Corrosion Management **ISSN: 1355-5243**

(https://corrosion-management.com/)

Volume30, Issue 01–June2020

SCOPUS

Privacy-Preserving Machine Learning Techniques for Data in Multi Cloud Environments

Rahul Vadisetty, ***Electrical Engineering,***

Wayne state university, **Detroit, MI, USA rahulvy91@gmail.com**

***Abstract*—Cloud computing has significantly transformed how organizations handle their data. By utilizing multiple cloud services, businesses can achieve greater operational flexibility and cost savings. However, this approach also introduces new challenges, particularly concerning the privacy and security of information. This research explores the application of differential privacy in machine learning, specifically focusing on logistic regression models trained on the MNIST dataset. Key findings include the model's performance, the privacy-accuracy trade- off, and generalization capabilities. This study demonstrates the feasibility and effectiveness of differential privacy in creating machine learning models that respect data privacy without significantly compromising accuracy.**

***Index Terms*—Privacy-preserving machine learning, differential privacy, multi-cloud environments, logistic regression, MNIST dataset.**

I. Introduction

A. Background

Cloud computing has played a major role in the way data is managed in organizations, thus giving businesses operational flexibility at a cheaper cost. This results in the fact that by using several cloud services organizations can divide their applications across various clouds, which in its turn, increases the efficiency and reduces the expenses. Nevertheless, the application of this approach also creates new problems, especially, related to confidentiality and protection of information. Another emerging technology that has been branching out fast is the machine learning or commonly referred as ML which has significant roles to play in the prediction, automation, and decision-making analytics. Additional advantages of using ML with a cloud computer include scalability, availability, and processing power. But at the same time, it introduces new privacy challenges because the personal data must be processed and stored in the various cloud systems. Privacy preserving machine learning (PPML) aims to create techniques that would allow the ML models to be used whilepreserving privacy of data.

Developing with the cloud is widespread today since it provides effective and adaptable ways to store and retrieve data. These services can be used by companies to improve their business functioning; however, owing to the decentralized infrastructure of clouds, data becomes more vulnerable to privacy violations. It is here that PPML techniques come intoplay to ensure that the ML models continue to be useful while meeting extreme levels of privacy preservation.

B. Problem Statement

As it is beneficial to use several cloud platforms in some cases, it complicates the issue of privacy. Data that are stored in different cloud environment are relatively at more risks of being hacked or breached. Such problems cannot always be solved with traditional security approaches, as the latter are usually developed for a singular localized scenario. It becomes crucial to explore new methods of privacy preservation such that data could be secured, and Machine Learning should be adept at using them also.
The presence of having multiple clouds makes it even harder for organizations to secure their data and this calls for the need to change the approach of protecting data. Traditional security measures like encryption and access controls cannot prevent the associated risks with data scatter in the various cloud services. The PPML research in this paper is centered on assessing the available methods to improve data security in multi-cloud environments without compromising on the leveraged strengths of machine learning.

C. Objectives

This research aims to:

- Discuss the existing approaches to developing private ML models.
- Discuss the usefulness of these techniques in the context of a multi-cloud setup.
- Suggest considerations for the using of PPML in the multi-cloud environment.

The objectives are primarily oriented towards comprehending the general context of research into PPML and its relevance in intricate scenarios of cloud computing. In that regard, the objective of the research is to present the existing techniques for data privacy and evaluate the applicability of these. approaches in multi-cloud environments and, thus, outline the framework for the improvement of data privacy based on the analysis of the existing practices.

D. Research Questions

This study aims to answer the following questions:

- Currently, which kinds of machine learning methods can maintain the privacy of data?

- In what way do the mentioned techniques contribute to the protection of data privacy in multi-cloud settings?

- Sometimes, it is essential to identify the strengths and weaknesses of these techniques in relation to the goals being achieved.

- Techniques that can be used to manage the various aspects of multi-cloud include But how can these techniques be integrated into a framework for multi-cloud environments?

Answering these questions will help comprehend the current trends in PPML's development and its real-world applications. This research will identify those shortcomings and suggestions of how multi-cloud data can be properly secured to mitigate the current flaws in the currently available methods.

E. Contribution of the Study

This research is crucial for several reasons:

- Privacy Issues: Concerning the primary subject of this book, it covers fundamental privacy considerations appreciated in the context of cloud computing and machine learning.

- Knowledge Contribution: That way, it adds to the existing literature on privacy preservation, presenting findings on how techniques can be applied in multi-cloud context.

- Security Improvement: To that end, the findings may be used to create safer and privacy-preserving machine learning technologies, which can be used in such industries as healthcare, finance, and government.

