

Figure 1. Effect on overall accuracy

I. Effects on Group Fairness

While the overall accuracy of the model was slightly improved in the private model compared to the non-private model, accuracy and fairness within groups and sub-groups is less straightforward. After running our models, fairness for sex improved (Figure 2). The accuracy of TPR, FPR, and Accuracy all improved in our model after applying DP-SGD. However, the same was not true when we ran our model on Race, where fairness became worse across all fairness metrics after we implemented DP-SGD (Figure 3).

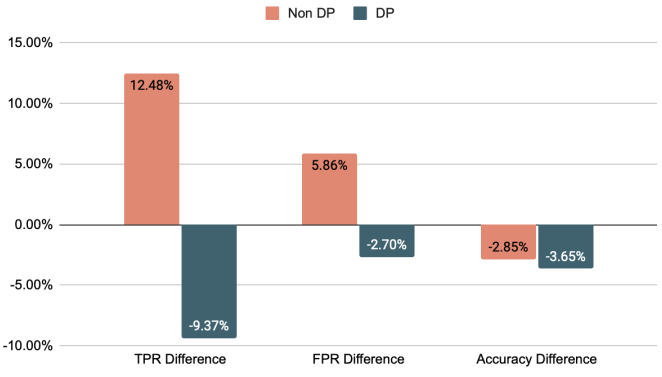

Figure 2. Overall effect on Sex

We see that within groups, fairness is also different across sub-groups. When looking at the sub-groups in Sex, the TPR decreased for the protected group (non-male), while the TPR for males improved (Figure 4). Similarly, the TPR for Black

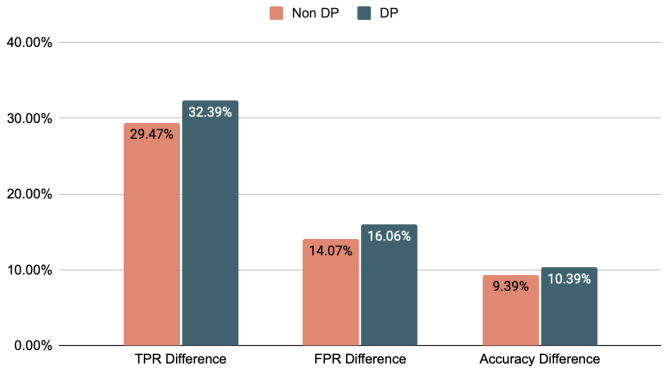

Figure 3. Overall effect on Race

and Hispanic / Latino patients worsened after applying DP-SGD compared to White patients (Figure 5).

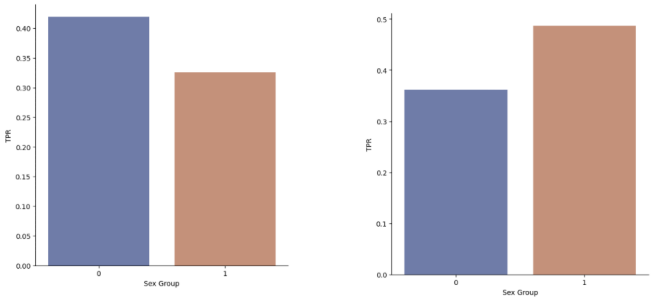

Figure 4. Effect on subgroups in Sex

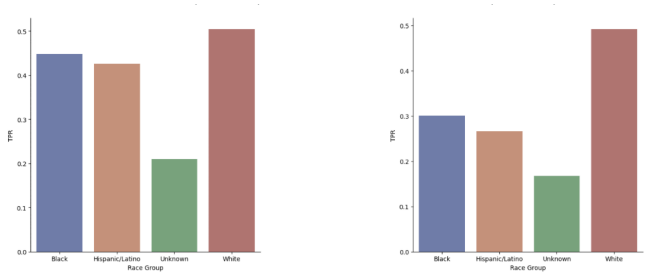

Figure 5. Effect on subgroups in Race

We discovered that the effects of improving privacy on fairness differ for various groups. Specifically, increasing ϵ-privacy resulted in a greater increase in the False Negative Rate (FNR) for Black and Hispanic/Latino patients compared to White patients (Figure 6). Furthermore, this increase was more pronounced for Black patients than for other groups, implying that the higher the level of privacy, the greater the negative impact on fairness, especially for groups most marginalized.

