A Study of Differentially Private Machine Learning in Healthcare

Abdur R. Shahid
Department of Computer and Information Systems
Robert Morris University, USA
ashahid@ieee.org

Sajedul Talukder
School of Computing
Southern Illinois University, USA
sajedul.talukder@siu.edu

***Abstract*—The field of Machine Learning (ML) has been engaged in intensive research for a while to build an efficient and effective intelligent system for the early identification of chronic diseases such as cancer and diabetes and has recently seen some promising findings. The bulk of the initiatives are aimed at classifying illness onset and minimizing cases of maltreatment. As a supervised learning problem, its accuracy is mostly determined by the training data, which is labeled data on actual patients that is highly privacy-sensitive. Privacy leakage can occur at any point in the machine learning lifecycle, from model training through model deployment, and can lead to a membership inference attack, model inversion attack, and reconstruction attack. As a result, safeguarding users' privacy is critical in healthcare issues, but little has been done to satisfy this demand. In this paper, we propose differential privacy-based Logistic Regression and Naïve Bayes models on breast cancer classification and diabetes prediction. We evaluate the two models using the popular Wisconsin Diagnostic Breast Cancer (WDBC) dataset, and Pima Indians Diabetes dataset and depict the privacy requirement and model accuracy trade-off.**

***Index Terms*—Machine Learning, Differential Privacy, Healthcare, Security, Privacy.**

I. Introduction

Over the last decade or so, Machine Learning (ML) has significantly changed the healthcare system. Breast cancer detection is just a single instance from a large pool of healthcare problems that have benefited fundamentally by ML[18], [14]. Breast cancer among women is one of the most common and deadliest cancers worldwide. Early diagnosis can help in timely treatment; leading to improved survival rates. ML has made the early detection of malignant tumors and the reduction in the possibility of mistreatment a reality. Similarly, Diabetes, one of the deadliest and chronic metabolic diseases occurs due to a high level of sugar in the blood over a long period. Many complications occur if diabetes remains untreated and unidentified. The risk factor and severity of diabetes can be reduced significantly if it is possible to detect diabetes at an early stage. In recent years, plenty of ML methods have been proposed and published for diabetes prediction[13]. Although ML-based solutions seem rather tempting, the vulnerabilities associated with the design of ML-based systems have not yet been fully understood. In prediction-based healthcare applications, the ML model responsible for the classification is considered to be a highly valuable intellectual property[17]. An attack on the model could seriously undermine the privacy of the patients' data that was used to train and test the model. This shows that ML models can capture information on their training data, and exploitation of the models could lead to membership inference attack[20], model inversion attack[15] and training data leakage using the model's prediction[10]. Going further, this data leakage combined with some knowledge of query distribution can even lead to a full reconstruction attack[2].

To address the privacy issues, differential privacy has gained significant attention to develop privacy-preserving machine learning models. Differential privacy-based approaches attempt to add statistical noise drawn from a probability distribution (e.g. Laplace distribution) to classifiers' parameters[23]. Formally, a randomized mechanism $\mathcal{M}$ is said to be ϵ-*differentially private* if for any two neighboring datasets D_1 and D_2 differing on at most one record, and any $S \subset Range(\mathcal{M})$, then $e^{-\epsilon} \leq \frac{Pr[\mathcal{M}(D_1) \in S]}{Pr[\mathcal{M}(D_2) \in S]} \leq e^{\epsilon}$ The sensitivity of a query q over the input datasets is computed as $q = \max ||q(D_1) - q(D_2)||$.

