Evaluating Optimal Differentially Private Learning - Shallow and Deep Techniques

Geetha Peethambaran, Chandrakant Naikodi, Suresh Lakshmi Narasimha Setty

Abstract: ***Data analytics is an evolving arena in today's technological evolution. Big data, IoT and machine learning are multidisciplinary fields which pave way for large scale data analytics. Data is the basic ingredient in all type of analytical tasks, which is collected from various sources through online activity. Data divulged in these day-to-day activities contain personal information of individuals. These sensitive details may be disclosed when data is shared with data analysts or researchers for futuristic analysis. In order to respect the privacy of individuals involved, it is required to protect data to avoid any intentional harm. Differential privacy is an algorithm that allows controlled machine learning practices for quality analytics. With differential privacy, the outcome of any analytical task is unaffected by the presence or absence of a single individual or small group of individuals. But, it goes without saying that privacy protection diminishes the usefulness of data for analysis. Hence privacy preserving analytics requires algorithmic techniques that can handle privacy, data quality and efficiency simultaneously. Since one cannot be obtained without degrading the other, an optimal solution that balances the attributes is considered acceptable. The work in this paper, proposes different optimization techniques for shallow and deep learners. While evolutionary approach is proposed for shallow learning, private deep learning is optimized using Bayesian method. The results prove that the Bayesian optimized private deep learning model gives a quantifiable trade-off between the privacy, utility and performance***

Keywords: Bayesian, Deep Learning, Privacy, Private Learning, Shallow Learning

I. INTRODUCTION

Data is the new lifeline. Technological progress has paved way for an electronic era and online activity has been a constant source of huge data generation. E-commerce applications, medical research, banking sectors and various business and educational sectors collect data in various forms. Businesses and organizations use the collected data for extracting useful insights, patterns and relationships from data which are in turn used for decision-making in the long-run. Data outsourced for analysis by organizations may contain personal details of end-users or data subjects. These may be at risk of disclosure when used for data analytics and hence privacy of end-users is compromised. Privacy preserving analysis of data is a field that is witnessing extensive research with the popularity of big data analytics. Analytics of data is intended to produce beneficial results to the analysts, but it comes at the cost of affecting privacy of the individuals who are part of the data under consideration. In order to protect privacy, while allowing analytics, numerous privacy models have been proposed and experimented in literature. k-anonymity [1] is one of the extensively studied privacy models for enhancing privacy of data. While research in k-anonymity based privacy preservation has shown to render acceptable results, the application of the algorithm in the big data scenario is challenging mainly due to two reasons. One is the optimization of the privacy-performance trade-off and the other is the ability of the privacy algorithm to sustain different attack scenarios. Nevertheless, the algorithm is still being experimented by researchers in a number of applications. Along similar lines, differential privacy [2] is an algorithm that has proven to provide strong mathematical guarantees for privacy protection. The algorithm proposed by Dwork[2] in the year 2006 has seen many improvements and is now being widely accepted by companies like Google [3], Apple [3] etc… in providing privacy to user's information. The technique has gained foothold in the recent times due to its feasibility in practical implementations [4]. Two such libraries DiffPrivLib [5] and Tensorflow privacy [6] developed by IBM and Google for differentially private machine learning are popular in the recent times .The work in this paper proposes optimization of differentially private learning. The impact of private learning is evaluated using optimized shallow and deep learning mechanisms on datasets containing sensitive data. The key challenge addressed in the paper is the quantification of privacy-utility-performance trade-off in privacy preserving big data analysis.

II. BACKGROUND

A. Differential Privacy

Differential Privacy was first proposed by DWork [2] in the year 2006. The work suggests that a data subject's(any individual entity) personal information involved in analysis, will not be harmfully affected despite the presence of other sources of data available in the form of externally available information or from data involved in other studies. The work is based on the underlying fact that analysis should allow useful learning about a group and nothing personal about an individual. The aim is to learn same facts about a group of individuals irrespective of a target individual's participation in the data

