ELEVATING PRIVACY: A DIFFERENTIAL PRIVACY INFUSED APPROACH TO GAN FOR ROBUST DATA SYNTHESIS IN DEEP LEARNING MODELS

U18AIP8601 - FELLOWSHIP II REPORT

Submitted by

RITHVIK PRANAO N – 20BAD032
SURIYA PRIYA J M – 20BAD043
YOGESWARI L – 20BAD050
GOKUL S – 20BAD206
SRI NANDHA S S – 20BAD213

in partial fulfillment for the award of the degree
of
BACHELOR OF TECHNOLOGY
IN

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

KUMARAGURU COLLEGE OF TECHNOLOGY
(An Autonomous Institution affiliated to Anna University, Chennai)
Post Box No: 2034, Coimbatore - 641049

APRIL 2024

KUMARAGURU COLLEGE OF TECHNOLOGY
COIMBATORE – 641049.

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

BONAFIDE CERTIFICATE

Certified that this project report **“ELEVATING PRIVACY: A DIFFERENTIAL PRIVACY INFUSED APPROACH TO GAN FOR ROBUST DATA SYNTHESIS IN DEEP LEARNING MODELS”** is the Bonafide work of “RITHVIK PRANAO N – 20BAD032, SURIYA PRIYA J M – 20BAD043, YOGESWARI L – 20BAD050, GOKUL S – 20BAD206, SRI NANDHA S S – 20BAD213” who carried out the project work under my supervision.

SIGNATURE

Dr.P.Shenbagam
HEAD OF THE DEPARTMENT
Department of Artificial Intelligence
and Data Science,
Kumaraguru College of Technology,
Coimbatore – 641049.

SIGNATURE

Dr. S Sangeetha
SUPERVISOR
Department of Artificial Intelligence
and Data Science,
Kumaraguru College of Technology,
Coimbatore – 641049.

The candidates with college roll/ register number 20BAD032, 20BAD043, 20BAD050, 20BAD206, 20BAD213 were examined in the Project Viva-Voce Examination held on …………….

Internal Examiner

External Examiner

DECLARATION

We affirm that the project work titled “ELEVATING PRIVACY: A DIFFERENTIAL PRIVACY INFUSED APPROACH TO GAN FOR ROBUST DATA SYNTHESIS IN DEEP LEARNING MODELS” being submitted in partial fulfillment for the award of B.Tech Artificial Intelligence and Data Science is the original work carried out by us. It has not formed the part of any other project work submitted for the award of any degree or diploma, either in this or any other University.

RITHVIK PRANAO N (20BAD032) **SURIYA PRIYA J M (20BAD043)**

YOGESWARI L (20BAD050) **GOKUL S (20BAD206)**

SRI NANDHA S S (20BAD213)

I certify that the declaration made above by the candidates is true.

Dr. S Sangeetha
Associate Professor,
Department of Artificial Intelligence and Data Science,
Kumaraguru College of Technology,
Coimbatore – 641049.

ACKNOWLEDGEMENT

We express our profound gratitude to the Management of Kumaraguru College of Technology for providing us with the required infrastructure that enabled us to successfully complete the project.

We extend our gratitude to our Principal, **Dr.M.Ezhilarasi,** for providing us with the necessary facilities to pursue the project.

We thank our Director, **Dr. V.R.Raghuveer,** for providing us with the necessary facilities to pursue the project.

We would like to acknowledge **Dr. P. Shenbagam,** Professor and Head, Department of Artificial Intelligence and Data Science, for her support and encouragement throughout this project.

We thank our project coordinator and guide **Dr. S Sangeetha,** Associate Professor**,** class advisor **Dr. D Sudharson,** Assistant Professor, Department of Artificial Intelligence and Data Science, for their constant and continuous efforts, guidance, and valuable time.

Our sincere and hearty thanks to faculty members of Department of Artificial Intelligence and Data Science of Kumaraguru College of Technology for their well wishes, timely help and support rendered to us during our project. We are indebted to our family, relatives, and friends, without whom life would have not been shaped to this level.

RITHVIK PRANAO N (20BAD032)
SURIYA PRIYA J M (20BAD043)
YOGESWARI L (20BAD050)
GOKUL S (20BAD206)
SRI NANDHA S S (20BAD213)

APPENDIX 3

TABLE OF CONTENTS

ABSTRACT

In today's data-driven landscape, the proliferation of deep learning models raises concerns about privacy vulnerabilities, particularly in scenarios where datasets are limited. One significant threat is posed by membership inference attacks, where adversaries exploit model outputs to discern whether specific data points were part of the training set, potentially leading to breaches of privacy. This problem becomes worse when models are trained on minimal data, as they are more susceptible to overfitting and may inadvertently leak sensitive information about individual data points.

