Exploring the Accuracy and Privacy Tradeoff in AI-Driven Healthcare Through Differential Privacy

Surabhi Nayak[1][a] and Sara Nayak[2][b]
[1]*Privacy Engineering, Google, New York, NY, U.S.A*
[2]*School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, U.S.A*

Keywords: Artificial Intelligence, Algorithmic Fairness, Differential Privacy, Prevent User Harm, Privacy by Design, Technology in Healthcare.

Abstract: With the increased integration of emerging AI capabilities into the healthcare landscape, the potential for user privacy violations, ethical concerns and eventual harm to the users are some of the foremost concerns that threaten the successful and safe adoption of these capabilities. Due to these risks - misuse of this highly sensitive data, inappropriate user profiling, lack of sufficient consent and user unawareness are all factors that must be kept in mind to implement 'privacy-by-design' when building these features, for a medical purpose. This paper aims to look at the top-most privacy and ethical concerns in this space, and provides recommendations to help mitigate some of these risks. We also present a technical implementation of differential privacy in an attempt to demonstrate how the addition of noise to health data can significantly improve its privacy, while retaining its utility.

1 INTRODUCTION

As medicine continues to evolve, its integration with Artificial Intelligence (AI) holds tremendous transformative power to enhance patient care, clinical decision-making, and overall healthcare outcomes. With its ability to analyze vast amounts of medical data, identify patterns and generate insights, AI offers medical practitioners innovative tools to navigate the complexities of modern healthcare. From diagnostic accuracy and personalized treatment plans to administrative streamlining and drug discovery, the application of AI in healthcare has countless possibilities.

However, the powerful convergence of AI and healthcare also raises ethical considerations regarding bias mitigation, patient data privacy, informed consent, algorithm transparency and equitable distribution of AI-enhanced healthcare services. By carefully understanding these ethical complexities, healthcare professionals and technologists can harness the growing potential of AI to advance patient care while upholding the values that define compassionate and responsible medical practice. This involves rethinking and restructuring the standard principles of AI algorithm deployment by prioritizing the alleviation of privacy and ethical concerns.

The purpose of this article is to explore some of these ethical considerations accompanying the integration of AI in the field of medicine, specifically - algorithmic fairness and privacy.

2 ALGORITHMIC FAIRNESS AND PRIVACY

The goal of algorithmic fairness is to ensure that the outcomes, decisions, and recommendations produced by AI systems do not perpetuate or exacerbate existing biases or disparities present within healthcare systems. This goal is inherently complex as it involves subjectivity in the definition of fairness. It leads to certain important concerns while designing AI systems for healthcare, such as - 'What should fairness mean?', 'Is ensuring fairness with respect to an individual the same as ensuring fairness with respect to a group?' A group refers to a set of individuals who share a common characteristic such

[a] https://orcid.org/0009-0004-0803-1423
[b] https://orcid.org/0009-0006-8821-1794

Nayak, S. and Nayak, S.
Exploring the Accuracy and Privacy Tradeoff in AI-Driven Healthcare Through Differential Privacy.
DOI: 10.5220/0013309500003899
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025) - Volume 1*, pages 349-354
ISBN: 978-989-758-735-1; ISSN: 2184-4356
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

as race, gender, economic background, demography, geography etc. Furthermore, ‘Even if the perfect notion of fairness is found, how should it be enforced?’

An obvious question that comes to mind is - why do standard machine learning techniques, when deployed directly, lead to outcomes which are unfair? While there are multiple explanations for the same, there are some which are widely known. Chouldechova et.al discuss several causes of unfairness in their work (Chouldechova and Roth, 2020). Firstly, bias could be encoded in the data. Consider an AI model designed to diagnose skin diseases from medical images, such as photographs of rashes or lesions. The model is trained on a dataset containing images of patients from various sources, including hospitals and clinics. In this scenario, an example of bias being encoded in the data could be the overrepresentation of lighter skin tones in the training data, leading to a situation where the AI model's predictions are unfair and less accurate for individuals with darker skin tones.