In this paper, we present the application of differential privacy versions of two popular and highly effective machine learning methods, Logistic Regression and Naïve Bayes, for the breast cancer classification and diabetes prediction problems. Figure 1 depicts the bird's eye view of the proposed framework for achieving differential privacy while training a machine learning model on a breast cancer dataset and a diabetes dataset. In summary, we introduce the following contributions:

- **Healthcare machine learning model vulnerabilities**. We focus on the privacy vulnerabilities for handwriting recognition by presenting two scenarios where the recognition model is trained and deployed on the cloud using the user-sent data. In both examples, the cloud can be compromised by malicious entities that can lead to leakage of sensitive personal information with serious consequences.
- **Differentially private Machine Learning**. We present a study of two differentially private machine learning algorithms, Logistic Regression, and Naive Bayes, to lay the foundation of developing privacy-preserving machine learning models for healthcare problems while preserving the privacy of the models.
- **Privacy-preserving Machine Learning models for Diabetes and Breast Cancer classification problems**. We trained four different differentially-private Logistic Regression and Naive Bayes models for Diabetes and Breast Can-

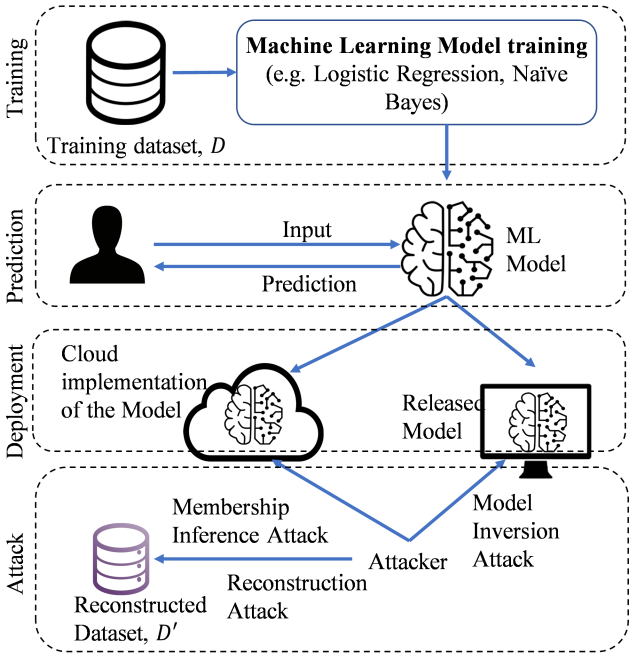

Fig. 1. Privacy-vulnerable machine learning framework in healthcare.

cer classification problems. We also trained four different non-private Logistic Regression and Naive Bayes baseline models.

Paper Outline: The rest of the paper is outlined as follows. In section II we describe the considered system model and its vulnerability against various machine learning model attacks. In section 3, we present background on the concept of differential privacy, logistics regression, and Naïve Bayes. In section 4, we present our proposed system model with differentially private logistic regression and Naïve Bayes. we describe the experimental design and the datasets, and experimental results in sections 5 and 6, respectively. Section 7 provides a guideline for integrating the research outcome into undergraduate and graduate curricula. Finally, section 8 concludes the paper with a hint to the future direction.

II. System and Threat Models

In this section, we discuss the system model, the machine learning model vulnerabilities in the system, an adversary's capabilities in exploiting such vulnerabilities, and his/her goals with the exploitation.

A. System Model

We consider a system where historical medical data are amalgamated into a database hosted on a cloud, yielding a training dataset for intelligent system design tasks (e.g. diabetes prediction and breast cancer classification). Using this training dataset, a machine learning model from a privacy-preserving machine learning algorithm (e.g. Logistic Regression, and Naive Bayes) with differentially privacy guarantee is trained and tested. Once, a stable and satisfactory model is developed, it can be deployed in two possible ways. The first deployment is hosting the model on a cloud and allow different entities remotely through queries. Machine learning as a service is such a practical cloud implementation which is gaining a lot of attention lately. The model can also be implemented on medical edge devices for local use). For example, a testing facility might have this model installed on their diabetes prediction device. The model can also be accessed through other ways, such as making it accessible for mass testing. We consider an honest-but curious attacker which has access to the both could model and edge models and is capable of some specific attacks which are discussed next. The attacker's goal from carrying out these attacks is multifaceted including reconstructing the dataset used to train the model and sell it to third parties for monetary gain. Figure 1 depicts the system and threat models considered in this paper.