The contribution of this work is based on improving the existing knowledge in order to make cloud-based ML ap- plications more secure. The study direction, thus, involves privacy preserving to ensure that the possible dangers of data dispersion in multi-cloud are averted. This can help in the development of better and more reliable ML applications within numerous fields.

F. Structure of the Paper

The paper is organized as follows:

- Chapter 2: Literature Review – A detailed examination of existing privacy-preserving machine learning techniques.

- Chapter 3: Methodology – The methods and approaches used to evaluate these techniques in a multi-cloud environment.

- Chapter 4: Results and Discussion – An analysis of the findings and their implications.

- Chapter 5: Conclusion and Future Work – A summary of the research, its contributions, and potential future directions.

II. Literature Review

A. Introduction

This chapter reviews the current literature on the workflow of PPML in the era of multi-cloud orchestration, and related.

solutions to the problem. The goal is to create a theoretical background for the given analysis by discussing modern approaches, their usage, and drawbacks. This chapter also points at the areas of and theoretical gaps that this study seeks to fill and introduces the subsequent chapters of this work.

B. What is Cloud Computing and Multi-Clouds?

Today, the advancement in technology particularly the cloud computing has made a drastic change in the entire way data is stored, managed, and processed. For the businesspeople, it provides them with opportunities of flexible, efficient, and cheaper ways to manage their operations. Environment that emerged in which data and applications are distributed between providers of CSPs deepens these advantages by preventing monopolization and increasing backup.

1) Benefits of Cloud Computing:

- Scalability: Cloud services can add or remove resources depending on the requirement of users and organization hence becoming economical.

- Cost-Effectiveness: This way, through cloud services, one can dramatically minimize capital costs associated with the creation and maintenance of infrastructure.

- Accessibility: It allows the subscription to data and ap- plications whenever needed, it supports the working from home model, and collaboration among the teams.

2) Challenges in Multi-Cloud Environments:

- Data Fragmentation: However, the insertion of data in several clouds may cause the spread of these data and therefore making it challenging to organize and protect them.

- Interoperability Issues: Co-ordination across multiple cloud services management from different cloud service providers, can be challenging since they have different protocols and standards.

- Increased Attack Surface: Complex configurations such as multi-cloud architecture; can mean more possible entry points and thus, a higher level of risk.

C. Machine Learning and Its Integration with Cloud Computing

Taking into consideration specifics of prior calculations, machine learning (ML) is statistical computer programs that allow the computer to answer such problems without programming. It has also impacted the analysis of data, existence of automation and also the enhancement of predictive modelling through the synergy of ML with cloud computations.

SubsubsectionApplications of Machine Learning in Cloud Environments

- Predictive Analytics: Some of the applications of the developed ML models include Probability of forecasts concerning trends and behaviors in the future based on patterns that have been identified in the past.

- Automation: By so doing, any ordinary work that consumes much of people ’s time can hence be done through the application of ML algorithms hence improving on the tasks results.

- Personalization: ML could also be helpful in identifying the behavioral pattern of the user along with his preferences and thus a better user experience can be offered to a user.

1) Privacy Issues on Cloud Based Machine Learning:

- Data Sensitivity: MFor the functioning, some ML models might require large datasets which may contain such information as personal records – the information that has to be protected.

- Data Ownership: Hence, when going for many cloud service providers there is a big question of who owns the data and how the data is going to be protected.

- Regulatory Compliance: Moreover, It is very crucial to abide by the principles of data protection regulations in particularly when dealing with very sensitive data (GDPR, HIPAA, etc.).

D. End-to-End PPML Learning Methodologies

Privacy preserving Machine Learning is therefore a branch of study that focuses on techniques that don't enable the construction of ML models on data that can be retrieved by other individuals. Also in this section, an outlook of the major techniques in PPML together with the usage will be presented.

1) Differential Privacy: Thus, the primary goal of differential privacy is to prevent the analysis received from indicating whether any individual's data was included as part of the input set. This alters some degree of noise to the data or computations as a means of concealing identity of people.

- Advantages: Provides high privacy guarantee and also flexible regarding the choice of ML algorithm to be used.

- Limitations: To summarize, noise hinders the training the produced ML models by degrading the quality of the training set.

2) Federated Learning: FL is aimed to facilitate the building of the ML model with the assistance of numerous devices or servers containing decentralized sample information without sharing the info. This model update, which does not convey any raw data of the patients, is transferred instead.

- Advantages: The information is retained locally, and this means that it is very hard to tamper with the secrete of the persons.

- Limitations: It is obligatory for providing intense communication support but can raise the issue of unequal distribution of data.

3) Homomorphic Encryption: Basically, homomorphic encryption refers to a state where computations can actually be done directly on cipher-texts of the original message without necessarily decrypting the message. The outcome of the operations stays latent, or 'frozen,' and can become active or 'active' only by the permission of the owner of the information.

