EasyChair Preprint

№ 11619

Analyzing the Influence of Medical Imbalanced Data on Performance and Fairness in Differentially Private Deep Learning

Benladghem Rafika, Hadjila Fethallah, Merzoug Mohammed and Belloum Adam

December 24, 2023

Analyzing the Influence of Medical Imbalanced Data on Performance and Fairness in Differentially Private Deep Learning

RAFIKA BENLADGHEM
Tlemcen University
LRIT
Tlemcen, ALGERIA
rafika.benledghem@univ-tlemcen.dz

2nd FETALLAH HADJILA
Tlemcen University
LRIT
Tlemcen, ALGERIA
fethallah.hadjila@univ-tlemcen.dz

3rd MOHAMMED MERZOUG
Tlemcen University
LRIT
Tlemcen, ALGERIA
mohammed.merzoug@univ-tlemcen.dz

4th ADAM BELLOUM
University of Amsterdam Informatics Institute Amsterdam
The Netherlands
A.S.Z.Belloum@uva.nl

***Abstract*— deep learning carries a significant potential for a paradigm shift in healthcare and medicine. Unfortunately, deep learning poses privacy risks, as various inference attacks have revealed. Differential Privacy offers robust guarantees and substantial defense against privacy threats, making it a prevalent approach for privacy-preserving deep learning lately. Many recent approaches to deep learning and differentially private deep learning assume identically Distributed data, which is often not the case in real-world situations. In our study, we examine the impact of imbalanced data on differentially private deep learning. We find that imbalanced data negatively affects both the model's performance and fairness. We explore the trade-off between privacy, usefulness, and fairness. Our findings underscore the challenges of using standard deep learning algorithms in a differentially private context to achieve reliable results for underrepresented groups.**

Keywords*— *Imbalanced medical data, privacy-utility/fairness trade-off, differentially private deep learning.

I. INTRODUCTION

Deep learning, a potent subset of machine learning, has spearheaded remarkable progress across diverse domains, particularly in healthcare and medicine. However, it is essential to note that deep learning's voracious appetite for data is a fundamental characteristic of its success. Nevertheless, the widening scope of deep learning's capabilities also opens the door to increased privacy risks. Among the primary concerns are privacy breaches facilitated by attacks like membership Inference attacks [1], [2], [3], which aim to determine whether a specific patient's record is part of the dataset utilized to train the model. and Inversion attacks [4], seek to reconstruct the complete patient data using only access to an intermediate layer within the deep network. These privacy challenges underscore the intricate trade-off between the immense potential of deep learning and the paramount necessity of protecting sensitive data. Due to the sensitive nature of patient data in medical records, harnessing the full capabilities of deep learning in healthcare necessitates an inventive strategy for constructing and using deep neural networks while preserving patient privacy. Of the various approaches introduced to offer quantifiable assurances of privacy, Differential Privacy (DP) [5] stands out for its ability to furnish algorithmic assurances of privacy in the face of several types of privacy threats.

While embracing DL and DP can offer privacy assurances, they come with associated drawbacks, including computational burdens [8], performance degradation [9], and fairness implications [10]. The last two issues are particularly accentuated in the context of deep learning when dealing with imbalanced class distributions within a classification dataset.

In our research, we explore the impact of imbalanced data on differentially private learning using a deep learning model designed for binary classification. Our primary focus is on evaluating the model's utility, as measured by commonly used metrics like F1-score and accuracy, along with its fairness. Initially, we establish baseline performance for non-differentially private deep learning models by assessing both utility and fairness in the presence of class imbalances. Subsequently, we extend our experiments to differentially private deep learning, quantifying the influence of imbalanced data on the model's utility and performance.

A. Contributions

In our study, our aim is to assess how various aspects of imbalanced data affect the deep, differentially private training of models used for binary classification in the context of imbalanced medical data. We aim to empirically investigate the influence of imbalanced data within a differentially private deep learning framework.