B. Threat Model

We consider an honest-but-curious attackers whose goal is reconstruct the privacy-sensitive dataset used to train the model. As outlined in system model (figure 1), the attacker can carry out the following attacks on a model to achieve its goal.

1) Membership inference attack: The membership inference attack exploits the age-old weakness of a typical machine learning models: it responds differently on trained data than on unseen data. In membership inference attack, an attacker builds an attack model in a black-box manner by studying only the output of a target model. This attack model can pinpoint this weakness in a target model and use them to differentiate members from non-members of the training dataset used to train the model. In a membership inference attack, the attacker develops a binary classifier to distinguish whether a query input x is in a target dataset D. If the attacker can establish with high confidence that $x \in D$ is true or that $x \notin D$ is false, the attack is effective. In The adversary only knows the architecture of the target model and has black-box access to it. For arbitrary input x the adversary can only obtain the prediction vector $p'(y|x)$ but cannot get the intermediate computations[16], [19].

2) Model inversion attack: Model inversion attack aims to reconstruct the training data from the access to a model. Let us look at the following scenario to understand the the context of this attack: An organization develops a model using health-related data gathered from a huge number of people. They intend to distribute the model to the public for broad usage (e.g., by a medical clinic that specializes in customized treatment) after restricting access to the model within the organization (say, under some tight access control). This model can be published in two ways: either releasing it as a black box for public usage or releasing it as a white box with information on its design and parameters. A model inversion attack intends to find the correlations contained within the model which is strong enough to be exploited to recover the sensitive training data. In this attack, the attacker utilize publicly available model and additional knowledge about the individuals in the training group. The model-inversion is defined as the estimated inverse function of $f_{\theta 1}$, trained with $v = f_{\theta 1}(x)$ as input and x as output. The attack is divided into three phases: (1) creating a

training set for the model-inversion; (2) training the model-inversion; and (3) querying the model-inversion to retrieve the input sample[11].

3) Reconstruction attack: Reconstruction attacks utilizes the only the features to reconstruct the the raw private data used to train a model[5], [3]. This type of attacks require white-box access to the ML model. In other words, to make this attack successful, the features in a model must be known to an attacker. Some machine learning algorithms store feature vectors in the model itself (e.g. Support Vector Machine (SVM), KNN), making them extremely vulnerable to reconstruction attacks. It is also showed that there is a linear correlation between generalization error and probability of inferring data attributes. However, such hypothesis might work if we assume that the adversary has knowledge of the prior distribution of the target features and labels [22]. Zhang et al. [24] extended this work by relaxing the assumption and showed that high predictive power is more susceptible to reconstruction attacks.

III. Preliminaries

A. Differential Privacy

Differential privacy is a technique for publicly disclosing information about a dataset by describing the patterns of groupings within the dataset while keeping information about individual dataset members private. Differential privacy, presented by Dwork et al. in 2006[7], ensures that the same conclusions will be reached for a query on a dataset, regardless of the presence of an individual in the dataset. In other words, we call an algorithm differentially private if its output on a dataset is independent of the presence of any single tuple in the dataset. This gold standard of privacy, unlike other "medieval" privacy notions, provides privacy protection for users regardless of the prior knowledge possessed by the adversaries. A randomized algorithm $\mathcal{M}$ satisfies ϵ-differential privacy if for any two datasets D and D' (differing only in one record) and any output $S \in Range(\mathcal{M})$, satisfies:

$$e^{-\epsilon} \leq \frac{\Pr[\mathcal{M} \in S]}{\Pr[\mathcal{M} \in S']} \leq e^{\epsilon} \tag{1}$$