B. Definition of Differential Privacy

A randomized algorithm A gives ε-differential privacy if for all data sets D and D′ differing on at most one row, and any S ⊆ Range (A)[2]

$$\frac{\Pr[A(D) \in S]}{\Pr[A(D') \in S]} \leq exp^{\varepsilon}$$

Revised Manuscript Received on July 20, 2020.
***** Correspondence Author
Geetha Peethambaran*, Computer Science, Cambridge Institute of Technology, Bangalore, India. E-mail: geetha.cse@cambridge.edu.in
Chandrakant Naikodi, Department of PG Studies and Research Centre Davangere University, Bangalore, India. E-mail: nadhachandra@gmail.com
Suresh Lakshmi Narasimha Setty,, Computer Science, Cambridge Institute of Technology, Bangalore, India. E-mail: suriakls@gmail.com

Three quantities that are of importance in the private learning are:

- **Epsilon for privacy (ε):** A metric to quantify privacy leak with change in data. Smaller values of epsilon indicate better privacy protection.
- **Utility**: The correctness of the differentially private output. For example, if analytic task is classification, then accuracy of classification can be used to quantify utility.
- **Performance**: Additionally, the computational efficiency of a private learning algorithm is determines its performance.

A differentially private learner governs the trade-off between privacy (P), utility (U) and performance (P) for any analytic task. In the rest of this paper, this trio combination is called PUP trade-off.

C. Terminology

Table 1: Nomenclature

GNB	Gaussian Naïve Bayes
LR	Logistic Regression
NSGA-II	Non-dominated Sorting Genetic lgorithm[12]
PUP	Privacy. Utility, Performance
MOOP	Multi-objective Optimization
GNB	Gaussian Naïve Bayes
LR	Logistic Regression
DPGNB	Differentially Private Gaussian Naïve Bayes
DPGLR	Differentially Private Gaussian Logistic Regression
DPDNN	Differentially Private Deep Neural Network
acc	Classification Accuracy
h	Initial Hyperparameter search space set
h_{opt}	Optimal Hyperparameter set
X	Input space
BO	Bayes Optimization

D. Dataset Description

The data for the current work adult [7] and diabetes [7] is obtained from the UCI machine learning repository[7] .The adult dataset is a benchmark tabular multivariate dataset used in the testing of a number of privacy approaches. The dataset contains 48000 instances, with information about education, employability and family status of individuals along with sensitive attributes such as race and salary. Diabetes is a clinical dataset containing around 1.5 lakh instances with patient data having personal information, medicinal doses and disease condition. Appropriate feature engineering is done to extract suitable features for the working of the algorithm.

III. PROPOSED METHODOLOGY

A. Differentially Private Learning

Private Learning is a multi-objective optimization problem [8]. Multi-objective optimization (MOOP)[9] involves solution strategies for complex problems which contain one or more conflicting objectives. Such problems cannot generate a single solution that simultaneously achieves all the stated goals. In the context of private learning, privacy, quality of learning and performance are conflicting, since maximizing privacy minimizes quality of learning and vice-versa. Besides, a complex strategy that achieves both, to a certain degree, may compromise on performance. Hence, the primary focus of this work is to empower efficient private learning, without influencing the isolation of any individual involved in the data. Privacy preservation is implemented using ε –differential privacy. Differential privacy works by adding noise to data. More noise implies less useful data. The measure of noise included is constrained by the boundary epsilon (ε). To render meaningful privacy, differential privacy requires epsilon in the range $0 < \varepsilon < 1$. This is a key challenge, since greater values of epsilon improve data analysis quality but with reduced privacy. In this work, two different approaches are proposed for optimized private learning. The first approach is the implementation of differentially private classification using shallow learners [10]. The benchmark Gaussian Naïve Bayes (GNB) [11] and Logistic Regression (LR) [11] are the chosen shallow learners. A balance between classification accuracy and epsilon is obtained by optimizing epsilon with a modified Non-dominated Sorting Genetic Algorithm NSGA –II [12] algorithm. The second approach proposes a deep learning model using Bayesian optimization that gives a quantifiable trade-off between privacy, utility and performance. The proposed workflow is shown in Fig 1.