To address these challenges, this project proposes a novel approach centered around the generation of synthetic data to augment the training dataset. By increasing the volume of available data, the model's susceptibility to overfitting and membership inference attacks can be significantly mitigated. Central to this solution is the utilization of Layer-wise Relevance Propagation (LRP), a technique that identifies and quantifies the relevance of features within the dataset. Leveraging the insights provided by LRP, the proposed methodology incorporates Differential Privacy Generative Adversarial Networks (DP-GANs) to generate synthetic data with enhanced privacy protections. By adding privacy-preserving noise to the synthetic data during generation, the DP-GANs ensure that sensitive information remains safeguarded while augmenting the training dataset.

The implementation of this solution involves the seamless integration of several cutting-edge technologies and methodologies. Layer-wise Relevance Propagation (LRP) serves as a critical tool for identifying relevant features within the dataset, enabling the extraction of meaningful insights that inform subsequent steps. Additionally, Differential Privacy Generative Adversarial Networks (DP-GANs) play a pivotal role in the data generation process, utilizing the relevance scores provided by LRP as input to add privacy-preserving noise. Together, these technologies form a robust framework for augmenting datasets, preserving privacy, and fortifying deep learning models against membership inference attacks in scenarios with limited data availability.

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATION	**DESCRIPTION**
LRP	Layer-wise Relevance Propagation
GAN	Generative Adversarial Network
DP-GAN	Differential Privacy Generative Adversarial Network
DOM	Document Object Model
SGD	Stochastic Gradient Descent
AI	Artificial Intelligence
API	Application Programming Interface
GPU	Graphics Processing Unit
TPU	Tensor Processing Unit

1. INTRODUCTION

In today's data-driven world, privacy concerns are large, especially with the rise of deep learning models and limited datasets. One significant threat is membership inference attacks, where adversaries exploit model outputs to discern if specific data points were used in training, potentially compromising privacy. This project proposes a novel solution: generating synthetic data to bolster the training dataset, thereby mitigating overfitting and the risk of such attacks. Central to this approach is Layer-wise Relevance Propagation (LRP), identifying crucial features, and Differential Privacy Generative Adversarial Networks (DP-GANs), adding privacy-preserving noise to synthetic data. By seamlessly integrating these technologies, the project aims to fortify deep learning models against membership inference attacks in scenarios with sparse data.

1.1 PROBLEM STATEMENT

Deep learning models face a critical challenge: vulnerability to privacy breaches, especially in scenarios with limited datasets. Membership inference attacks, where adversaries exploit model outputs to determine if specific data points were part of the training set, pose a significant threat to individual privacy. Additionally, overfitting in models trained on minimal data exacerbates the risk of unintentional data leakage.

Given the urgency of safeguarding privacy in deep learning applications, there is a pressing need for innovative solutions that can augment training datasets, mitigate overfitting, and fortify models against membership inference attacks while preserving individual privacy. Therefore, the problem at hand is to develop a comprehensive solution that integrates advanced techniques like Layer-wise Relevance Propagation (LRP) and Differential Privacy Generative Adversarial Networks (DP-GANs) to generate synthetic data, mitigate overfitting, and enhance privacy protections in deep learning models, particularly in scenarios with limited data availability.

1.2 PRODUCT SCOPE

The proposed solution seeks to create a comprehensive framework that effectively addresses privacy vulnerabilities encountered in deep learning models trained on limited datasets. This framework encompasses several key components:

- Development of algorithms and methodologies: The project involves designing algorithms and methodologies for seamlessly integrating Layer-wise Relevance Propagation (LRP) and Differential Privacy Generative Adversarial Networks (DP-GANs) into the deep learning model training pipeline.

- Creation of tools and utilities: The development process includes creating tools and utilities specifically tailored for data augmentation using synthetic data generated by DP-GANs. These tools facilitate the incorporation of synthetic data into the training process, enhancing model robustness and privacy protections.

- Integration of privacy-preserving mechanisms: The framework incorporates privacy-preserving mechanisms to mitigate overfitting and safeguard against membership inference attacks. By strategically applying techniques such as differential privacy, the solution ensures that sensitive information remains protected throughout the training process.

