Tutorial on Fair and Private Deep Learning

Manisha Padala
manishap@iisc.ac.in
Dept. of CSA
IISc, Bangalore
India

Sankarshan Damle
sankarshan.damle@research.iiit.ac.in
Machine Learning Lab
IIIT, Hyderabad
India

Sujit Gujar
sujit.gujar@iiit.ac.in
Machine Learning Lab
IIIT, Hyderabad
India

Figure 1: A fair and private neural network to overcome bias (ProPublica) and privacy leak [18]

ABSTRACT

Deep Learning (DL) finds application in several prominent fields, including computer vision, natural language processing, and bioinformatics. The proliferation of DL-based methods has brought to notice critical issues about bias (or unfairness) in classification and weak privacy guarantees of the training data. It is crucial to prioritize addressing these issues to prevent the potentially significant negative impact on users. While there has been progress, majority of the works focus on independently resolving fairness and privacy. We propose a tutorial on "Fair and Private Deep Learning" – aimed to provide an exhaustive discussion on (i) reasons behind unfair classifications and lack of privacy, (ii) fairness notions in literature and methods to ensure them, (iii) differentially private DL, and (iv) algorithms that address fair and private DL simultaneously. Moreover, in this tutorial, we not just limit our attention to classical, centralized DL models but also to the fairness and privacy challenges in distributed (or federated) DL.

CODS-COMAD 2024, January 4–7, 2024, Bangalore, India

ACM ISBN 979-8-4007-1634-8/24/01…$15.00
https://doi.org/10.1145/3632410.3633294

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

Deep Learning, Federated Learning, Group Fairness, Differential Privacy

ACM Reference Format:
Manisha Padala, Sankarshan Damle, and Sujit Gujar. 2024. Tutorial on Fair and Private Deep Learning. In *7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD) (CODS-COMAD 2024), January 4–7, 2024, Bangalore, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3632410.3633294

1 INTRODUCTION

With ease in accessibility of a wide variety of data and computational efficiency, deep learning (DL) algorithms have gained popularity. However, racial, gender-based, and other forms of prejudice remain prevalent in the data we collect to train such DL-based models. DL algorithms retain (and may further amplify) the data bias through predictions [4, 6, 10]. For instance, ProPublica studied the risk assessment tool widely used by the judiciary system in the USA[1]. The study showed a bias in criminal risk scores – for similar profiles, the 'black' defendants received a higher risk score

[1]From ProPublica: propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

than 'white' defendants. As such, it is imperative that DL algorithms not only focus on standard measures (e.g., accuracy) but also incorporate notions of *group fairness* [9] – which requires equal treatment of every demographic group. Governments worldwide are also adopting laws to enforce the same. E.g., USA's labor laws have introduced the 80% *Disparate Impact* rule [16].

Approaches to ensure a group fairness notion require knowledge of *sensitive attributes* such as gender or race. These attributes often comprise the most critical information, reflected in prohibiting such attributes for developing DL-based models. E.g., the EU General Data Protection Regulation prevents the collection of sensitive attributes [37]. Thus, addressing discrimination while preserving the leakage of sensitive attributes from the data samples is crucial. As such, higher accuracy and improvement in fairness must not come at the expense of the loss of *privacy* of the training data. To ensure holistic and productive solutions, the evaluation of a model must depend on (i) primary performance measures (e.g., accuracy), (ii) the fairness notion that it satisfies, and (iii) the privacy guarantees it provides.

2 TARGET AUDIENCE AND PREREQUISITES

The tutorial is relevant to the broad CODS-COMAD community, especially researchers, practitioners, and data scientists working on DL-based applications. In particular, they may benefit from (i) an overview of the fairness and privacy notions in DL and (ii) a summary and walk-through of implementations of prominent techniques for fair and private DL.

The tutorial does not require participants to know any specific prerequisites. While a basic understanding of DL-based classification will be handy, we plan to explain all the required preliminaries for the results presented.