Secondly, different groups can have significantly different distributions. Next, it is possible that features are less predictive on some groups as opposed to other groups. Consider an AI model designed to predict heart disease risk in patients based on various health indicators. The model is trained on a diverse dataset that includes individuals from different demographic groups, including both men and women. Imagine a scenario in which a specific health indicator ‘X’ is more strongly correlated with heart disease risk in men compared to women. Due to the stronger correlation between health indicator ‘X’ and heart disease in men, the model might assign a higher risk score to a woman with elevated levels of ‘X’, even if other risk factors for heart disease are less significant in women. Therefore, the unequal predictive strength of certain features for different groups—stronger for men compared to women—has led to a situation where the AI model's predictions are less accurate and fair for female patients.

Lastly, it is possible that some groups are inherently less predictable. Consider an AI model trained on a diverse dataset, designed to assist in diagnosing mental health disorders. It is commonly known that individuals from culturally distinct groups may express symptoms of mental health disorders in ways that are not well-captured by standardized assessments. Cultural norms, beliefs, and communication styles can significantly influence how symptoms manifest and are reported. Thus, the inherent unpredictability of symptom expression among culturally distinct groups can lead to a situation where the AI model's predictions are less reliable and equitable.

Some of these scenarios can have relatively ‘simple’ solutions such as collecting more representative data or including features which are more predictive on all groups, which in itself is a challenging and expensive process. While some algorithms such as bolt-on postprocessing methods (introducing randomization to ensure fairness) have been proposed by researchers, other scenarios are more complex to solve and are still open areas of research.

Since it is evident that there is a need to augment standard principles of AI algorithm deployment to account for algorithmic fairness, we revisit the process of defining ‘fairness’. Kearns et.al state that, based on the vast majority of work done on fairness in machine learning, various definitions of fairness can be divided into two broad categories: statistical definitions and individual definitions (Kearns, Neel, Roth and Wu, 2019). Statistical definitions focus on fixing a small number of protected groups (such as race) and defining fairness as the equality of a statistical measure across all the subgroups (Asian, Hispanics, African-American - not an exhaustive list) in the identified group. An example of such a statistical measure in healthcare could be the False Positive Rate of a medical diagnosis. Fairness, in this case, would mean that the probability of mispredicting the presence of a disease should be approximately equal across all subgroups.

Individual definitions of fairness focus on satisfying each person’s perspective of fairness. Algorithmically, this can be viewed as a constraint satisfaction problem in which each person’s perspective of fairness is a constraint which must be satisfied while we improve the performance of our AI model. Satisfying individual definitions of fairness is an open research question because it does not scale well. This means that the feasibility of simultaneously solving each of these fairness constraint satisfaction problems reduces as the number of individuals involved increases.

Till date, more research has been done on the statistical definitions of fairness due to its comparatively lower complexity and simpler validation. The first step is to identify which groups or attributes we wish to ‘protect’ when we deploy our algorithm. By protect, we mean that we want to identify which are the vulnerable or minority groups in our dataset. The next step focuses on defining what constitutes ‘harm’ in a system. As an example, in case of medical diagnosis, harm with respect to fairness could be a higher misprediction of the absence of a disease in a certain group in a population (referred to as False Negative Rate). Comprehending a definite notion of harm should be an essential component of the medical problem statement for which an AI solution is being developed.

There are some challenges to statistical definitions of fairness. First, what makes achieving fairness challenging is the subjectivity involved in defining ‘protected groups’ and ‘harm’. Second, the concept of intersectionality - when a person can belong to more than one minority subgroup (such as race: Asian and gender: Female), adds to the complexity of the problem as now the definition of fairness must hold over all subgroups the individual belongs to. Third, there are certain cases where violating statistical definitions does not necessarily mean unfairness. For example, in shared decision-making scenarios, patients' preferences play a significant role in treatment choices. AI algorithms might need to prioritize recommendations based on patient preferences even if it leads to varied outcomes across different groups.

Exploring algorithmic fairness in healthcare AI has revealed an essential crossroads where technology and ethics meet. By acknowledging the nuanced facets of fairness, we can strive to innovate more responsibly.