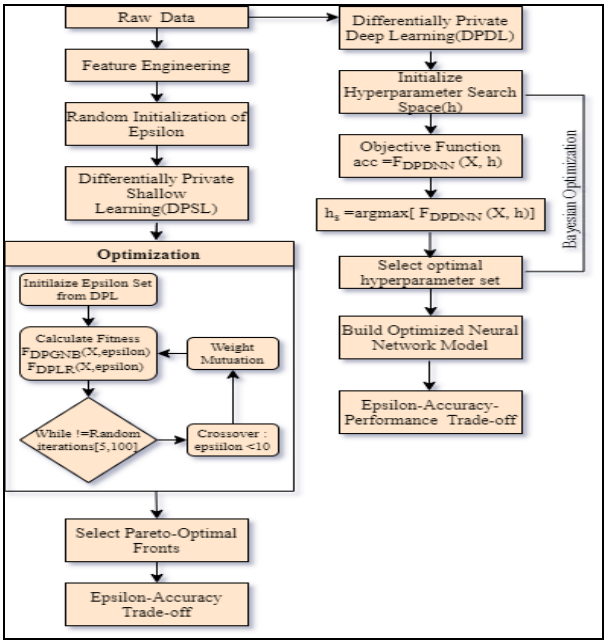

Fig 1 Proposed Workflow

B. Differentially Private Shallow Learning

In the first approach, private learning is split into two phases. The initial phase is differentially private learning and next phase is optimization for enhancing performance. During the first phase, as shown in Fig 1, two differentially private models are trained using the learners, Gaussian Naïve Bayes and Logistic Regression. The datasets used for the learning are adapted from the UCI machine learning repository [7]. Adult [7] and Diabetes [7] are the datasets used for implementation. The raw data is first pre-processed and trained using the shallow learners.

IBM's library DiffPrivLib[5] is used for implementation. GaussianNB requires specification of epsilon and bounds on attribute values during learning. The training is done on data using different random ranges of epsilon. Similarly, logistic regression is trained with l2 norm for regularizing the fit and pre-defined epsilon ranges. The test accuracies generated by the private learners are different on every random run of the model and all such {epsilon, accuracy} pairs are recorded. These are used to initialize the random population of the modified NSGA-II algorithm for optimization in the subsequent phase of the approach.(Section C)

C. Optimizing Epsilon using Evolutionary Approach

In the second phase of shallow learning, evolutionary approach based on modified NSGA –II[18] algorithm is proposed to optimize epsilon during learning. The modified NSGA-II algorithm generates a set of pareto-optimal fronts [13] from which an optimal value is chosen. The pareto-optimal set consists of {epsilon, accuracy} pairs for differentially private classification analysis. The steps of the algorithm are as described in Section D.

D. Modified NSGA II Algorithm

1. Assume random population size(s)
2. Initial epsilon set populated from random runs of differentially private learning (Peps)
3. Calculate Fitness value (F):
   a. acc = F_{DPGNB}(X, epsilon)(Differentially Private Gaussian Naïve Bayes)
   b. acc= F_{DPLR}(X, epsilon) (Differentially Private Logistic Regression)
4. While (! =randomly generated no of iterations in range [5,100])
   a. Implement crossover rule: select individuals with epsilon< 10
   b. Perturb crossed over individuals by weight mutating each individual with random weight β using Gaussian distribution
   c. Calculate fitness

E. Differentially Private Deep Learning

The second proposed approach uses a deep learning neural network for implementing learning. Deep learning frameworks for differential privacy have been gaining lot of importance due to its practical feasibility. The proposed private deep learner for classification analysis is implemented using the open source library tensorflow privacy. A deep neural network is modeled on data with accounting of epsilon that determines the privacy guarantee offered by the model. Adult and Diabetes datasets are used for analysis. When compared to shallow learning, deep neural network gives better epsilon-accuracy trade-off. Accuracy is found to improve with increasing epochs, but at the cost of running time. This basic model is referred to as the standard deep learner. The learner's performance can be improved by fine-tuning the hyperparameters to facilitate faster training. Precisely, optimizing the parameters such as learning rate and number of epochs enhances learning and computational efficiency. Hence this standard deep learner is optimized using Bayesian method to reduce the training time while balancing both accuracy and epsilon.