Our research makes the following key contributions:

- Establishing a performance benchmark and investigating the trade-offs between privacy, utility, and fairness in the context of imbalanced data. We demonstrate the adverse effects of differential privacy (DP) on both the fairness and utility of both non-private and DP deep learning models.
- Examining the intricate relationship involving data distribution, privacy, utility, and fairness in differentially private deep learning. We simulate varying degrees of privacy for DP deep learning and observe that an increase in the variability of data distribution tends to have a more

pronounced adverse effect on utility and fairness, especially for underrepresented classes.

II. BACKGROUND

This section offers fundamental foundational knowledge for the primary concepts and algorithms employed in our analysis.

A. Deep learning models for classification problems

Deep learning (DL) has become a pivotal methodology for addressing various classification problems, including binary classification. It leverages artificial neural networks, which are inspired by the structure and function of the human brain, to make predictions and decisions based on data. DL encompasses a class of machine learning techniques that excel in tasks involving the classification of data into one of two categories, commonly referred to as binary classification. These models are particularly suited for scenarios where discerning between two distinct outcomes is crucial, such as spam detection, disease diagnosis, sentiment analysis, and fraud detection.

Key Components of Deep Learning:

1. **Neural Networks:** Deep learning models are built upon neural networks, which consist of interconnected layers of artificial neurons. These networks learn to recognize patterns and relationships in data through a process known as backpropagation.
2. **Deep Architectures:** The "deep" in deep learning refers to the presence of multiple hidden layers within a neural network. These deep architectures allow the model to learn complex, hierarchical representations of the input data.
3. **Activation Functions:** Activation functions introduce non-linearity into neural networks, enabling them to model complex relationships. Common activation functions include the sigmoid, ReLU (Rectified Linear Unit), and softmax functions.
4. **Loss Functions:** Loss functions measure the disparity between the predicted output and the actual target. In binary classification, commonly used loss functions include binary cross-entropy and hinge loss.
5. **Optimization Techniques:** To train deep learning models effectively, optimization algorithms like stochastic gradient descent (SGD) and Adam are employed to adjust the model's parameters iteratively.

Challenges in Deep Learning for Binary Classification:

Despite their effectiveness, deep learning models can face challenges in the context of binary classification. These challenges may include data imbalance, where one class significantly outnumbers the other, and the need to strike a balance between model performance, fairness, and interpretability.

Deep learning offers a potent framework for tackling binary classification problems by learning intricate patterns and relationships within the data. With the right architecture, optimization, and regularization techniques, deep learning models can deliver impressive results in discerning between two distinct classes, making them a valuable tool in a wide range of applications.

B. Differential Privacy

The original notion of ϵ-differential privacy (ϵ-DP) was initially proposed by [8]. Later, the same researchers introduced a revised form referred to as (ϵ, δ)-DP [6] [7]. In this adaptation, they incorporated the parameter δ as an additional element within the original definition. This adjustment was made to accommodate privacy protection in the context of the Gaussian distribution.

1. **Definition 1: (ϵ, δ)-differential privacy** [6] [7]:

   A randomized mechanism, represented as K, achieves (ϵ,δ)-differential privacy if, for any pair of neighboring data samples B and B', and for all possible outcome subsets Z within the set K:

   $$\forall \mathrm{B}, B' \in B^n, \forall Z \sqsubseteq W:$$

   $$\Pr[K(B) \in Z] \leq e^{\epsilon} \Pr[K(B') \in Z] + \delta \quad (1)$$

   Where: W is the set of all possible outputs,

   $$\delta \ll 1/|\mathrm{Z}|$$

The interpretation of mechanism K as meeting (ϵ, δ)-differential privacy implies that it attains ϵ-differential privacy with a probability of 1-δ. However, (ϵ, δ)-DP is not suitable when the privacy-sensitive set Z consists of just one element. It's crucial to emphasize that the value of δ must be exceedingly small in comparison to the size of set Z (i.e., $\delta \ll 1/|\mathrm{Z}|$) to prevent the unfavorable scenario where privacy is consistently compromised for a significant portion of the dataset represented by δ.

One of the prominent techniques for introducing differential privacy (DP) into the field of machine learning is differantially private stochastic gradient descent DP-SGD, which was introduced by Abadi et al. in 2016 [11]. The DP-SGD method operates by limiting the gradients to manage the sensitivity of the mechanism and incorporating precisely calibrated noise into the gradient values. To monitor the privacy budget expenditure, an accounting mechanism has been suggested. Additionally, other accounting methods have been put forward in existing literature, offering more stringent estimates of privacy costs, such as the Rényi-DP-based accountant [12].