The healthcare industry generates an enormous amount of patient data. AI-driven algorithms and models excel at extracting meaningful insights from this large amount of data. However, utilizing patient data such as medical records, images, genetic information and wearable device data, for research can lead to data leakage and loss of privacy of the patient. Simple aggregation or de-identification of patient data does not suffice as multiple data sources can be linked to re-identify data related to a patient. The concept of differential privacy tackles this very issue. Dwork and Roth formally define Differential Privacy as: A randomized algorithm M with domain $N|X|$ is (ε, δ)-differentially private if for all $S \subseteq \mathrm{Range}(M)$ and for all $x, y \in N|X|$ such that $\|x - y\|1 \leq 1$:

$$\Pr\,[M(x) \in S] \leq \exp(\varepsilon)\,\Pr\,[M(y) \in S] + \delta \tag{1}$$

where the probability space is over the coin flips of the mechanism M. If $\delta = 0$, we say that M is ε-differentially private (Dwork and Roth, 2014).

It aims to protect the sensitive information of individuals while allowing useful insights to be extracted from data. It provides a way to ensure patient data used in medical research and analysis remains private.

Once health data or health-related data comes into scope, the privacy risk profile of a system or product increases exponentially. As a result, health data is often classified as sensitive personally identifiable information (SPII). Simple raw data of a person (like resting heart rate, disease condition, exercise history), can be used to infer sensitive medical information about a user once this data is input into an algorithm. The table below shows a summarization of the study by Ribeiro et.al about how health-related information can be derived from something as simple as device sensors (Ribeiro, Singh and Guestrin, 2016):

Table 1: Summarization of methods used to derive healthcare information from mobile sensors.

Derived Data	Raw data and combinations from sensors
Demographics	Motion - can determine gender by gait with 94% accuracy
	Touchscreen - can distinguish child vs adult with 99% accuracy
	Network, Location - obtained marital status and state of residence with 80% accuracy
Activity and Behavior	Motion - classified drinking behavior of young adults using nightlife physical motion with 76% accuracy
	Network, Location - determined whether the user was standing, walking, or using other transportation with 97% accuracy
Health Parameters and Body Features	Motion - estimated the continuous BMI value from the accelerometer and the gyroscope data with a maximum accuracy of 94.8%
	Touchscreen - determined if a person has Parkinson disease by analyzing their keystroke writing pattern with accuracy of 88%
	Network, Location - identifying periods of depression using geolocation patterns acquired from mobile phones of individuals with 85% accuracy
Mood and Emotion	Motion - determined mood with 81% accuracy
	Touchscreen - Based on keystroke metadata and accelerometer data, they reported a 90.31% prediction accuracy on the depression score
	Network, Location - Recognized the composite emotions (happiness, sadness, anger, surprise, fear, disgust) of users with 63% accuracy

This means that a lot of innocuous and irrelevant seeming data, once ingested into an algorithm could end up revealing a lot more about a user, creating a profile on the user and even uniquely identifying a user. Given the scale at which users generate data today (specifically on their devices), we can classify data as structured or unstructured. The approach to protect both these types of data, before and during its usage as a training dataset for an algorithm, can be described as follows (Delgado-Santos, Stragapede, Tolosana, Guest, Deravi and, Vera-Rodriguez, 2022):

Table 2: Data Protection Approaches categorized by Type of Data.

	Structured data	Unstructured data
Examples	Fingerprint, location, weather parameter, physiological signals, personal attributes	Face images, activity signals, biometrics
Data Modification	Traditional data modification techniques work well with structured data	Machine-learning-based data modification techniques work better
Privacy Enhancing Mechanisms	Perturbation - replacing with added noise for location data Aggregation - compression algorithms for HR (Yang, Zhu, Xiang and Zhou , 2018) Sampling - based on conditional probability distribution like gender K-anon on server-side and synthetic data on device-side (Ren, Wu and Yao, 2013)	Differentially private stochastic gradient descent (DP-SGD) already exist in practice, DP based auto-encoders can be used for biometrics (Liu, Chen, Zhou, Guan and Ma, 2019) Generative Adversarial Network (GAN) to sanitize motion data (Phan, Wang, Wu and Dou, 2016) Semi-adversarial network (SAN) to sanitize faces and selective obfuscation (Boutet, Frindel, Gambs, Jourdan, 2021)

Consider an example where a large set of patient health records are used for medical research. With differential privacy, before this data is released or used, a controlled amount of noise or randomness is added to the data in a way that makes individual patient information indistinguishable. This means that any specific patient's information is hidden within the noise. It's important to note that achieving the right balance between privacy and data utility (accuracy of results) requires careful parameter tuning.