DISCUSSION

While the overall accuracy of the model improved, analyzing fairness at the group and subgroup levels revealed that

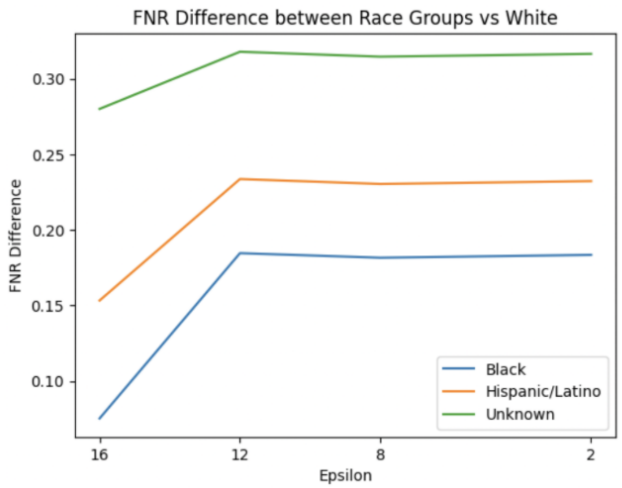

Figure 6. Effect of changing privacy on Race

accuracy and fairness were not necessarily equal for all. As machine learning models are increasingly used to make decisions that impact personal, societal, and economic outcomes, it is crucial that the addition of privacy techniques does not disproportionately worsen outcomes for certain groups over others. While we acknowledge that the results of our comparisons may be influenced by the dataset distribution and attributes, our project highlights the importance of evaluating fairness before and after applying modifications to such a dataset to ensure fair outcomes for all individuals and groups regardless of their representation in data. Future research could involve comparing fairness outcomes on different models or datasets with different data distributions.

Although unfairness and bias in DP models are still not fully understood [1], there are recommendations and considerations that can be applied to improve fairness when using differential privacy techniques. Our results highlighted the importance of evaluating fairness at both the group and subgroup levels. While we observed an overall improvement in fairness for the Sex group, a closer look at the subgroup level revealed that non-male patients were more negatively affected compared to male patients. Although complete fairness may be unattainable when using privacy techniques [11], privacy engineers should consider adding fairness constraints to their privacy model to penalize unfairness while optimizing for privacy [12]. Additionally, it is crucial to identify and address bias and unfairness early in the algorithm development process through a Privacy through Awareness framework [5].

We also found that increasing fairness for all groups at the same rate can disproportionately reduce fairness for some groups, emphasizing the importance of including similar evaluations when designing for privacy (Figure 6).

This points towards the fact that the solution may not simply be to decrease privacy for underserved subgroups as those are the groups most at risk of privacy disclosures. Ultimately, balancing privacy and fairness requires a nuanced understanding of the complex interactions between data, technology, and human values. It is important to approach privacy and fairness as complementary goals that should be achieved together.

REFERENCES

[1] F. Fioretto, C. Tran, P. Van Hentenryck, and K. Zhu, "Differential privacy and fairness in decisions and learning tasks: A survey," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022.

[2] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness in machine learning," *arXiv preprint arXiv:1610.02413*, 2016.

[3] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, and I. Mironov, "Opacus: User-friendly differential privacy library in pytorch," *arXiv preprint arXiv:2109.12298*, 2021.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021. [Online]. Available: https://doi.org/10.1145/3457607

[5] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and machine learning," *Fairness, Accountability, and Transparency in Machine Learning*, pp. 161–187, 2018.

[6] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf

[7] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "A survey of bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[8] V. M. Suriyakumar, N. Papernot, A. Goldenberg, and M. Ghassemi, "Challenges of differentially private prediction in healthcare settings," in *Proceedings of the IJCAI 2021 Workshop on AI for Social Good*, 2021.

[9] CDC, "Covid-19 case surveillance public use data," https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf, accessed in April 2023.

[10] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *CCS*, 2016.

[11] H. Chang and R. Shokri, "On the privacy risks of algorithmic fairness," in *2021 IEEE European Symposium on Security and Privacy (EuroSP)*, 2021, pp. 292–303.

[12] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern, "On the compatibility of privacy and fairness," in *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 2019, pp. 309–315.