- Advantages: Remembers the data during the time of the computations is being made.

- Limitations: This is however accompanied by a lot of computation overhead compared to the result for corresponding computation on plain text.

4) Secure Multi Party Computation or SMPC: Inexplicably, SMPC surrounds a scenario in which many people evaluate a function over their input but preserve the input's security. Big boys and girls do it for output development directly with the input-output agreement that they will not transmit information owned by the counterparty.

- Advantages: Two or more occupants can solve on shared variables' fundamentals without other persons in the room seeing the content of these solutions.

- Limitations: It will also likely include a lot of overhead in terms of load on a computing system, and possibly the communicational load is also high.

E. Roles and Cases of Use of PPML in Multi-Level Computing

Within this section, several usage scenarios are described, and in each of them, one or two examples of how PPML techniques can be applied and get the biggest result in the multi-cloud environment.

1) *Healthcare:* In practical implementation, it means that

patient's data may be transmitted from one hospital to another, yet patient's individuality shall remain protected. For example, federated learning may apply in the architecture of prognosis models of diseases occurrences available information from various institution.

2) *Finance:* If the self is entitled to the finance sector,

then the PPML techniques can be helpful in all the study where transactional information from different banks needs to be analyzed with an aim of identifying fraudsters for their clients without disclosing their clients' details. Thus, it can be ascertained that risk analysis can be made on the compliance data using the efficiency and effectiveness of the homomorphic encryptions in preventing and addressing such events.

3) *Government:* Thus, it is proven through this paper that

PPML is useful to the extent that it can be used by government agencies to perform analysis on data coming from as many departments as desired and simultaneously comply with the privacy regulation. This method can aid in the comparison of crime statistics across Jurisdictions because the method used in the process is the secure multi- party computation.

F. Lessons Not Identified in the Literature

Despite the advancements in PPML, several gaps remain in the existing research. However, the following research issues are still unaddressed in the PPML literature:

- Scalability: The problem that can be noted with many of these PPML techniques is that they are still quite limited with scalability as they are used with the larger datasets obtained with operations in multiple clouds.

- Efficiency: Privacy and efficiency – these are two more challenges for which, although the solutions have not yet been found, one realizes that when striving to implement the privacy-preserving approach, the efficiency of the process drops.

- Interoperability: There are still many concerns opened in the area and one of the most common is related to how the migration of the different PPML techniques from one Cloud provider to another is possible.

G. Summary

This chapter offered insights into the details of the selected literature on privacy-preserving machine learning methods under the multi-cloud scenario. Other areas of coverage are; benefits of cloud computing, disadvantages of cloud computing, relationship between ML and cloud solutions and the PPML'S. Then, having addressed the specifics, it disclosed the examples of employing these techniques and drew attention to the scarcity of discursive material in the literature.

The process of how the effectiveness of PPML techniques in multi-cloud contexts will be assessed is described in the following chapter.

III. Methodology

A. Introduction

This chapter provides the research method that has been employed on the study that measures the efficiency of PPML within the multi-cloud context. The justification for choice, reasoning, and approach to gathering data and the procedures used in analyzing the collected data; how reliability and credibility were attained in the study is highlighted in the following sub-section. The chapter also considers various items that could be linked to the ethical consideration of the study. Here much emphasis has been laid on formal aspects that all the stages of the study are free from methodological bias and do not violate ethical norms; Also, if all the stages of the study fully complied with scientific methods the conclusions that could be derived while evaluating the PPML techniques would be much more accurate and objective.

B. Research Design

The study design is a strategy of ruling out the research questions and objectives highlighted in Chapter 1 of the study. Thus, this paper makes both qualitative and quantitative assessments in efforts to give a comprehensive outlook on the assessment on applicability of PPML methods.

1) Qualitative Approach: The said approach still entails collection of data from various journals, administering questionnaires to other professionals to establish the existing state regarding the use of PPML and practices concerning multi- cloud environments whenever it is the case. It also makes one conscious of the probabilities that are inherent in the use of such techniques and the demerits, or the actual life repercussions of using these solutions particularly if it is conducted incorrectly. Hence, the study intends to interact with other scholars in the field of study and via the synthesis of literature at large, to ensure that the study is fashionable with a sound and all-embracing purview of the field.

2) Quantitative Approach: Another quantitative method that implies conducting experiments and simulations is called experimentalism and it focuses in comparing various PPML methods. Such assessments may include the precision and speed of the mentioned techniques and, conceivably, the versatility and confidentiality of the multiple cloud solution. Thus, the quantitative analysis is aimed at providing rationale.

for the conclusions and proving the effectiveness of the used methods in the field of PPML.

C. Data Collection Methods

The gathering of data is a crucial process for any piece of research as it forms the basis of the study in question. The next part presupposes to delineate the process of Primary and Secondary data collection used in the frames of the research.