There are two important concepts that we need to delve into: the notion of neighboring datasets and function sensitivity.

2. **Definition 2: Neighboring datasets:** Consider two datasets, B and B', both belonging to the dataset domain B^n. These datasets, B and B', are regarded as neighboring when they differ by a single data point. In simpler terms, B' is obtained by either adding or removing a single data point from B.
3. **Definition 3: Sensitivity [7]:** Sensitivity measures the most significant change in the output resulting from the alteration of a single data point within the database. The sensitivity of a query function, denoted as f, is expressed as:

$$\forall \mathrm{B}, B' \in B^n, \quad f: B^n \rightarrow \mathbb{R}^d$$

$$\Delta_f = \max_{\mathrm{B},B'} \| f(B) - f(B') \|_2 \quad (2)$$

where $\| \cdot \|$ denotes the $l_2$norm.

In this paper, to attain (ϵ,δ)-differential privacy, we select the Gaussian mechanism, which utilizes L2 norm sensitivity. It involves the introduction of Gaussian noise to each dimension of the output f(B):

$$\mathrm{f(B)} + \mathcal{N}(0, \Delta^2 \sigma^2 \mathbf{I}) \quad (3)$$

4. Imbalanced data in deep learning

Imbalanced data in the context of deep learning refers to a situation where the distribution of classes in a dataset is highly skewed, with one class significantly outnumbering the other(s). This imbalance can pose challenges for machine learning models, including deep neural networks, as they may have difficulty learning to accurately classify the minority class. The model tends to be biased toward the majority class, resulting in poorer performance for the minority class and potential unfairness in predictions.

Consider a medical diagnosis scenario where a deep learning model is trained to classify X-ray images as either "normal" or "disease present." In this dataset, there is a significant class imbalance, with 90% of the images being "normal" and only 10% showing "disease present." In this case, the deep learning model may perform exceptionally well in correctly identifying "normal" cases due to their abundance in the dataset. However, it may struggle to accurately detect "disease present" cases, as they are underrepresented. This imbalance can lead to life-critical errors, where the model fails to identify individuals with the disease, posing serious consequences.

III. Methodology

In this study, the experiments investigate the training of two types of models: non-private models, specifically employing the original "vanilla" Stochastic Gradient Descent (SGD), and differentially private deep learning (DP-DL) models using DP-SGD. These models are trained on two distinct medical datasets, namely, the PIMA dataset and the Breast Cancer Wisconsin dataset. The experimentation encompasses varying privacy levels and employs two different network architectures and settings for each of the two datasets.

A. Experiment setup

All experiments were carried out using Python version 3.8.16. The tests were conducted on Google Colab, which offers access to the NVIDIA Tesla GPU and CUDA compilation tools version 11.2.152. The advantage of this setup is the ability to execute code at a significantly higher rate of operations per second compared to a CPU. Two open-source libraries suitable for training deep learning models with DP are Opacus for PyTorch [13] and TensorFlow Privacy for TensorFlow [14]. Both of these libraries are compatible with CUDA. For this research, the choice was made to utilize Opacus.

A. PIMA

1) Datasets:

The objective is to predict the likelihood of a patient having diabetes using the Pima database, which contains specific diagnostic measurements. Notably, all individuals in the dataset are female, aged at least 21 years, and of Pima Indian heritage. The dataset comprises various medical predictor variables and a binary target variable, where "1" represents diabetic and "0" represents non-diabetic outcomes. The Pima database is of size (768x9), and it is characterized by an imbalanced distribution of classes as depicted in Figure 1.

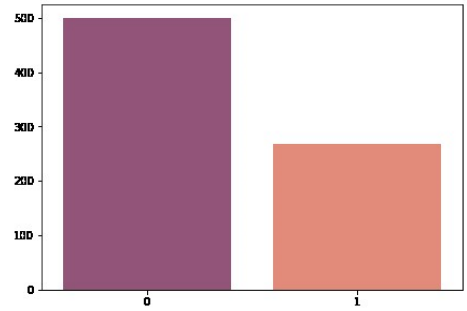

Fig. 1. Class distribution in the PIMA database.