Here, ϵ denotes the privacy level of $\mathcal{M}$. Generally, differential privacy can be achieved by adding noise, drawn from a probability distribution, to the results of the query function. A popular way of guaranteeing differential privacy is Laplace mechanism[8]. By definition, for a query function $f : D \to \mathcal{R}$, a randomized algorithm $\mathcal{M}$ satisfies ϵ-differential privacy if

$$\mathcal{M}(D) = f(D) + Lap\Big(\frac{\mathcal{S}(f)}{\epsilon}\Big) \tag{2}$$

Here, $\mathcal{S}(f)$ is the sensitivity of f and $Lap(\frac{\mathcal{S}(f)}{\epsilon})$ is the amount of noise from Laplace distribution with center 0 and scaling $\frac{\mathcal{S}(f)}{\epsilon}$[12]. The sensitivity $\mathcal{S}(f)$ of a query f is defined as $\max_{\text{adjacent}D,D'} |f(D) - f(D')|$.

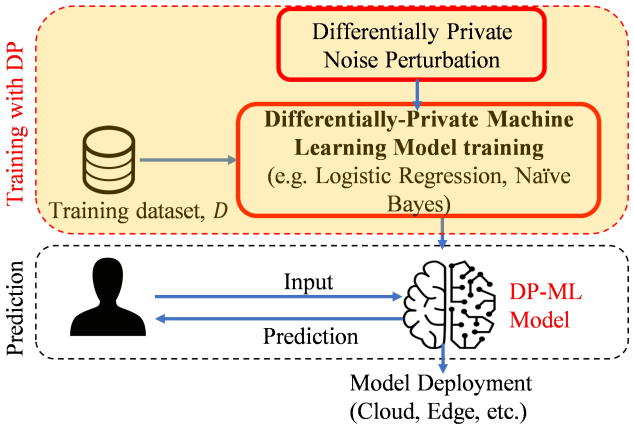

Fig. 2. The proposed system model for privacy-preserving machine learning with Logistic Regression and Naïve Bayes in Healthcare.

B. Logistic Regression

Logistic regression is a very popular statistical model used in medical research to describe data and to explain the relationship between one dependent binary variable and one or more independent variables. For a given dataset $\mathcal{D} = \{(x_i, y_i)\}_{1 \leq i \leq n}$, where $(x_i, y_i) \in \mathcal{R}^d \times \{-1, 1\}_{1 \leq i \leq n}$, logistic function computes the weighted sum of the input features and output the logistic of the results $P(y|X) = \frac{1}{1+e^{-w^T X}}$, where w and X stand for the observed outcome and the d-dimensional covariates of a sample respectively, and each dimension refers to an attribute in a dataset. The parameters w are trained by minimizing negative log-likelihood $\sum_i \log(1+e^{-y_i w^t x_i})$ over the training set.

C. Naïve Bayes

Naïve Bayes, based on the Bayes theorem, is simple but highly effective method used frequently as a baseline standard for classification tasks. Under curated tuning, Naïve Bayes is found be very accurate and used for various practical applications. Consider a given dataset $D = \left\{\left(\boldsymbol{x}^{(1)}, y^{(1)}\right), \ldots, \left(\boldsymbol{x}^{(n)}, \gamma^{(n)}\right)\right\}$ with $d+1$ attributes $X_1, X_2, \ldots, X_d, Y$ where $Y \in Y$ is the output and $y = \{c_1, \ldots, c_k\}$ is a set of responses. The Naive Bayes model is known as a classification method based on the Bayes theorem and the conditional independence assumption $P(X = x \mid Y = c_k) = \prod_{i=1}^{n} P(X_j = \boldsymbol{x}_j \mid Y = c_k)$. The Bayes theorem is used to find the output y with the largest posterior probability [12]:

$$y = \operatorname{argmax} P(Y = c_k) \times \prod_{i=1}^{n} P(X_j = \boldsymbol{x}_j \mid Y = c_k) \tag{3}$$