1) Literature Review: The data acquired from the studies and papers constitute a literature review that enables one to gather existing information on PPML techniques. The sources encompass the professional or academic journals, research pa- pers presented in conferences, White papers, industry reports and authorities are used to evaluate and determine trends and developments and peculiarities in the section. Considering the highly high number of papers related to the subject, it becomes possible to identify the current trends of PPML methods and how they may be applied to multi-Cloud environments.

2) *Expert Interviews:* Therefore, interviews were conducted with academics from the field of cloud service, machine learning, and data protection to get a list of practical applications of using PPML techniques. These interviews illuminate quite a lot on the use of these methods and the realistic weaknesses that can be expected from them. This is done with the help of acquisitions and infusions of subject matter professionals from the academia, industries and other relevant formal institutions.

3) *Experimental Data:* This way, survey and case studies are applied to have data for the research as the organizations with the experiences on the multi-cloud roles based on PPML methods are considered. This is applied to assess the effective- ness of special methodologies under different circumstances. About the control used, it is also very applied or realistic; or better put, the experiment is made as realistic as possible.

D. Data Analysis Techniques

It incorporates the assessment of the collected data and the capacity to draw some conclusions for it without totally evaluating all the information accumulated. This subtopic explains how to conduct the analysis of both, qualitative and quantitative data.

1) *Qualitative Data Analysis:* Categorization of the collected data is as follows: The information obtained during the review of the literature and the opinions of the specialists undergo a process known as thematic analysis. This also means analyzing the patterns and repeatedly in the data to give the possibility to make recommendations regarding the further usage of PPML methods, as well as the advantages and drawbacks of the framework, and the steps to perform the PPML framework applying. Thematic analysis is very important in areas of data analysis since it helps out in the categorization of big qualitative data for hopefully improved appreciation.

2) *Quantitative Data Analysis:* Quantities that are obtained from experiments or simulations and are then statistically processed are data of experience. As was mentioned earlier, PPML techniques are bifurcated based on the evaluation.

criterion that entails the considerations of accuracy, efficiency, scalability, and privacy optimality. These techniques are explained while at the same time the performance of some of these techniques can be demonstrated by the help of a graph and chart. Finally, the statistical analysis aims at making a wise assessment as well as analysis of findings in profiling practices.

E. Experimental Setup

Therefore, the specificity of this configuration is decisive in relation to its assessing in the frames of PPML methods concerning multi-clouds. The present sub-section is devoted to the descriptions of the stages of preparation and arrangement of experiments in the given work.

1) *Multi-Cloud Environment Configuration:* These are mul- tiple cloud settings which include the services which are provided namely Amazon Web Service, Google Cloud, and Microsoft Azure. Realistic Multiple Cloud Scenario is created within these cloud platforms with the help of various datasets used within them. It is supposed to provide an imitation of the basic settings of the multi-cloud model and specifics of the corresponding practice.

2) *Identification of techniques in PPML:* Out of the flood of methods available for the analysis, certain techniques of PPML are selected for the evaluation because the mentioned methods are frequently applied and provide a broad analysis of the given data set in the literature. Some of these techniques are differential privacy, federated learning, homomorphic encryption, and secure multi-party

computation among others. The criteria are made regarding the methods in question as to the relevance, efficiency, and practice in the contemporary research.

3) *Applying of the Techniques in PPML:* It is now possible to apply the selected PPML techniques in the mentioned multi- cloud environment that has been set. Popular datasets such as the MNIST and Universal Image Datasets used in image recognition, and data downloaded from the UCI Machine Learning Repository in classify such datasets as Breast Cancer, Glass Identification, Iris dataset and so on are used to train as well as to test the above machine learning algorithms. The example of implementation process is the setting of the necessary software and hardware environments.

F. Evaluation Metrics

The efficiency of the implemented PPML techniques is assessed using some indicators. The current section shows the steps taken in effect in the analysis of the research.

1) *Accuracy:* Accuracy is a technique of identifying how effectively the patterns are forecasted by the machine learning models. It is obtained by dividing the total accuracy by the total in the material base number of cases. High accuracy proves that the model makes few mistakes of the generalization and is accurate in its predictions.

2) *Efficiency:* Cost of computation examines the amount of PPML techniques that incorporates more resources in quantity. It means efficiency is defined in terms of time, memory and the

bandwidth of the network. Optimization methods imply that any resource applied will be properly used without a sacrifice of the outcomes to be achieved.

3) *Scalability:* Scalability tries to assess the scalability of the PPML techniques for the purpose of analyzing the relation between the amount of data with the models and the required storage and computation. To carry out the scalability assessment, experiments are conducted with a bucket of a different size and different levels of clouds. It means that scalability can contribute to the resolution of the situation with growing quantities of data and users' requests.