2) Model architecture:

The model architecture can be described as follows:

- Input Layer (8 features)
- Hidden Layer 1 (20 neurons) with ReLU activation
- Hidden Layer 2 (5 neurons) with ReLU activation
- Hidden Layer 3 (5 neurons) with ReLU activation
- Output Layer (2 neurons)

This architecture is designed for binary classification tasks, such as those commonly encountered in the medical field, including the prediction of diseases like diabetes. The model takes 8 input features, passes them through the hidden layers with ReLU activations, and produces binary classification output.

B. Breast Cancer Wisconsin:

1) Datasets:

This dataset provides details regarding the attributes of cell nuclei found in images. These features are derived from digitized images of fine needle aspirates (FNA) of breast masses. For each image, the mean, standard error, and the worst or largest values (mean of the three largest) of these features were calculated, resulting in a total of 30 features. The Breast Cancer Wisconsin database is sized at (569x32), and it exhibits an imbalanced distribution, with approximately 37% of cases diagnosed as malignant (cancerous) and around 63% classified as benign (non-cancerous) as depicted in Figure 2.

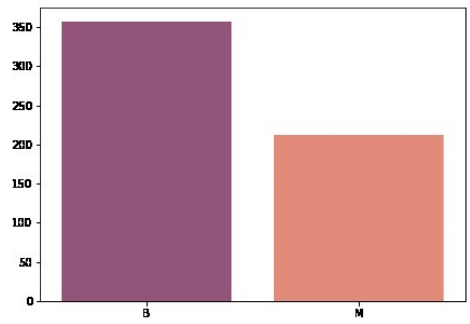

Fig. 2. Class distribution in the BREAST CANCER WISCONSIN database.

2) Model architecture:

The model architecture can be described as follows:

- Input Layer (30 features)
- Hidden Layer 1 (20 neurons) with ReLU activation
- Hidden Layer 2 (10 neurons) with ReLU activation
- Hidden Layer 3 (10 neurons) with ReLU activation
- Output Layer (2 neurons)

This architecture is designed for binary classification tasks, such as the diagnosis of breast cancer (malignant or benign). The model takes 30 input features, processes them through the hidden layers with ReLU activations, and produces binary classification output. It's a common architecture for tasks involving medical diagnosis and classification.

C. Metrics

a) **Precision (pre):** Precision measures the accuracy of positive predictions. It is the ratio of true positives to the sum of true positives and false positives. A high precision indicates a low rate of false positive predictions.

b) **Recall (rec):** Recall, also known as Sensitivity or True Positive Rate, measures the ability of the model to correctly identify positive instances. It is the ratio of true positives to the sum of true positives and false negatives. A high recall indicates a low rate of false negative predictions.

c) **Specificity (spe):** Specificity measures the ability of the model to correctly identify negative instances. It is the ratio of true negatives to the sum of true negatives and false positives.

d) **F1-Score (f1):** The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is useful when there is an imbalance between the two.

e) **Geometric Mean (geo):** The geometric mean of precision and recall is used to calculate the G-mean. It is a metric that takes into account both false positives and false negatives, making it suitable for imbalanced datasets.

f) **Index of Balanced Accuracy (iba) [15]:** The Index of Balanced Accuracy is a metric that combines the sensitivity (recall) and specificity of a classification model to provide a balanced measure of accuracy. It's particularly useful for imbalanced datasets.

These metrics are especially important when dealing with imbalanced datasets because they provide a more comprehensive evaluation of the model's performance, taking into account both the majority and minority classes.

The hyper-parameter settings for training both the private and non-private models are provided in Table 1, for both datasets.

Dataset	**Settings**	***Value***
PIMA	Learning rate	0.01
	Loss function	CrossEntropyLoss
	Batch size	64
	Epochs	500
BREAST CANCER WISCONSIN	Learning rate	0.001
	Loss function	CrossEntropyLoss
	Batch size	64
	Epochs	500

Table.1. Hyper-parameters of the model that was trained on the Datasets.