F. Bayesian Optimized Differentially Private Deep Learning

The performance of the deep learner is improved by tuning the hyperparameters using the Bayesian method [14][15]. Bayesian method efficiently tunes hyperparameters by using the probabilistic Bayesian theory [15][17] assuming a Gaussian distribution of the variables. In comparison to the grid search [16] and random search[16] methods, which searches the entire parameter space, Bayesian method reduces the time for search, since each subsequent iteration of optimization selects parameters from the past results[18]. The search space size is reduced and hence converges faster. The optimal selection of hyperparameters for the learning model, improves training time, reducing the number of iterations required for training.

The steps for optimization using Bayesian Method are as follows:

Let h be initial hyperparameter search space
acc be accuracy of classification
F_{DPDNN} (X, h) be objective function of differentially private deep neural network(DPDNN) with X as the input space

1. Initialize the search space with random default hyperparameters to be tuned. The search space consists of the following hyperparameters
   h= {learning rate, batch size, number of dense layers, number of dense nodes, Adam decay}
   where h is the set of hyperparameters
2. Define the objective function
   acc =F_{DPDNN} (X, h)
   h_{opt} =argmax[F_{DPDNN} (X, h)]
   The objective function values are assumed to follow a Gaussian Distribution.
3. Select the most effective sample set of hyperparameters with the highest posterior probability P(acc| hopt)

IV. EXPERIMENTAL EVALUATION

A. Evaluation of optimized shallow learners

This section describes the evaluations of private learners for both the proposed approaches. Both shallow and deep private models are evaluated on the datasets Adult and Diabetes adapted from the UCI machine learning repository. Shallow learning is implemented using the IBM DiffPrivLib library. The graphs in Fig 2 and Fig 3 show the epsilon-accuracy trade-off using Naïve Bayes and Logistic Regression for the two different datasets. The plot is generated by fitting the objective function after crossover and mutation as described in the algorithm. The crossover rule selects epsilon individuals in range [0,1] which is then weight mutated to generate diverse individuals. The algorithm is repeated for pre-specified random runs in the range [5,100]. The GaussianNB model (Fig 2) shows an accuracy range of [0.4, 0.7] for epsilon values less than 1. Further, accuracy increases and stabilizes beyond epsilon value of 1.

The logistic regression model's (Fig 3) accuracy is low for smaller ranges of epsilon. The linear increase in accuracy is seen for values of epsilon beyond 1. NB performs better for both datasets in comparison to logistic regression. Logistic regression performs poorly for smaller values of epsilon, and shows a linear rise in accuracy only for values greater than 10. Shallow learning causes greater privacy leaks to enhance learning performance.

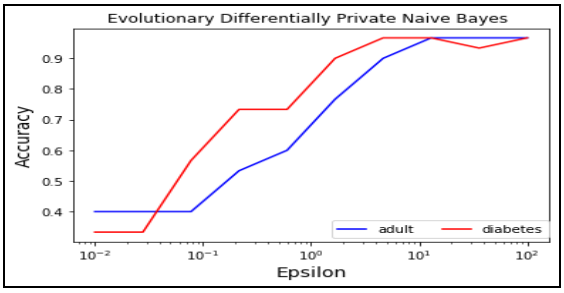

Fig 2: Epsilon –Accuracy Trade-off - NB

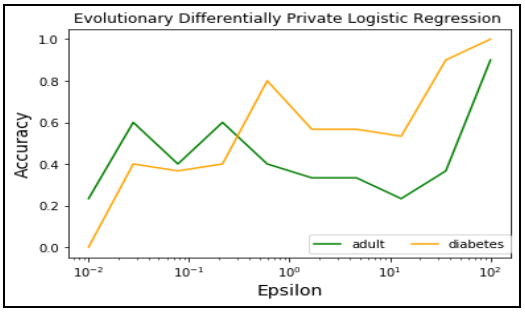

Fig 3: Epsilon –Accuracy Trade-off - LR

B. Evaluation of optimized deep learner

This section compares the efficiency of the standard deep learner with the Bayesian optimized(BO)learner. Architecturally, the standard model has 1000 dense nodes, normalized and activated using the sigmoid function. The same model is used for the optimized variant. The optimization procedure is initiated with a random sample of default hyperparameters. The prior probability of the objective function is observed with the initial set and posterior is updated. The function which returns the largest accuracy is chosen for the selection of the optimal hyperparameter set. The tabulation (Table 1 and Table 2) shows values of epsilon, accuracy and execution time for adult and diabetes datasets respectively. The values are recorded for different iterations. The graphs in Fig 4,5,6,7 show comparison of results between the standard and optimized learners for the adult and diabetes datasets.