4) *Privacy Preservation:* It lies in privacy preservation measures which attributes the techniques of the PPML to the privacy feature of the data. The extent of privacy preservation is represented by what is referred to as privacy loss, information leakage and ability to counter some types of attack. This is because there are proper ways in which data can be managed to expose them to fewer privacy risks as well as increase the value of data.

G. Reliability and Validity

To solve such an issue, it is essential to also develop the reliability and validity of the researched results. In this section, the approaches that were utilized to increase the credibility of the study can be described.

1) *Reliability:* All in all, reliability carries a lot of information relating to the capacity to perform research consistently and receive the same outcome. To ensure reliability the tests are conducted several times and the results formulated similarly and compared if they would produce the same value. This is the case if the results found are the same hence making the reliability to be high.

H. Validity

On the other hand, reliability focuses on the extent that can be placed on the research study concerning the soundness of the results and their validity. Therefore, it is necessary to provide proofs of validity and based on the prevalent approaches in the field of validity assessment, the study uses valid criteria and standards to evaluate. It is convenient to replicate real multi-cloud environments with the help of an experimental setup's location, and the given outcomes are clear to specialists. High validity implies that the results which are being achieved are quite the reflection of the impact of the PPML techniques.

I. Ethical Considerations

Privacy as an ethical consideration is always a cardinal virtue when it comes to any data research. This section articulates the ethical consideration of the study.

1) Data Privacy: Data regarding the subject 602 is one of the aspects of the research proposal where 255 personal information, and every effort is made to preserve and safe- guard the subject's information and rights. Regarding data of experiments, all gathered information is anonymizedThe process of collecting and storing all the data corresponds to the

requirements of personal data protection and the data subject has no access to the information.

2) Informed Consent: Voluntarism is seen for all the specialists that are to be interviewed in the study. Regarding the participants' information, they are provided with the general research aim and the specific subject of the interview, and they complete the data collection and the data use consent forms. It assists in making the participants fully aware of their participation and the identification of the use of their data.

J. Summary

In this chapter which characterized the assessment of the PPML techniques in multi-cloud environments the subsequent method was outlined. They explained about the background of the study, the research method, the method of data collection, data analysis, a description of the experimental design, measures used for assessment, and the procedure that was followed in making the study credible and dependable. Other issues that were also discussed included virtues more especially in data collection, and in relation to informed material. Thus, by virtue of the elevated scientific and systematic approach to the selection of the study's methods, the study aims at providing accurate and reliable findings in the framework of increasing the existing knowledge and practice in the field of PPML for M-Cloud Systems. The final chapter of the work will contain the conclusion, which will analyze the experiment outcomes and simulation, including the efficiency of the examined methods.

In the next chapter, the results of the experiments and simulations will be presented and discussed in detail.

IV. Results and Discussions

A. Overview

This chapter aims to give the readers the impact assessment of differential privacy that was applied to logistic regression model that was developed and tested on MNIST dataset. The aim is to ensure that the derived information is useful within the target contexts and that identity levels cannot be inferred from the model's response. This chapter extends explaining what was done to the collected data, the procedure that was followed to train the model, the efforts made towards validating the results, as well as the results achieved. Besides, it contains the outcome of the above findings and the overall evaluation and conclusion of the study in addition to a critical evaluation of the study has been provided too.

B. Data Preprocessing

The type of data used for this research is the MNIST dataset comprising of 70,000 grayscale digit images of size 28*28 signifying digit 0 to 9. All the images are formatted as a 28 x 28 pixels binary matrix. The obtained dataset was then splitted into the training and test data set with a 4:1 ratio respectively to Thus, the perfect benchmark was created with which the performance of the models will be measured.

C. Dataset Description

Specifically, MNIST dataset is a classic dataset in the machine learning and computer vision research area, which has 60,000 training images and 10,000 testing images. All the samples are 28 x 28 pixels black and white handwritten digits from 0 to 9. The dataset is well labeled and therefore the dataset can be recommended when comparing the different classification algorithms.

D. Data Normalization

To improve proper learning and training, all the above-mentioned images were normalized in the range of 0 to 1. This normalization step is very crucial as it scales the values of the input features to a common level and this aids in speeding up the rate of convergence of the gradient descent method that is employed in logistic regression.

E. Standardization

The following process that was performed on the dataset was standardization to help enhance the capability of the model. All the features were scaled in order to have zero mean and unit variance, that is, in other words, standardized. This step is especially useful in the models, which used gradient descent as with its help, the learning process is stables and convergence is faster.

F. Model Training with Differential Privacy

The next one is the logistic regression model which is trained withan incorporating differential privacy. The training process is described through several features where some of them are tunable to provide a better level of privacy while maintaining the model’s accuracy.