Dataset	**Settings**	***Value***
PIMA	Learning rate	0.01
	Loss function	CrossEntropyLoss
	Batch size	64
	Noise parameter(σ)	Variable
	Gradient_clipping_norm(C)	1
	Delta(δ)	10^{-4}
	Epochs	500
BREAST CANCER WISCON-SIN	Learning rate	0.001
	Loss function	CrossEntropyLoss
	Batch size	64
	Noise parameter(σ)	Variable
	Gradient_clipping_norm(C)	1
	Delta(δ)	10^{-4}
	Epochs	500

Table.2. Hyper-parameters of DP-model that was trained on the Datasets.

IV. RESULT AND DISCUSSION

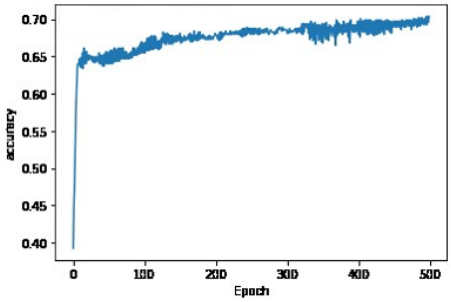

Fig. 3. Accuracy of the Non-DP Model Trained on the PIMA Dataset.

For the model trained on the PIMA dataset the model exhibits reasonable accuracy, with an overall accuracy rate of 72% as shown in Figure 3. The model trained on the Breast Cancer dataset showcases outstanding accuracy, with an overall accuracy rate of 92% as shown in Figure 4.

The difference in model performance can be traced back to the class distribution in the datasets. The PIMA dataset has a class imbalance, with a significantly larger number of samples in one class (class 0) compared to the other (class 1) – 498 samples for class 0 and 270 for class 1. In contrast, the Breast Cancer dataset features a more balanced distribution, with 357 samples for non-cancerous (class B) and 212 samples for cancerous (class M).

This class imbalance in the PIMA dataset has a notable impact on accuracy. In imbalanced datasets, models tend to perform well on the majority class (class 0 in this case) but struggle with the minority class (class 1). The model trained on the PIMA dataset is influenced by the dominance of class 0 samples, which can lead to accuracy results skewed in favor of that class. Conversely, the Breast Cancer dataset's balanced class distribution contributes to the exceptional accuracy achieved by the model trained on it.

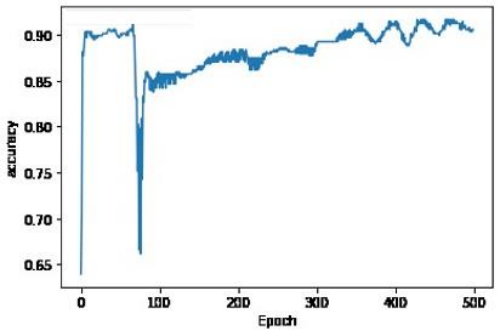

Fig. 4. Accuracy of the Non-DP Model Trained on the Breast cancer Wisconsin SET.

In our analysis, we observed distinct performance characteristics in two different datasets as shown in table 3. In the PIMA dataset, Class 0, representing non-diabetic cases, demonstrated reasonably balanced precision and recall, albeit with a lower specificity, indicating its ability to effectively classify negative cases but also raising concerns about potential false positives. Conversely, Class 1, representing diabetic cases, exhibited lower precision and recall, implying challenges in correctly identifying positive cases. While weighted metrics indicated a balanced overall performance. In contrast, the Breast Cancer dataset displayed notable performance attributes. Class 0, representing benign cases, showcased exceptional precision, recall, and specificity, signifying its proficiency in identifying negative cases. Similarly, Class 1, representing malignant cases, achieved high precision, balanced recall, and specificity, reflecting an overall commendable performance. Weighted metrics reinforced the excellence of the model on this dataset, which benefits from a more balanced class distribution. Our findings underscore the substantial impact of class distribution on model accuracy, with balanced datasets, like the Breast Cancer dataset, leading to superior performance. Achieving fairness in class representation remains essential, as it directly influences precision and recall, making the Breast Cancer model a more reliable classifier.