Table 1 Epsilon-Accuracy-Efficiency-Adult Data

Standard Model - Adult Dataset				Optimized Model –Adult Dataset			
Epochs	Accuracy	Epsilon	ExecTime(s)	Epochs	Accuracy	Epsilon	ExecTime(s)
10	0.81	5.68	61.2	1	0.83	0.782	3.2
30	0.82	6.23	102.3	2	0.84	0.891	4.89
50	0.83	6.59	150.3	3	0.842	0.923	6.312
70	0.87	7.21	253.25	4	0.853	1.02	7.89
90	0.87	7.86	310.32	5	0.86	1.15	8.21
110	0.8723	8.3	340.89	6	0.87	2.25	8.29
130	0.8756	10.21	412.12	7	0.892	3.523	9.35
150	0.879	11.19	490.11	8	0.89	4.563	9.783
170	0.8792	11.56	509.12	9	0.89	5.34	10.56

Table 2 Epsilon-Accuracy-Efficiency-Clinical Data

Standard Model - Clinical Dataset				Optimized Model –Clinical Dataset			
Epochs	Accuracy	Epsilon	ExecTime(s)	Epochs	Accuracy	Epsilon	ExecTime(s)
10	0.76	4.36	65.2	1	0.82	0.87	3.2
30	0.772	5.11	112.3	2	0.84	0.899	4.89
50	0.778	6.78	167.3	3	0.83	0.91	6.312
70	0.79	6.923	292.33	4	0.843	1.56	7.89
90	0.81	7.23	320.32	5	0.87	3.78	8.21
110	0.822	7.89	348.77	6	0.88	5.25	8.29
130	0.83	8.23	467.82	7	0.881	6.23	9.35
150	0.842	8.45	590.71	8	0.892	7.563	9.783
170	0.84	9.23	599.42	9	0.897	8.34	10.56