IV. Proposed Differentially-Private Machine Learning-based System for Healthcare

In this section, we present our proposed framework to train privacy-preserving classification model for healthcare problem. This current version of the framework utilizes differentially-private version of two very popular machine learning algorithms: logistic regression and Naïve Bayes. In the future, we will incorporate different other differentially

private algorithms to perform not only classification, but also regression and unsupervised tasks for healthcare. Figure 2 presents the workflow of the proposed system for healthcare problems. The focus of work presented in this paper is cancer and diabetes prediction. Initially, the input is the dataset. This dataset is then pre-processed that implements methods to detect missing and incorrect data, and error correction. Then, feature selection and feature transformation are done on the pre-processed dataset. The dataset is then used to train differentially-private machine learning models. The privacy parameter ϵ is tuned for each of logisitc regression and naïve bayes algorithms.

1) Differentially-private Logistic Regression: Chaudhuri and Monteleoni [4] proposed a differentially private regularized logistic regression algorithm based on solving a perturbed optimization problem. The input to our algorithm is a set of examples $x_1, \ldots, x_n$ over $\boldsymbol{R}^d$ such that $||x_i|| \leq 1$ for all i, a set of labels $y_1, \ldots, y_n$ for the examples, a regularization constant λ and a privacy parameter ϵ and the output is a vector $w^\star$ in $\boldsymbol{R}^d$. Note that for a vector x, $||x||$ denotes its Euclidean norm. The algorithm works as follows. First, it picks a random vector b from the density function $h(b) \propto e^{-\frac{\epsilon}{2}||b||}$, where the norm of b is picked from the $\Gamma(d, \frac{2}{\epsilon})$ distribution and the direction of b is uniformly random. Next, given examples $x_1, \ldots, x_n$ with labels $y_1, \ldots, y_n$ and a regularization constant λ, it computes $w^\star = \text{argmin}_w \frac{1}{2}\lambda w^T w + \frac{b^T w}{n} + \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i w^T x_i})$ and output $w^\star$. This approach solves a convex programming problem very similar to the logistic regression convex program, and therefore it has running time similar to that of logistic regression. It can be shown that, the output of this algorithm preserves ϵ-differential privacy [4].

2) Differentially-private Naïve Bayes: Vaidya et el. [21] proposed the first and one of pioneered methods of differentially private Naive Bayes classification. The central idea of their method is to derive the sensitivity of classifier parameters and utilize it to add noise (e.g. Laplace noise [9]) to guarantee differential privacy in the trained classifier. For categorical attributes, the sensitivity computation is done on the counts. For a given attribute X with r possible attribute values $x_1, \ldots, x_r$, the probability is $P(X = x_k | Y = c_j) = (n_{kj}/n)$ where n is the total number of training examples and n_{jk} is the number of the training examples that also have $X = x_k$. As the difference in the counts due a new record is simply 1, the sensitivity of each n_{kj} is 1 for all attribute values x_k and class values c_j. or numeric attributes, the probability $P\left(X = \boldsymbol{x} \mid Y = c_j\right)$ depends on mean μ_j and variance σ_j^2, hence we need to derive the sensitivity of both mean and standard deviation. Assume that the values of attribute X_j lie in the range $[l_j, u_j]$, the sensitivity for the mean is $\left(u_j - l_j\right)/(n+1)$ and the sensitivity for the standard deviation is $\sqrt{n} \times \left(u_j - l_j\right)/(n+1)$. Following this, Laplace noise is added to parameters (counts for categorical attributes, mean and standard deviation for numeric attributes) to preserve the privacy.