3 TUTORIAL STRUCTURE

Overview. This tutorial discusses the following ethical concerns about DL models: (i) *fairness* of the decisions made by the model and (ii) *privacy* of the training data. First, we provide the different group fairness notions and their corresponding algorithms. We then describe how these fair algorithms may potentially reveal sensitive information from the training data. In order to provide specific guarantees against such breach of privacy, we discuss the importance of differential privacy. Next, we discuss works demonstrating the challenges of imposing fairness and privacy restrictions simultaneously. Finally, we focus on existing approaches for fair and private classifiers in two settings. The first is a centralized setting, where the entire data is available on a central server where the model is trained. The data is distributed across clients in the second setting, and a central model is trained using federated learning. In detail, the tutorial is structured as follows.

3.1 Fairness in Deep Learning

Generally, standard notions for fairness in classification tasks depend on false positive and false negative rates, i.e., FPR and FNR. More concretely, the notions depend on each demographic group a's FPR and FNR value. We informally define popular notions next for classifying a data entry as $\{x, y\}$ for a sensitive attribute denoted by set A.

Fairness Measures. First, we have *equalized odds* (EO), which ensures that a model's classification is independent of the attribute, i.e., $\text{FPR}_a = FPR$ and $\text{FNR}_a = FNR$, $\forall a \in A$. Second, *accuracy parity* (AP) states that the classification error must be equal $\forall a \in A$, i.e., $\text{FPR} + \text{FNR} = \text{FPR}_a + \text{FNR}_a$, $\forall a$. Third, with *equality of opportunity* (EOpp), we ensure that a model classifies an input with x or y with equal probability $\forall a$ or $\text{FNR} = \text{FNR}_a$, $\forall a$. Last, *demographic parity* (DemP) states that a model's classification rate of the 'positive' class equals $\forall a$.

Fair DL. The approaches fall into three broad categories for fairness in classification: i) The first body of work focuses on pre-processing, i.e., modifying the input data [13, 16, 20] to ensure fairness. Adversarial learning of fair representations to achieve DP, EO, or DI is also well known [7, 14, 26]. ii) The second body of work can be classified as in-processing approaches, where the fairness measure is introduced into the training objective. One way would be to introduce penalty functions or regularizers based on fairness measures [5, 8, 31, 40]. Zhang *et al.* [41] uses neural network-based adversarial learning to predict the sensitive attribute based on the classifier output to learn an equal opportunity classifier. Another approach is the reductionist approach, in which the task of fair classification is reduced to a sequence of cost-sensitive classification [29], and [2], which can then be solved by a standard classifier. iii) Finally, post-processing approaches modify the output obtained from a trained model to ensure fair predictions [23–25, 39].

These approaches above can also be extended to the case of Federated Learning (FL) [15, 19, 42]. In FL, clients locally train their models, which get broadcasted to an aggregator. It then uses an aggregation heuristic to derive the global model. Kanaparthy *et al.* [21, 22] note that these approaches will not readily generalize to the FL setting where the clients' data is heterogeneous. Techniques to ensure fairness in such a setting include fair aggregation heuristics [21, 22] and aggregation using Momentum [35].

3.2 Private Deep Learning

In the broad AI/ML literature, *differential privacy* (DP) has emerged as the gold standard of privacy [1, 12, 34, 36, 38]. Private DL algorithms using DP have found widespread adoption [27]. In particular, DP-SGD [1] and PATE [32] are state-of-the-art techniques for classification tasks. Approaches for ensuring DP in DL algorithms focus on perturbing inputs, gradients and prediction, subsampling, and bagging [27].

In this tutorial, we will primarily focus on DP-SGD [1] due to (i) its broad applicability and (ii) a stronger adversarial model. DP-SGD combines Stochastic Gradient Descent (SGD) with a differentially-private mechanism. The algorithm ensures that an individual's data record remains hidden through a calibrated addition of Gaussian noise to the gradients computed during each iteration. The authors also introduce the *moments accountant* – a privacy accountant that estimates the privacy budget of the training process. This estimate helps practitioners set appropriate (hyper) parameters to achieve the desired level of privacy and utility (e.g., the model's accuracy).