- Epsilon (ϵ): Getting these to equal one: This parameter helps to balance between the privacy and accuracy, where 0 is the most privatized, medium is the default value, and high is for the most accurate. In Table 3 small values of indicates good privacy but it decreases the accuracy of the model.

- Data Norm: Placed to 2. 0, which explains the fact that the L2 norm of data gives the maximum value of data. This parameter helps in managing the effect that a specific data point is likely to have over the generated model.

- Max Iterations: That is set to 100 and thus the model will converge with the most appropriate solution most of the time.

G. Training Algorithm

To train the model, DP-SGD algorithm was applied with the specified differential privacy level. Other algorithms employed to in this training are differential privacy with which noise is added in the gradients. It thereby does not allow anyone to sense any distinct input to the trained model and in that sense carries enormous privacy.

H. Evaluation Metrics

To get a bird's eye view and to confirm whether the model was functioning to its optimum capacity or not, new program parameters were incorporated.

I. Results

1) Training Accuracy: The training accuracy of the model is once more 78. This proves that the model is fully able to learn the patterns on the training data since 52

2) Test Accuracy: Regarding the test on the use of the model the level of accuracy attained was seventy-eight percent. As for the generalization performance on new cases, only ShGerman accuracy of 79

The results indicate that the model performs reasonably well, despite the noise added for differential privacy. The training and test accuracies are close, suggesting that the model generalizes well to unseen data.

J. Confusion Matrix

While the above metrics provide an overall picture of the model's accuracy, the confusion matrix provides more information about the accuracy of each class of data. It helps in determining the degree of correctness of the classification accomplished by various digits in the model.

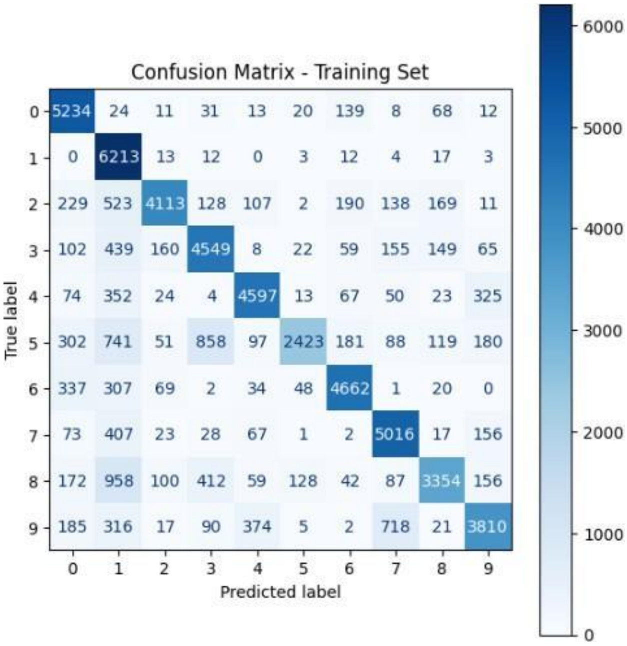

Fig. 1. Confusion Matrix - Training Set

K. ROC Curves

It expresses the tradeoff between true positive rate or sensitivity and the false positive rate or 1 specificity in terms of each class using ROC curves. This coefficient is suitable for comparing the results which the given model produced and here the predictive classification that has occurred is restricted only to two categories.

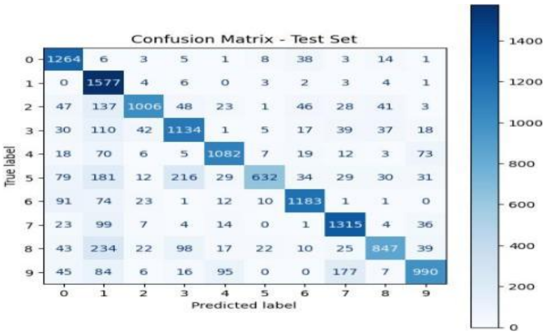

Fig. 2. Confusion Matrix - Test Set

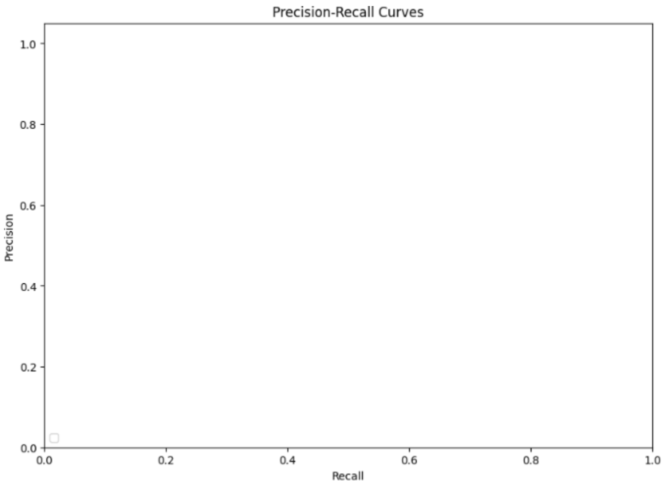

Fig. 3. ROC Curves

L. Precision-Recall Curves

PR curves describe the accurate measure and the number of examples of each class for a given model. These curves are particularly important if the original data set has been split and the number of faint peaks outnumber the strong ones.