Results							
Non Private							
Datasets		**Pre**	**Rec**	**Spe**	**F1**	**Geo**	**Iba**
PIMA	**0**	**0.79**	**0.81**	**0.51**	**0.80**	**0.64**	**0.43**
	1	**0.55**	**0.51**	**0.81**	**0.53**	**0.64**	**0.40**
Avrage/Toral	**/**	**0.72**	**0.72**	**0.60**	**0.72**	**0.64**	**0.42**
Breast cancer Wisconsin	**B**	**0.88**	**1.00**	**0.81**	**0.94**	**0.90**	**0.82**
	M	**1.00**	**0.81**	**1.00**	**0.89**	**0.90**	**0.79**
Avrage/Toral	**/**	**0.93**	**0.92**	**0.89**	**0.92**	**0.90**	**0.81**

Table.3. Non-Private Binary Classification Model Evaluation Metrics for Pima and Breast Cancer Wisconsin Datasets.

In Table.4. The provided results illustrate how adjusting the privacy budget ε impacts the performance metrics of differentially private deep learning models trained on imbalanced datasets (PIMA and Breast Cancer). ε governs the level of noise introduced to the model updates to ensure differential privacy.

For the PIMA dataset at ε = 8, both classes exhibit reasonable precision, recall, and F1-scores, signaling a balanced performance. The lower specificity for Class 0 compared to Class 1 indicates a potential imbalance favoring Class 1. The Index of Balanced Accuracy at 0.64 suggests a relatively fair model. At ε = 2, the model's performance diminishes, resulting in lower precision, recall, and F1-scores. Substantial decreases in specificity for Class 0 highlight an imbalance in favor of Class 1, reflected in an Index of Balanced Accuracy of 0.50 a less balanced model. At ε = 1, further reduction in ε leads to significant performance deterioration, with Class 0 specificity reaching 0, indicating severe unfairness. The Index of Balanced Accuracy drops to 0.32, portraying a highly unfair model.

Results								
DP- Private								
Dataset	**(ε,δ)-dp**		**Pre**	**Rec**	**Spe**	**F1**	**Geo**	**Iba**
PIMA	**$(8,10^{-4})$-dp**	**0**	**0.81**	**0.77**	**0.51**	**0.79**	**0.63**	**0.40**
		1	**0.45**	**0.51**	**0.77**	**0.48**	**0.63**	**0.38**
	$(2,10^{-4})$-dp	**0**	**0.77**	**0.69**	**0.31**	**0.73**	**0.46**	**0.24**
		1	**0.23**	**0.31**	**0.69**	**0.27**	**0.46**	**0.21**
	$(1,10^{-4})$-dp	**0**	**0.75**	**0.63**	**0.00**	**0.68**	**0.03**	**0.01**
		1	**0.00**	**0.00**	**0.63**	**0.00**	**0.03**	**0.008**
Breast Cancer Wisconsin	**$(8,10^{-4})$-dp**	**B**	**1.00**	**0.79**	**1.00**	**0.88**	**0.89**	**0.80**
		M	**0.62**	**1.00**	**0.79**	**0.76**	**0.89**	**0.78**
	$(2,10^{-4})$-dp	**B**	**0.96**	**0.70**	**0.87**	**0.81**	**0.78**	**0.63**
		M	**0.43**	**0.87**	**0.70**	**0.57**	**0.78**	**0.60**
	$(1,10^{-4})$-dp	**B**	**0.91**	**0.63**	**0.67**	**0.74**	**0.65**	**0.43**
		M	**0.25**	**0.67**	**0.63**	**0.36**	**0.65**	**0.40**

Table.4. DP-Private Binary Classification Model Evaluation Metrics for Pima and Breast Cancer Wisconsin Datasets.