Therefore, understanding differential privacy can help medical professionals appreciate the importance of safeguarding patient information while still contributing to medical advancements through responsible data sharing and analysis.

3 METHODS AND RESULTS

An illustrative technical implementation has been performed to evaluate the accuracy-privacy tradeoff, which is the balance between the accuracy of data analysis and the level of privacy provided to individuals whose data is being used. Enhancing privacy often involves adding noise to the data, aggregating data, or using encryption techniques, which can reduce the accuracy of the analysis or model. Conversely, maximizing accuracy typically requires more detailed and precise data, which can compromise individual privacy.

Data related to various health parameters of about 68,000 patients is analyzed and preprocessed to build two binary classification models, which categorize patients based on the presence or absence of cardiovascular disease. Both models utilize a Neural Network architecture, the difference being in the optimizers used to train the models. The first model is trained with the Adam Optimizer and the second model is trained with the Differentially Private Adam Optimizer. The Adam Optimizer computes individual adaptive learning rates for each parameter based on the estimates of first and second moments of the gradients (Kingma and Ba, 2014). The Differentially Private Adam Optimizer is a variant of Adam which includes gradient clipping and noise addition to ensure that individual training examples remain private.

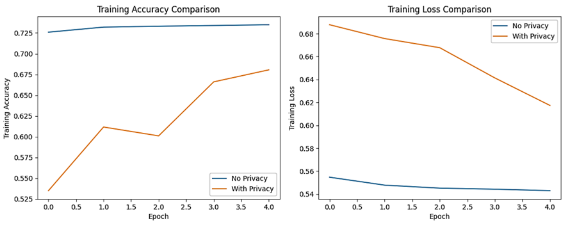


Figure 1: Comparison of Training Accuracy and Loss of Models Trained with and without Differential Privacy Implementation.

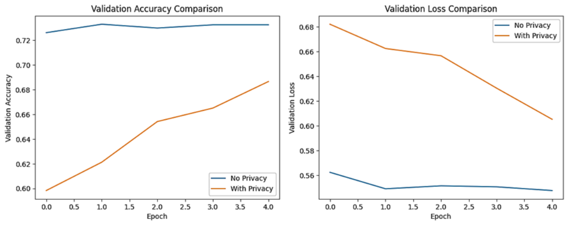


Figure 2: Comparison of Validation Accuracy and Loss of Models Trained with and without Differential Privacy Implementation.

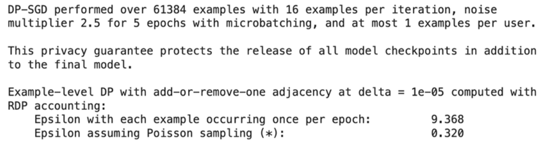
The table below (Table 3) shows the comparison of test accuracy and test loss between models trained with and without privacy considerations.

Table 3: Comparison of performances of model 1 and model 2.

Performance Indicator	Model 1 - Neural Network with Adam Optimizer	Model 2 - Neural Network with DP Adam Optimizer
Test Accuracy	0.729	0.691
Test Loss	0.548	0.602

By comparing the test performance of Model 1 (without privacy implementation) and Model 2 (with privacy implementation) is highly comparable. The accuracy of Model 2 is less than that of Model 1 by 0.038 and the loss of Model 2 is greater than that of Model 1 by 0.054.

To quantify the privacy loss in the algorithm, the privacy budget is analyzed as shown in the figure (Fig. 3) below. The privacy budget, often denoted by epsilon (ε) measures the strength of the privacy guarantee by bounding how much the probability of a particular model output can vary by including/excluding a single training point.. Based on the tiers of privacy stated in the paper (Ponomareva, Hazimeh, Kurakin, Xu, Denison, McMahan, Vassilvitskii, Chien and Thakurta, 2023), the currently undocumented but commonly implemented aim for DP-ML models is to achieve an ε ≤ 10 to provide a reasonable level of anonymization. The value of ε under the assumption that each data point is used exactly once per epoch in the training process is 9.368.



```
DP-SGD performed over 61384 examples with 16 examples per iteration, noise
multiplier 2.5 for 5 epochs with microbatching, and at most 1 examples per user.

This privacy guarantee protects the release of all model checkpoints in addition
to the final model.

Example-level DP with add-or-remove-one adjacency at delta = 1e-05 computed with
RDP accounting:
    Epsilon with each example occurring once per epoch:          9.368
    Epsilon assuming Poisson sampling (*):                       0.320
```

Figure 3: Privacy guarantee generated by performing DP-SGD over data.