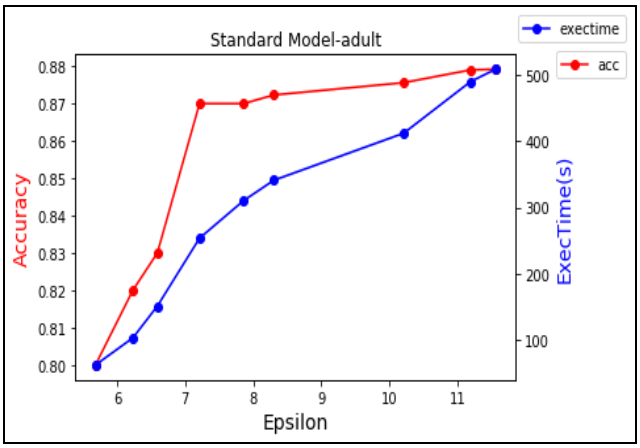

Fig 4: PUP trade-off in Standard Learner-adult

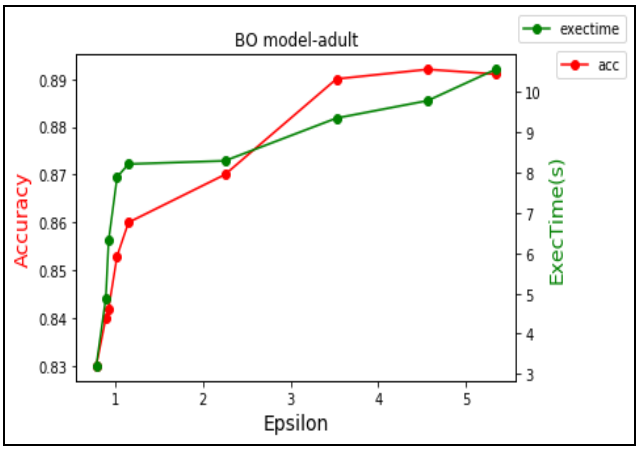

Fig 5: PUP trade-off in BO Learner-adult

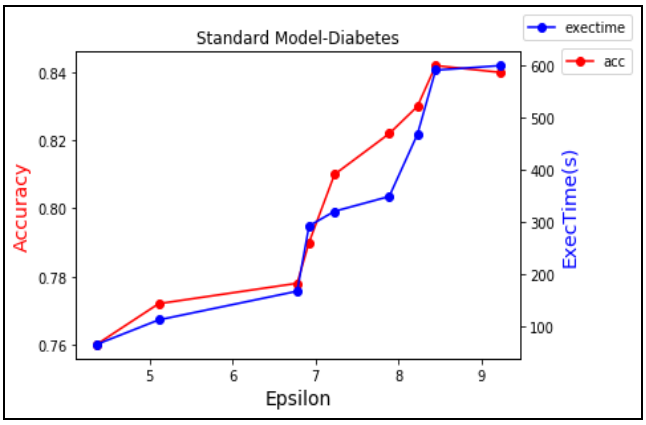

Fig 6: PUP trade-off in Standard Learner-diabetes

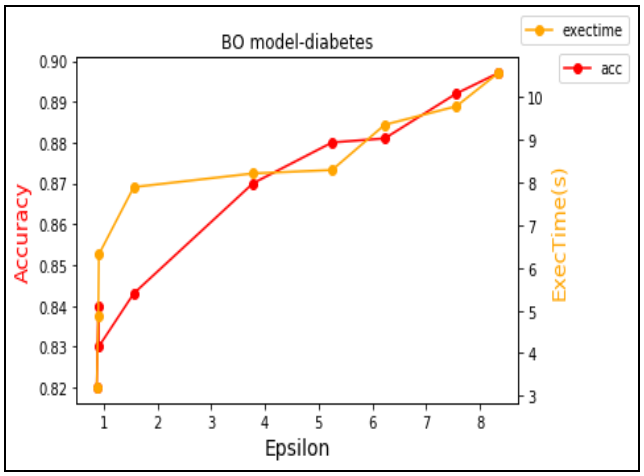

Fig 7: PUP trade-off in BO Learner-diabetes

The graphs in Fig 4 and Fig 6 show the standard learner's performance for the datasets adult and diabetes respectively. Epsilon is found to increase with increasing accuracy. So is the case with learning speed. The BO learner's performance is shown in Fig 5 and Fig 7. The proposed BO learner performs better than the standard learner with respect to the epsilon range and accuracy. Epsilon values for optimized learner ranges between <0.78, 5.34> for adult and <0.87, 8.34> for diabetes. It can be perceived that the standard model's learning initiates with larger values for epsilon in the range <5, 11>. This range is greater than the corresponding optimized learner. The enhanced learning speed of the optimized learner is attributed to the optimization of the number of epochs for training the model. The standard learner requires the model to be trained for atleast 10 epochs to give a starting accuracy of 80%. In contrast, by fine-tuning the hyperparameters, greater accuracy is achieved due to the probabilistic learning of Bayesian method. The difference in the number of epochs required for the training both models is shown in the graphs (Fig 8, Fig 9, Fig 10, Fig 11).As can be seen from the tabulation, number of epochs has definitely influenced the training time. While the standard model requires 10 epochs to give an initial accuracy of approximately 80 percent, the same is achievable in lesser number of epochs. Similarly, the choice of an optimal learning rate and activation function facilitates faster and improved learning.