3.3 Fair and Private Deep Learning

Works that concurrently study fairness and privacy in DL are limited [3, 11, 28, 37]. The authors in [3] show that privacy has a

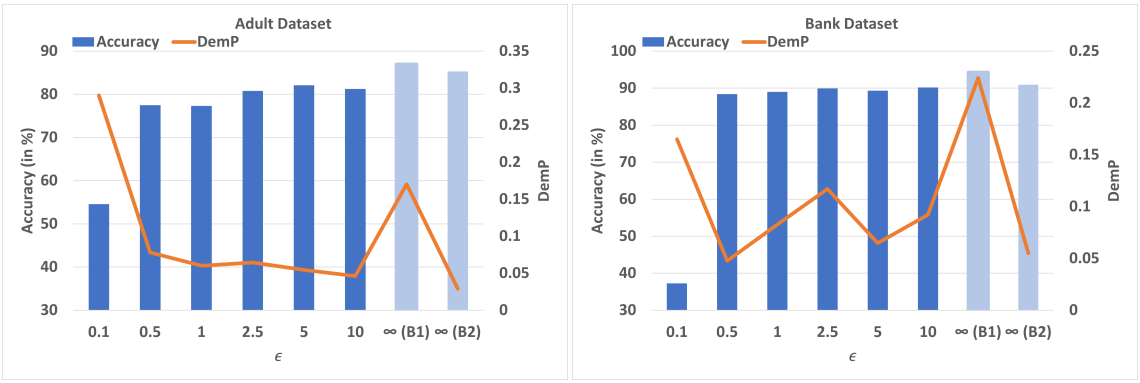

Figure 2: The trade-off between Demographic Parity, Differential Privacy and Accuracy for Adult and Bank Datasets [30]

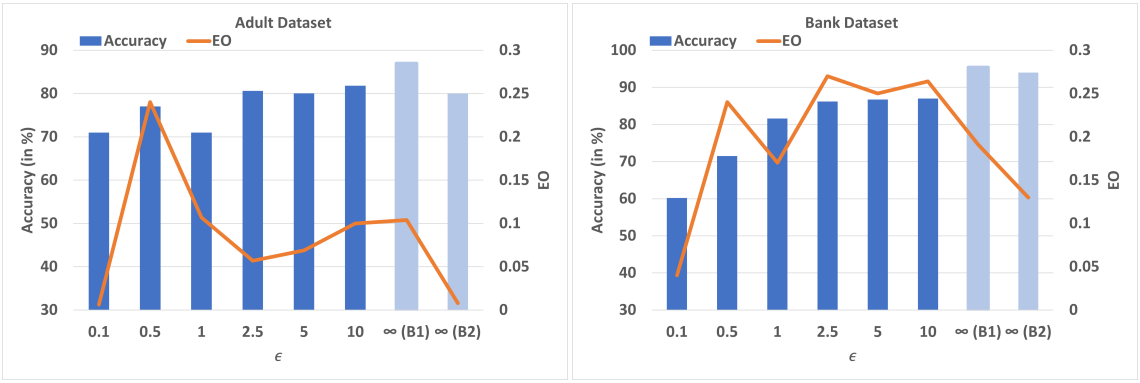

Figure 3: The trade-off between Equalized Odds, Differential Privacy and Accuracy for Adult and Bank Datasets [30]

negative impact on a model's fairness. Tran et al. [37] look at the confidentiality of the sensitive attribute. The authors only add noise to the gradients from the fairness loss to not comprise the resulting accuracy. The implication is that their approach only preserves the sensitive attribute and not the training data. FPFL [30] decouples the training process by first improving fairness followed by ensuring DP, thus protecting both – the sensitive attribute and the training data. Some of the trade-offs we discuss in FPFL are presented in Figure 2 and Figure 3.

For further details, we refer the reader to a recent survey on fair and private DL [17].

3.4 Implementation Walkthrough

We intend to demonstrate intuitive online tools for fairness and privacy, including:

- What-If (pair-code.github.io/what-if-tool) by Google
- AI-Fairness 360 (ai-fairness-360.org)
- Diffprivlib (github.com/IBM/differential-privacy-library) by IBM
- PINQ (github.com/LLGemini/PINQ) by Microsoft

Additionally, we plan to walk through the codes of specific state-of-the-art algorithms for all three of the above categories. More concretely, for fair DL, we intend to focus on FNNC [31], DP-SGD [1] for private DL, Fair-LD [37] and FPFL [30] for fair and private DL.