4 CONCLUSION

In summary, integrating AI and medicine promises a groundbreaking journey ahead. However, this convergence requires careful consideration of its privacy and ethical ramifications. The study described in this paper, evaluates the balance between data analysis accuracy and privacy protection, using health data from 68,000 patients to create two neural network models for predicting cardiovascular disease. One model is trained with the Adam Optimizer, while the other uses the Differentially Private Adam Optimizer to ensure individual data privacy. Performance comparisons reveal that the privacy-enhanced model has a slightly reduced accuracy (by 0.038) and increased loss (by 0.054). The privacy budget, quantified by epsilon (ε), achieves a value of 9.368, indicating a reasonable level of anonymization according to commonly accepted standards.

As research continues and even though several questions are yet to be answered, it is certain that striking a balance between innovation and safety is imperative. Safeguarding patient privacy and ensuring fairness are not just checkboxes; they define the conscientious application of AI in medicine. By weaving ethics into the AI-medical narrative, we ensure that progress and compassion walk hand in hand, paving the way for a future where cutting-edge technology and unwavering medical ethics coexist harmoniously.

REFERENCES

Chouldechova A, Roth A. The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810. 2018 Oct 20.

Kearns M, Neel S, Roth A, Wu ZS. An empirical study of rich subgroup fairness for machine learning. InProceedings of the conference on fairness, accountability, and transparency 2019 Jan 29 (pp. 100-109).

Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science. 2014 Aug 10;9(3–4):211-407.

Delgado-Santos P, Stragapede G, Tolosana R, Guest R, Deravi F, Vera-Rodriguez R. A survey of privacy vulnerabilities of mobile device sensors. ACM Computing Surveys (CSUR). 2022 Sep 10;54(11s):1-30.

Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the predictions of any classifier. InProceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 1135-1144).

Yang M, Zhu T, Xiang Y, Zhou W. Density-based location preservation for mobile crowdsensing with differential privacy. Ieee Access. 2018 Mar 19;6:14779-89.

Ren J, Wu G, Yao L. A sensitive data aggregation scheme for body sensor networks based on data hiding. Personal and Ubiquitous Computing. 2013 Oct;17:1317-29.

Liu C, Chen S, Zhou S, Guan J, Ma Y. A novel privacy preserving method for data publication. Information Sciences. 2019 Oct 1;501:421-35.

Phan N, Wang Y, Wu X, Dou D. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. InProceedings of the AAAI Conference on Artificial Intelligence 2016 Feb 21 (Vol. 30, No. 1).

Boutet A, Frindel C, Gambs S, Jourdan T, Ngueveu RC. DYSAN: Dynamically sanitizing motion sensor data against sensitive inferences through adversarial networks. InProceedings of the 2021 ACM Asia conference on computer and communications security 2021 May 24 (pp. 672-686).

Mirjalili V, Raschka S, Ross A. PrivacyNet: Semi-adversarial networks for multi-attribute face privacy. IEEE Transactions on Image Processing. 2020 Sep 21;29:9400-12.

Rigaki M, Garcia S. A survey of privacy attacks in machine learning. ACM Computing Surveys. 2023 Nov 10;56(4):1-34.

Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. InProceedings of the 2016 ACM SIGSAC conference on computer and communications security 2016 Oct 24 (pp. 308-318).

Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014 Dec 22.

Ponomareva N, Hazimeh H, Kurakin A, Xu Z, Denison C, McMahan HB, Vassilvitskii S, Chien S, Thakurta AG. How to dp-fy ml: A practical guide to machine learning with differential privacy. Journal of Artificial Intelligence Research. 2023 Jul 23;77:1113-201.