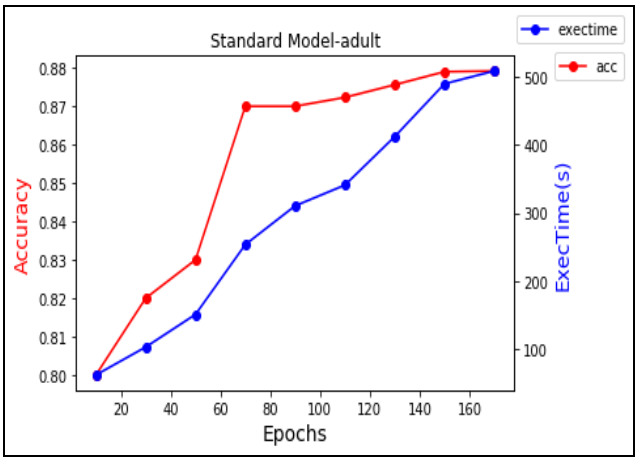

Fig 8 Efficiency Vs Epochs –Standard Model-adult

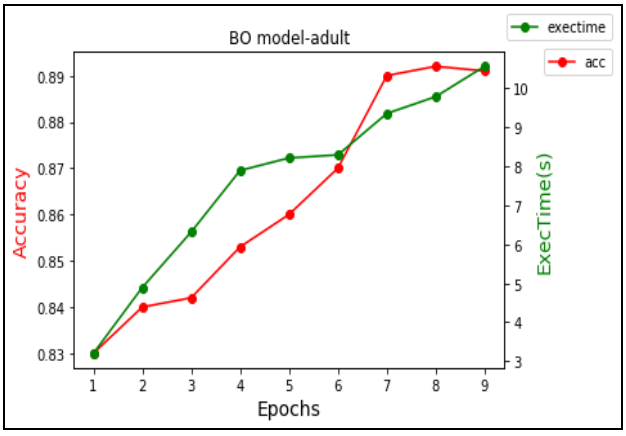

Fig 9 Efficiency Vs Epochs –BO Model-adult

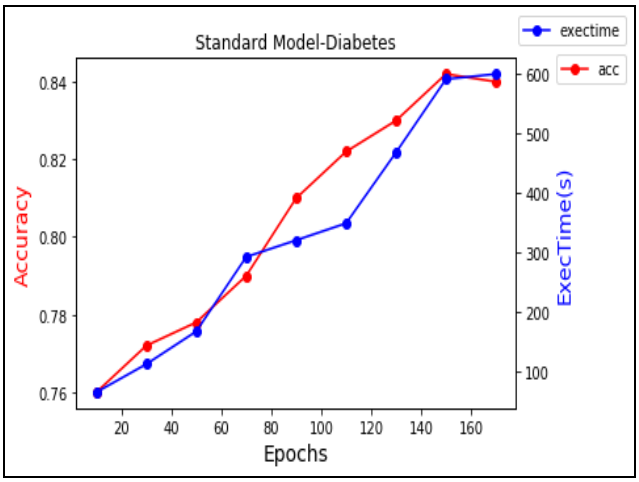

Fig 10 Efficiency Vs Epochs –Standard Model-diabetes

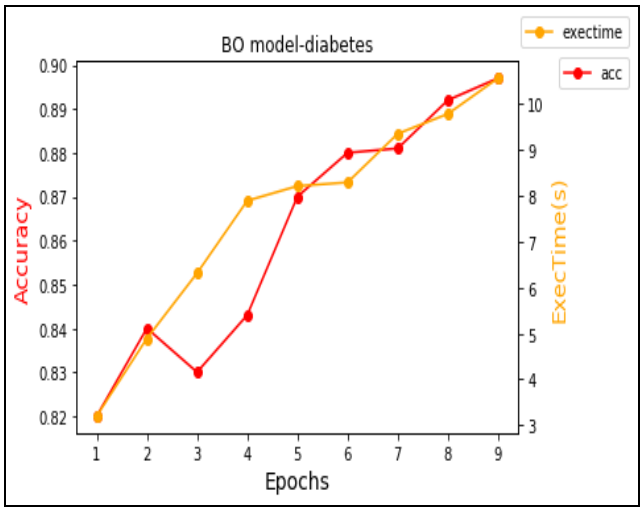

Fig 11 Efficiency Vs Epochs –BO Model-Diabetes

V. CONCLUSION

Privacy preserving data analysis focuses on protecting sensitive information of individuals involved in data analysis. While data analysis is beneficial in a number of applications for gaining useful insights, it is vital to respect the privacy of individuals in the data. Differential privacy is a robust technique with strong scientific guarantees that assures privacy of data. Since analysis on privacy protected data diminishes its usefulness, it is required to adopt machine learning techniques that can enable private learning efficiently. Specifically, since the algorithm works by adding noise, lesser the noise, better the utility and vice-versa. Considering that differential privacy uses epsilon to quantify the extent of privacy, the primary objective is to minimize epsilon, while maximizing utility and performance of the analytic technique. The work in this paper proposes two approaches for privacy preserving learning. Firstly, shallow learning for classification is implemented using evolutionary approach to minimize epsilon. Classification accuracy using the optimized evolutionary method improves at the expense of greater epsilon values and increased learning speed. Secondly, a Bayesian optimized deep learning approach is proposed for private learning, using which an optimal set of hyperparameters is generated for the learning model. The tuned model learns faster and hence causes lesser privacy leaks. The proposed deep learning neural network trains faster with lesser privacy leaks and increased accuracy. Coupling the learning ability of deep learning with a robust privacy technique such as differential privacy provides a wide scope for efficient privacy preserving analytics in big data.