4 PROPOSERS & DISTRIBUTION OF TOPICS

Here, we present each presenter's bio, followed by the distribution of topics. All three presenters will be available to present the tutorial **in-person**.

4.1 Presenters' Bio

- **Manisha Padala** is a Post-doctoral researcher at the Indian Institute of Science (IISc), Bangalore, with the Department of Computer Science and Automation (CSA) under Prof. Siddharth Barman. She completed her Ph.D. in CSE at the Machine Learning Lab, International Institute of Information Technology (IIIT), Hyderabad, under Prof. Sujit Gujar. Her Ph.D. thesis titled, "Fairness in AI-based Decision Making," primarily addresses fairness issues in Machine Learning and Resource Allocation. Her research interests include Machine Learning, Game Theory, and Mechanism Design. She has held research internship positions at Adobe Research India and Google Research India. Her work has also received a Best Paper (Runner Up) award at PRICAI '22.
- **Sankarshan Damle** is a Ripple-IIITH Ph.D. Fellow at Machine Learning Lab, International Institute of Information Technology (IIIT), Hyderabad. He is also a Bachelor of Technology from IIIT, Hyderabad. He has held research internship positions at The Hong Kong University of Science and Technology (HKUST),

Ecole Polytechnique Federale de Lausanne (EPFL), Samsung Research India, and Microsoft Research India. His main research interests include Game Theory and Mechanism Design, Applied Cryptography, Trustworthy AI/ML, and Blockchain.

- **Sujit Gujar** is an Associate Professor at the International Institute of Information Technology (IIIT), Hyderabad. He also holds the CA Technologies Faculty Chair position at IIIT, Hyderabad. Earlier, he was a Post-doctoral researcher with Prof. Boi Faltings, LIA, EPFL, Lausanne (Jan'14-Oct'15) and a Senior Research Associate with Prof. Y. Narahari (Nov'15-Apr'16). He was a Research Scientist at Xerox Research Centre India (Jan'11-Nov'13). He completed his Ph.D. in the Department of Computer Science and Automation (CSA) at the Indian Institute of Science (IISc), Bangalore. He worked with Prof Y. Narahari at the Game Theory Lab. He is a recipient of the Alumni Medal of IISc for the Best Thesis in CSA for the academic year 2011-12 for his Ph.D. Dissertation, "Novel Mechanisms for Allocation of Heterogeneous Items in Strategic Settings". He has co-authored more than 100 research peer-reviewed articles, primarily in AI, fairness in AI, game theory, privacy, and blockchains.

Distribution of Topics. First, Sujit will give an introduction to the notions of fairness and privacy. Next, Manisha will introduce DL-based classification and discuss Fair DL. Sankarshan will then present Private DL (30 minutes), followed by Fair and Private DL and implementation.

5 LIST OF IMPORTANT REFERENCES

Our tutorial aims to cover the following papers in detail.

- Fair DL: [2, 26, 31, 33]
- Fair and Federated DL: [22, 35]
- Private DL: [1, 32]
- Private and Federated DL: [38]
- Fair and Private DL: [37] (Non-federated) and [30] (Federated)

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *ACM SIGSAC CCS*. 308–318.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *ICML*. 60–69.

[3] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. *NeurIPS* 32 (2019).

[4] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[5] Yahav Bechavod and Katrina Ligett. 2017. Learning Fair Classifiers: A Regularization-Inspired Approach. *CoRR* abs/1707.00044 (2017).

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018).

[7] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed Huai hsin Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *CoRR* abs/1707.00075 (2017).

[8] M. Bilal Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. 2015. Fairness Constraints: Mechanisms for Fair Classification. *ArXiv e-prints* (July 2015). arXiv:1507.05259 [stat.ML]

[9] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).

[10] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[11] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. 2019. On the compatibility of privacy and fairness. In *UMAP*. 309–315.