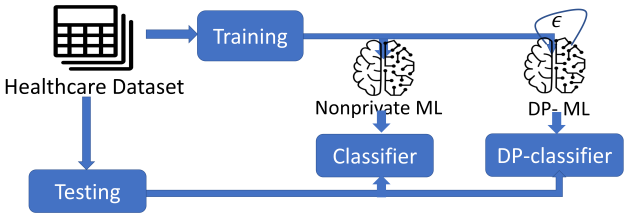

Fig. 3. Experiment Design

Parameter	Values
ϵ	$0.5, 1.0, 1.5, 2, 2.5, 3.0, 3.5, 4.0$
Data norm	$5, 7, 9, 11, 13, 15, 17, 19$

TABLE I
PARAMETERS AND THEIR VALUES USED IN THE EXPERIMENT.

V. EXPERIMENTAL DESIGN

Our experiment design includes dataset preparation, training, and testing different models on the datasets, tuning different privacy parameters to evaluate the balance between privacy and utility. For comparison purposes, we implemented four different models:

- **Non-private Naive Bayes.**
- **Non-private Logistic Regression.**
- **Differentially private Naive Bayes.**
- **Differentially private Logistic Regression.**

We used the scikit-learn and IBM diffPrivLib libraries for implementing the vanilla and differentially private models, respectively. private models, respectively. We used the R and Python programming languages for developing the scripts to create and analyze the models. We used accuracy as the primary metrics to evaluate the models. In the experiment, we evaluated two relationships:

1) Amount of privacy versus accuracy (for all the models)
2) Amount of privacy versus data norm versus accuracy (for logistic regression)

Here, data norm defines the spread of the data that is protected. In the experiment, we used the values presented in table I for privacy parameter (ϵ) and data norm. following parameters:

Since the differentially private Logistic Regression and Naïve Bayes add random Laplace noise, 20 iterations have been run in all experiments and the average was taken as the final result.

A. Dataset

In the experiment, we use two popular datasets: Wisconsin Breast Cancer Dataset [6] and Pima Indians Diabetes Dataset [1]. In these datasets, the values of different features have different ranges. Hence, normalization was applied to both datasets to transform their values into a common scale.

1) The Wisconsin Dataset: The Wisconsin dataset includes 569 instances, 32 features, and 2 targets; either malignant or benign. The dataset has no missing values. In this dataset, the features describe the characteristics of the cell nuclei present in breast mass. As described in the dataset, ten real-valued features represent the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension of each cell nucleus.

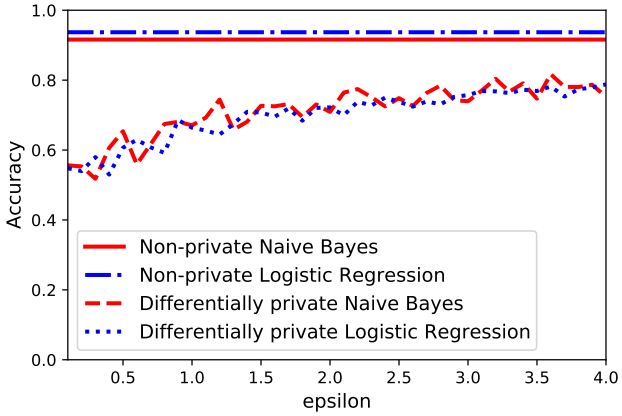

Fig. 4. Breast Cancer classification problem: Comparison among the four models (data norm in differentially-private Logistic Regression = 20)

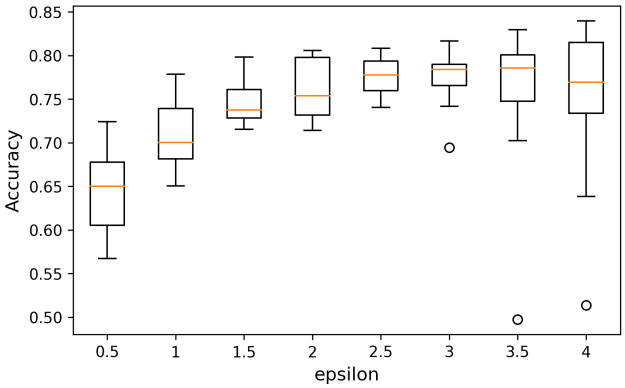

Fig. 5. Breast Cancer classification problem: Impact of the data norm in differentially-private logistic regression over accuracy for different epsilon values.