REFERENCES

1. Latanya Sweeney "Achieving k-anonymity Privacy Protection Using Generalization and Suppression",May 2002, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588.
2. Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3 4):211–407, 2014.
3. https://www.macobserver.com/analysis/google-apple-differential-privacy/
4. Kobbi Nissim, et al. Differential Privacy: A Primer for a Non-technical Audience. February 14, 2018.https://diffprivlib.readthedocs.io
5. www.tensorflow.org
6. https://archive.ics.uci.edu/ml
7. Marler, R. T. and Arora, J. S. (2004). Survey of multi-objective optimization methods for engineering. Structural and multidisciplinary optimization, 26(6):369–395
8. parvizi, Mahdi & Shadkam, Elham & jahani, Niloofar. (2015). A Hybrid COA/ε-Constraint Method for Solving Multi-Objective Problems. International Journal in Foundations of Computer Science & Technology. 5. 27-40. 10.5121/ijfcst.2015.5503. www.medium.com
9. www.towardsdatascience.com
10. Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA–II. IEEE Transactions on Evolutionary Computation 6, 2 (2002), 182–197.
11. Rinku Dewri, Darrell Whitley, Indrajit Ray and Indrakshi Ray," A Multi-Objective Approach to Data Sharing with Privacy Constraints and Preference Based Objectives", Proceedings of GECCO 2009
12. www.cs.torontu.edu
13. Bargav Jayaraman and David Evans," Evaluating Differentially Private Machine Learning in Practice", Proceedings of the 28th USENIX Security Symposium, 978-1-939133-06-9, August 14–16, 2019
14. Thanh Dai Nguyen(B), Sunil Gupta, Santu Rana, and Svetha Venkatesh," A Privacy Preserving Bayesian Optimization with High Efficiency", LNAI 10939, pp. 543–555, 2018.
15. Borja Balle, B. Avent, J. Gonzalez, T. Diethe and A. Paleyes," Learning the Privacy-Utility Trade-off with Bayesian Optimization"
16. Guillermo Campos Ciro, Frederic Dugardin, Farak Yalaoui, Russell Key, "A NSGA-II and NSGA-III comparison for solving an open shop scheduling problemwith resource constarints,2405,8963,2016

AUTHOR PROFILE

Geetha P. is currently pursuing her Ph.D from Visveswaraya Technological University, Karnataka. She is a postgraduate in Computer Science from Visvesvaraya technological University. Machine Learning, Data Analytics and Cyber Security are her areas of interest. She has publications in various National/International conferences and journals. Her teaching experience of 12 years has gained her expertise in courses like data structures, algorithm design and operating systems. She is currently working as Assistant Professor in Cambridge Institute of Technology, Bangalore, Karnataka.

Chandra Kant Naikodi has over 14 years of involvement in Software Development industry. He got his Ph.D. from UVCE, Bangalore. He has published papers in many popular national and international journals. He has created 16 specialized reading material distributed by Tata McGraw Hill, SCHAND. His areas of interest include Computer Networks, MANETS, WSN, and Big Data. He is associate professor in the branch of Studies and Research in Computer Science (PG), Davangere University. Chandrakant Naikodi has worked as a Principal Applications Engineer in a MNC, Bangalore, India.

Suresh L has received his Ph.D in the year 2010. His academic experience is around 29 years. His services in teaching field are highly appreciated by the student community. Suresh has authored books on Data Structures, Algorithms, Java, Python and other programming languages. He has published more than 70 papers in many reputed national and international journals. His expertise is in the core areas of Data Mining, Data Structures and Algorithms and Database Management Systems. He is currently the Principal in Cambridge Institute of Technology, Bangalore, Karnataka.