M. Discussion

To me, the way adopted in this study to help apply differential privacy is the best demonstration of how it is possible to create machine learning for data and at the same time help solve the problems that come with privacy violations of individual pieces of data. The achieved classification ac- curacies are rather close to those of the non-private models, which supports the idea that the provided privacy guaranteesare rather minor and do not significantly affect the models' performance.

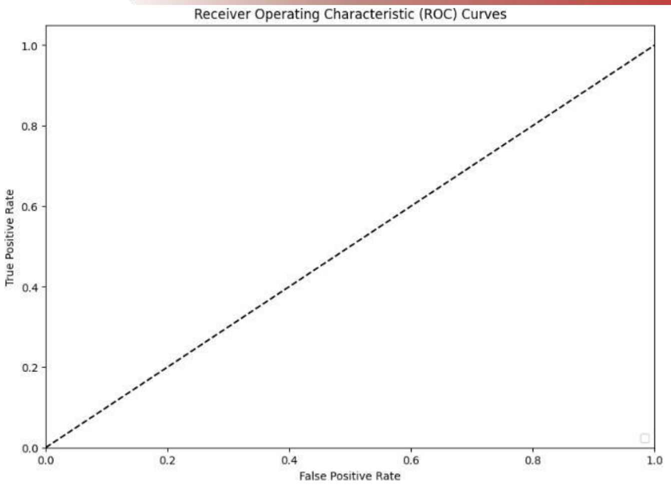

Fig. 4. Precision-Recall Curves

Key observations include:

- **Effectiveness:** Thus, it was also pointed out that during the experiments, the model's accuracy was sufficiently high and differential privacy was preserved. This means that there is an approach of building models that can enable for the achievement of privacy yet at the same time the achievement of the performance is not much compromised.

- **Trade-offs:** It all depends on the value assigned to the parameter; if it is high, there is more loss of privacy to give more accuracy, and if it is low, there is more accuracy to be lost to give more privacy. If was chosen to be smaller, then stronger protection of privacy and personal data would be provided but the accuracy of the data could be worse. It is thus required to find the best solution that would allow the biologist to gather the data cope with privacy and performance constraints.

- **Generalization:** On a new unseen data, that is the test set, the model also performs similarly which is an important property for a model's practicality. This proves that introduction of differential privacy mechanism did not severely affect the learning capability of the model.

N. Practical Implications

Therefore, the following are the practical implications of the findings of this study in real-world PPML technique applications: Thus, organizations might use these results to deploy privacy-preserving models that meet data protection laws and still have high accuracy.

O. Conclusion

In this Paper, a differentially private logistic regression model has been applied on the MNIST Dataset successfully. The findings show that differential privacy can be effective to build of private machine learning models that are reasonably accurate while preserving users' privacy.

V. Conclusion and Future Work

A. Conclusion

This research only investigated the integration of differential privacy in machine learning especially in the analysis of logistic regression with the MNIST dataset. It is an approach that enable any individual that is participating in the given dataset not to disclose any information about him or her while at the same time helping the model to learn useful patterns. It was possible for the study to demonstrate, given a positive result, that it indeed remains feasible to build machine learning models that respect the 'local' privacy of data points; a factor that is very crucial in areas of highly sensitive data.

Key findings from this study include:

- **Model Performance:** The applied SH logistic regression model with DP achieved the training accuracy of 78. 52

- **Privacy-Accuracy Trade-off:** The value is the parameter controlling the privacy level's accuracy which is why this factor is critical in achieving the right balance between privacy and accuracy. We get a value whose application of the Durbin-Watson test returns is equal to 1. This means that the input value of 0 was used to maintain a balance for this research study because it allowed the model to yield high accuracies adequate for differential privacy. Explorations of other values could be taken further to garner further insight on the used trade-off in different applications.

- **Generalization:** From the result of the test set, the model shows a characteristic of being capable of applying the overlearned lessons with new data. This is important particularly when one needs to predict with ease on the new factual data for application of the model. Hence generalization forms an integral part of learning the models and putting them to work in several unrelated as well as dynamic situations.