[12] Sankarshan Damle, Aleksei Triastcyn, Boi Faltings, and Sujit Gujar. 2021. Differentially Private Multi-Agent Constraint Optimization. In *WI-IAT*. 422–429.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *ITCS*. 214–226.

[14] Harrison Edwards and Amos Storkey. 2016. Censoring Representations with an Adversary. In *ICLR*. arxiv.org/abs/1511.05897

[15] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. 2021. Fairfed: Enabling group fairness in federated learning. In *NeurIPS Workshop on New Frontiers in Federated Learning (NFFL)*.

[16] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *KDD*. 259–268.

[17] Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. 2022. Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey. In *IJCAI*. 5470–5477.

[18] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC*.

[19] Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. 2022. Provably Fair Federated Learning via Bounded Group Loss. In *ICLR Workshop on Socially Responsible Machine Learning*.

[20] F. Kamiran and T. Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. 1–6.

[21] Samhita Kanaparthy, Manisha Padala, Sankarshan Damle, and Sujit Gujar. 2022. Fair federated learning for heterogeneous data. In *CODS-COMAD*. 298–299.

[22] Samhita Kanaparthy, Manisha Padala, Sankarshan Damle, Ravi Kiran Sarvadevabhatla, and Sujit Gujar. 2023. F3: fair and federated face attribute classification with heterogeneous data. In *PAKDD*. 483–494.

[23] Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *AIES*. 247–254.

[24] Pranay Lohia. 2021. Priority-based post-processing bias mitigation for individual and group fairness. *arXiv preprint arXiv:2102.00417* (2021).

[25] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. 2019. Bias mitigation post-processing for individual and group fairness. In *ICASSP*. 2847–2851.

[26] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *ICML*. 3384–3393.

[27] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh. 2020. Privacy in deep learning: A survey. *arXiv preprint arXiv:2004.12254* (2020).

[28] Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. 2020. Fair learning with private demographic data. In *ICML*. 7066–7075.

[29] Harikrishna Narasimhan. 2018. Learning with Complex Loss Functions and Constraints. In *AISTATS*. 1646–1654.

[30] Manisha Padala, Sankarshan Damle, and Sujit Gujar. 2021. Federated Learning Meets Fairness and Differential Privacy. In *ICONIP*. 692–699.

[31] Manisha Padala and Sujit Gujar. 2020. FNNC: Achieving Fairness through Neural Networks. In *IJCAI*. 2277–2283.

[32] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. 2017. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. In *ICLR*. openreview.net/forum?id=HkwoSDPgg

[33] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems 30*. 5680–5689.

[34] Kritika Prakash, Fiza Husain, Praveen Paruchuri, and Sujit Gujar. 2022. How private is your RL policy? An inverse RL based analysis framework. In *AAAI*, Vol. 36. 8009–8016.

[35] Teresa Salazar, Miguel Fernandes, Helder Araújo, and Pedro Henriques Abreu. 2023. FAIR-FATE: Fair Federated Learning with Momentum. In *International Conference on Computational Science*. 524–538.

[36] Sambhav Solanki, Samhita Kanaparthy, Sankarshan Damle, and Sujit Gujar. 2022. Differentially Private Federated Combinatorial Bandits with Constraints. In *ECML PKDD*. 620–637.

[37] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. 2021. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. In *AAAI*. 9932–9939.

[38] Aleksei Triastcyn and Boi Faltings. 2019. Federated learning with bayesian differential privacy. In *IEEE Big Data*. 2587–2596.

[39] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P Calmon. 2020. Optimized score transformation for fair classification. *PMLR* 108 (2020).

[40] Yongkai Wu, Lu Zhang, and Xintao Wu. 2018. Fairness-aware Classification: Criterion, Convexity, and Bounds. *CoRR* abs/1809.04737 (2018).

[41] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. *CoRR* abs/1801.07593 (2018). arXiv:1801.07593 http://arxiv.org/abs/1801.07593

[42] Daniel Yue Zhang, Ziyi Kou, and Dong Wang. 2020. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *IEEE Big Data*. 1051–1060.