2) Pima Indians Diabetes Dataset: The dataset[1] comes from women of Pima Indian heritage having at least 21 years of age residing in or near Phoenix, AZ, and is originally a part of a larger dataset held by the National Institutes of Diabetes and Digestive and Kidney Diseases in the US. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. The dataset consists of 768 observations (500 samples from non-diabetic patients and 268 from diabetic patients) of medical details for Pima Indians patients that contain eight medical predictor variables and one target variable, Outcome. Predictor variables are all numeric attributes that include the number of pregnancies the patient has had, their plasma glucose concentration, blood pressure, BMI, triceps skinfold thickness, insulin level, diabetes pedigree function, and age. The class attribute is a binary variable that takes the values "0" or "1", indicating a negative and positive diagnosis for diabetes within 5 years of the measurements, respectively.

VI. Experimental Results

A. Results on Wisconsin Dataset

Figures 4 and 5 represent the results of the two relationships in the Wisconsin dataset. From figure 4 the difference between the private and non-private in terms of accuracy

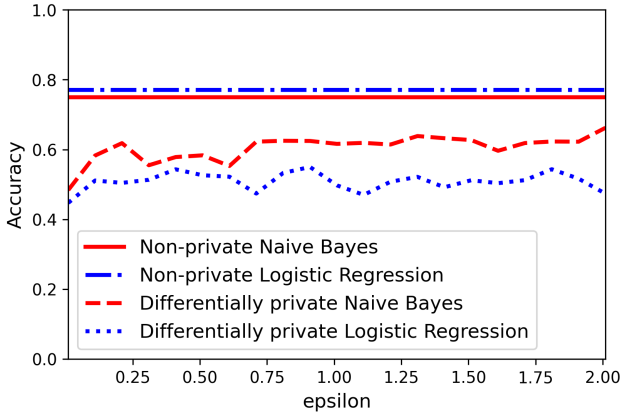

Fig. 6. Diabetes: Comparison among the four models (data norm in differentially-private Logistic Regression = 20)

is clear. The accuracy of non-private logistic regression and Naive Bayes were 90.21% and 91.61%. Not surprisingly, both non-private models achieve better accuracy (90%) than the differentially private models. For the differentially private models, with the increase in the value ϵ (in other words, the decrease in privacy), the accuracy increases. The major difference in the accuracy of private and non-private models underlines the main trade-off between privacy and utility. For instance, when ϵ is 0.1 the value of accuracy of both privacy-preserving models is $\approx 57\%$. The meandering nature of curves for differentially private Naive Bayes and Logistic Regression is due to the randomness in the noise. One possible way to control this is to continue generating noise until the accuracy of a model meets a specific threshold. It is interesting to see that, while both the non-private go side by side, a similar trend was seen for their privacy-preserving versions, regardless of the values of ϵ. However, the rate of growth seems to be slowed down with the increase in value of ϵ. In our future work, we will investigate this phenomenon to increase the accuracy of the privacy-preserving models. In differentially private Logistic Regression, there is an impact of data norm on accuracy. The result presented in figure 5 has some randomness for ϵ's values 3-4, as the accuracy dropped for some values of data norm. However, it is important to note the data norm is usually set based on the domain expertise, not at random. Yet, this result should be helpful in deciding which value to pick to maintain higher accuracy, if given multiple options.