For the Breast Cancer dataset at ε = 8, both classes (B and M) demonstrate high precision, recall, and F1-scores, indicating a balanced performance. The Index of Balanced Accuracy at 0.89 suggests a well-fair model. At ε = 2, decreasing ε results in lower precision, recall, and F1-scores. Reduced specificity for Class B implies an imbalance in favor of Class M, reflected in an Index of Balanced Accuracy of 0.79—a less fair model. At ε = 1, further reduction in ε leads to significant performance deterioration, with Class B specificity decreasing to 0.67, indicating a substantial imbalance. The Index of Balanced Accuracy drops to 0.65, depicting a less fair model.

a) Impact of ε on Metrics and model quality:

Higher values of ε generally lead to better model performance, while lower values result in increased privacy and reduced utility. As ε decreases, the models become less accurate and more unfair.

b) Fairness - Distribution Trade-off:

Fairness is reflected in the balance of performance metrics across classes. in other words A well-balanced model should have similar metrics for both classes. Notably, adding

privacy has a more pronounced effect on the class with fewer samples.

c) Privacy-Utility/Fairness Trade-off:

There exists a trade-off between privacy (controlled by ε) and model utility/fairness. Smaller ε values provide stronger privacy guarantees but may sacrifice model quality.

Class distribution is a pivotal factor in model performance. The class imbalance in the PIMA dataset can result in accuracy biases, favoring the majority class. This is why the model for breast cancer performs better, benefiting from a more even class distribution. In practical scenarios, selecting an appropriate value of ε involves considering the desired level of privacy and the acceptable trade-off with model performance. Achieving a balance between privacy and utility is crucial, especially in applications where fairness is a key consideration.

V. Conclusion

In this study, we conducted a thorough empirical analysis to examine how imbalanced data influences the effectiveness and equity of differentially private deep learning models. Utilizing a real-world dataset encompassing imbalanced sampels, we delved into the intricate trade-offs involving utility, privacy, and data distribution. Our findings indicate that differential privacy (DP) generally exerts a negative influence on both the utility and performance of deep models, particularly for underrepresented classes. Moreover, we observed that imbalanced data exacerbates the disparity in utility and fairness between minority and majority classes. Our choice of metrics allowed us to comprehensively investigate the diverse aspects of imbalanced data's impact on our experiment with privacy-preserving deep models. While our results align with our initial expectations regarding the effects of DP and imbalanced data, it's important to note that this study represents a constrained exploration of the intricate interplay between fairness and utility, considering their trade-offs with privacy and data distribution. Nonetheless, our consistent findings emphasize that the presence of imbalanced data has an adverse effect on the utility and fairness of privacy-preserving deep models.

References

[1] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP), pages 3–18, 2017.

[2] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. Ml-doctor: Holistic risk assessment of inference attacks against machine learning models, 2021.

[3] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. ` Membership inference attacks from first principles. arXiv, 2021a.

[4] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pages 1322–1333, 2015.

[5] Cynthia Dwork. Differential privacy: A survey of results. In International Conference on Theory and Applications of Models of Computation, Xi'an, China, April 2008.

[6] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Advances in Cryptology—EUROCRYPT, pages 486–503, 2006a.

[7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Proc. of the Third Conf. on Theory of Cryptography (TCC), pages 265–284, 2006b.

[8] T. T. Cai, Y. Wang, and L. Zhang, "The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy," arXiv preprint arXiv:1902.04495, 2019.

[9] BLANCO-JUSTICIA, Alberto, SANCHEZ, David, DOMINGO-FERRER, Josep, et al. A critical review on the use (and misuse) of differential privacy in machine learning. ACM Computing Surveys, vol. 55, no 8, p. 1-16, 2022.

[10] T. Farrand, F. Mireshghallah, S. Singh, and A. Trask, "Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy," in Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice, pp. 15–19, 2020.

[11] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep learning with differential privacy", October, In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communica tions Security, Vienna, Austria, 2016.

[12] I. Mironov, "Renyi differential privacy," in ´ 2017 IEEE 30th Computer Security Foundations Symposium (CSF), pp. 263–275, IEEE, 2017.

[13] Facebook. Opacus: Train pytorch models with differential privacy, September 2021. [Online; posted 01-September-2021].

[14] Google. Tensorflow privacy, September 2021. [Online; posted 01-September-2021].

[15] García, V., Mollineda, R.A., Sánchez, J.S. Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions. In: Araujo, H., Mendonça, A.M., Pinho, A.J., Torres, M.I. (eds) Pattern Recognition and Image Analysis. IbPRIA 2009. Lecture Notes in Computer Science, vol 5524. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-02172-5_57. (2009).