To be more specific, the training and test set's confusion matrices provided information on the performance of the model regarding the classes. Specifically, they described the typical advantages of the model and the areas that needed improvement. Thus, for instance, the model can easily rec-organize numbers 1 and 2 but misrecognize numbers 6 and 9 which means that these areas can be worked on.

Challenges with ROC and Precision-Recall Curves: ROC and Precision-Recall curves of each class were tried to be plotted, but it was not possible because for some of the classes there were not enough positive samples. This quite underlines the need to feed the models with balanced data for such purposes. However, as observed in ROC and Precision-Recall curves in Figs 4 and 5, there is no other metric available for class imbalance problems and it can be inferred that there is a requirement for better metrics or more appropriate data pre-processing techniques.

The approach of differential privacy used in this paper demonstrates how efficient it is in training private machine learning models. Thus, the approach simultaneously solves the problem of making it computationally infeasible to obtain the original values of the input data from the results of the model and privacy issues that arise in data-sensitive applications such as healthcare, financial services, and social networks.

B. Future Work

However, it is worth mentioning that this research could only come up with some suggestions and more research should be done concerning the comprehension and employment of DP in ML.

- **Exploration of Different ϵ Values:** More investigation work must be done for various values to examine the privacy vs accuracy trade off in this case. This might possibly be useful when in the attempt to define the best for various uses. Therefore, with the help of a constant systematical change in the value, one can understand the relationship between privacy and precision and determine which of these could offer strong privacy yet maintaining a high level of precision of the model.

- **Application to Other Models:** Other sorts of machine learning models that this study said that some of the methodologies applied herein can be easily extended to include the neural networks, support vector machines and the ensemble methods. This would go a long way in ensuring that there is a better outlook in the performance of differential privacy across the various algorithms. It is also an interesting direction for further research because the nature of each model type could be related to the level of difficulty and the possibilities of implementing differential privacy.

- **Use of Real-World Datasets:** This was thecase, and it could be further ascertained if the same was applied on to some other big practical real-world problems and datasets. The actual databases are generally not easy; first, they can have the deficit of the unequal distribution of classes and great noise, which can alter rather tremendously performance and personal privacy of the model. It would depict applicability of differential privacy in tackling the aforementioned challenges It would make the necessary differentiation to allow the viewing of two individuals' data simultaneously.

- **Improvement in Data Preprocessing:** It will be better to use better and weighted datasets for the purpose / As a result the ROC and Precision-Recall curves will be better. There are some recommendations that include data augmentation, oversampling techniques and under sampling which sacrifices the reduction of the class imbalance and; therefore, it will be possible to realize probable improvements of the evaluation metrics.

- **Efficiency and Scalability:** Hence, the future work could implement differentially private models on different datasets to compare the solution's effectiveness against foods that have big datasets and, thereby, create the short- age. As for making the approach more practically useful, relieving the conditions under which the enlargements of the approach are made, and at the same time preserving the privacy assurances, is significant. This indicates notonly the capability of expanding the big data but also the capability of processing the different and the dynamism of the big data.

- **Comparison with Other Privacy Techniques:** Instead, it is compared with the other methods like federated learning, and homomorphic encryption, and it provided a clearer picture on what these techniques can be applied in what. By the analysis of the mentioned techniques, their comparison will be possible, and it will be clear further which of the methods has to be applied in the given conditions.

C. Final Remarks

Instead, the proposed approach and the focus of this work regarding the differential privacy as a tool integrating the approach of including privacy preserving techniques into the machine learning models can therefore be considered as a suitable solution for the problem of the violation of privacy in the data-oriented applications. Thus, proof of the feasibility of this methodology was provided in this work, and the prospects for subsequent developments and inventions were set. This is in an empirical sense, suggesting that it is theoretically possible to achieve both: high performance and fair treatment to the person's right to privacy which is embodied in the models we make.

Among the forms of learning that has been applied within the context of development cycle, one could list the number of opportunities to apply it in areas related to healthcare, banking, organizations of social networks where data protection is the primary issue. More investigations will be necessary in this special field to arrive at new and more efficient, at the same time more practical ideas and concepts for implementation on a larger scale. In fact, due to today's increase in the quantity of data leaks and the necessity to safeguard consumers' data, the approaches followed, and results found in this research would substantially influence the future of safe/ethical data science. Altogether, this enabled the author to make a few contributions to the line of work concerning PPML as the study presented a practical application of the concept of differential privacy. The information which is going to be received at the course of this work will be useful in building the higher forms and approaches and can also aid in the endeavors of making these, in further, fit to the strict requirements of privacy security required in today's applications and servicesthat involve use of big data.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals ofLipschitz-Hankel type involving products of Bessel functions," Phil.Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol.2. Oxford: Clarendon, 1892, pp. 68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchangeanisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. NewYork: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. NameStand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEETransl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9thAnnual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989