B. Results on Pima Indian Diabetes Dataset

We present our experimental study on the two relationships for the diabetes classification problem in figures 6 and 7. Similar to the comparison among the four different models in the breast cancer classification problem, both non-private models achieve higher accuracy than the privacy-preserving models. Here, both non-private models' accuracy is less than 80%. However, the main difference here we observe in figure 6 is the correlation between ϵ and the accuracy of the two differentially private models. Both models do not show significant improvement in accuracy with the change in the value of ϵ. For

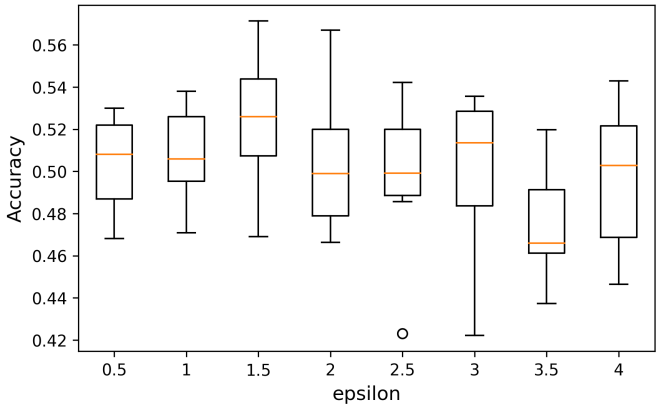

Fig. 7. Diabetes: Impact of the data norm in differentially-private logistic regression over accuracy for different epsilon values.

example, for $\epsilon = 0.1$ and $\epsilon = 2.0$, the accuracy of differentially private Logistic Regression are almost same. There is a scope for investigating this problem further to improve the accuracy of the differentially private models.

VII. Conclusion

In this paper, we have developed a differentially private Naïve Bayes and Logistic Regression classifiers for breast cancer classification and diabetes prediction. We have tested both classifiers on the real world datasets and the results show that it is possible to achieve high accuracy with both models, compared to baseline models. The future direction of this work should look at how different feature engineering methods, such as data augmentation, can improve the accuracy of the models, and explore how to use differentially private deep learning to develop more robust classifiers.

References

[1] Pima indians diabetes dataset: UCI machine learning repository, 2017.

[2] David Cash Akshima, Francesca Falzon, Adam Rivkin, and Jesse Stern. Multidimensional database reconstruction from range query access patterns. *IACR Cryptol. ePrint Arch*, 2020:296, 2020.

[3] Mohammad Al-Rubaie and J Morris Chang. Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy*, 17(2):49–58, 2019.

[4] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[5] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. PODS '03, page 202–210, New York, NY, USA, 2003. Association for Computing Machinery.

[6] Dheeru Dua and Casey Graff. Breast cancer wisconsin (diagnostic) data set: UCI machine learning repository, 2017.

[7] C. Dwork. Differential privacy. In *ICALP*, 2006.

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[9] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[10] 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.

[11] 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*, CCS '15, page 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery.

[12] Maoguo Gong, Yu Xie, Ke Pan, Kaiyuan Feng, and Alex Kai Qin. A survey on differentially private machine learning. *IEEE Computational Intelligence Magazine*, 15(2):49–64, 2020.

[13] Md Kamrul Hasan, Md Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan. Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, 8:76516–76531, 2020.

[14] R. Hazra, M. Banerjee, and L. Badia. Machine learning for breast cancer classification with ann and decision tree. In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0522–0527, 2020.

[15] Zecheng He, Tianwei Zhang, and Ruby B Lee. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 148–162, 2019.

[16] Hongsheng Hu, Zoran Salcic, Gillian Dobbie, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *CoRR*, abs/2103.07853, 2021.

[17] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 399–414. IEEE, 2018.

[18] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):1–12, 2019.

[19] 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. IEEE, 2017.

[20] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 2019.

[21] J. Vaidya, B. Shafiq, A. Basu, and Y. Hong. Differentially private naive bayes classification. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 571–576, 2013.

[22] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

[23] Farzad Zafarani and Chris Clifton. Differentially private naive bayes classifier using smooth sensitivity. *arXiv preprint arXiv:2003.13955*, 2020.

[24] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 253–261, 2020.