- Implementation of testing and validation procedures: Rigorous testing and validation procedures are conducted to assess the effectiveness and performance of the solution. This includes evaluating the model's resilience to privacy attacks and its ability to maintain accuracy while preserving privacy.
- Documentation and user guides: Comprehensive documentation and user guides are provided to facilitate the seamless integration and utilization of the framework across various deep learning applications.
- Continuous monitoring and updates: The framework is designed to continuously monitor and update to remain compatible with evolving privacy standards and emerging threats. This ensures that the solution remains effective and robust in the face of evolving privacy challenges.

Overall, the product aims to deliver a scalable, adaptable, and user-friendly solution for enhancing privacy protections in deep learning models. By enabling the responsible deployment of these models in real-world scenarios with limited data availability, the framework contributes to advancing the field of privacy-preserving deep learning.

2. LITERATURE REVIEW

In recent years, there has been a surge in the adoption of data-driven technologies, accompanied by an increased emphasis on privacy preservation within deep learning and artificial intelligence systems. Generative Adversarial Networks (GANs) have emerged as powerful tools for generating synthetic data, but concerns about privacy breaches have spurred researchers to explore innovative techniques for privacy preservation. This literature survey aims to scrutinize existing approaches for privacy-preserving GANs and highlight their limitations, with a focus on addressing these drawbacks in our proposed model, which combines Layer-wise Relevance Propagation (LRP) with differential privacy (DP) GANs.

PPGAN[1]: Introduces a privacy-preserving GAN model but faces challenges in balancing privacy preservation with data utility. Additionally, the model may lack scalability for larger datasets due to computational constraints and limitations in handling complex data distributions. pGAN[2]: Aims to generate synthetic data while preserving privacy but struggles with handling continuous or time-series data effectively. It also requires further validation to assess its performance on diverse datasets and may not generalize well to real-world scenarios.

DPMI[3]: Improves the quality of synthetic images but may suffer from increased computational complexity and resource requirements, limiting its scalability for large-scale applications. Additionally, the differential privacy mechanisms employed may impact model scalability and efficiency. MEGAN[4]: Provides direct control over information leakage but risks compromising the utility of generated data, particularly in scenarios with high-dimensional or heterogeneous datasets. Moreover, the model's effectiveness in preserving privacy across various domains needs to be validated through rigorous experimentation.

PATE-GAN[5]: Offers a privacy-preserving approach to synthetic data generation but relies on the original GAN framework, limiting its applicability to diverse datasets and use cases. Furthermore, its performance on real-world datasets needs to be evaluated to assess its practical utility. ECG Synthesis[6]: Successfully generates realistic ECG signals but requires further testing to assess privacy risks effectively. Moreover, the model's performance on diverse ECG datasets should be investigated to ensure its reliability and robustness in real-world applications.

RDP-GAN[7]: Effectively addresses information leakage concerns but faces challenges in balancing overfitting and may require fine-tuning to achieve optimal performance. Additionally, the model's scalability to large-scale datasets needs to be explored to assess its suitability for practical deployment. Adaptive Laplace Mechanism[8]: Preserves privacy but may increase computational overhead and training time, limiting its practical applicability for real-time or resource-constrained environments. Moreover, its effectiveness across different neural network architectures needs to be evaluated to ensure robustness and scalability.

Neuron Noise-Injection Technique[9]: Narrows accuracy gaps but lacks empirical validation for privacy preservation, necessitating further experimentation and validation on diverse datasets. Additionally, its scalability to complex data distributions and high-dimensional feature spaces needs to be investigated. Layer-wise Perturbation[10]: Effective in privacy preservation but requires enhancement for comprehensive privacy protection, particularly in scenarios with deep neural networks and multi-layered architectures. Moreover, its scalability to large-scale datasets needs to be investigated to ensure its practical viability.

Explaining Deep Learning Models[11]: Promising for feature subset selection but needs further verification and validation to assess its effectiveness across different domains and datasets. Additionally, its scalability to high-dimensional datasets should be explored to ensure its practical utility in real-world applications. Ensemble of Random Decision Trees[12]: Enhances performance but faces challenges in complexity and interpretability, particularly in scenarios with large and heterogeneous datasets. Moreover, its effectiveness in preserving privacy across different machine learning tasks needs to be evaluated to ensure its reliability and robustness in real-world applications.

MPCD[13]: Offers enhanced privacy and efficiency but encounters challenges related to complexity and limitations in evaluation, particularly in real-world social network datasets. Additionally, its performance on diverse datasets and use cases should be assessed to ensure its practical applicability. ADPPL[14]: Preserves privacy adaptively but introduces complexity and computational overhead, necessitating further optimization and fine-tuning of hyperparameters. Moreover, its effectiveness across different datasets and domains needs to be evaluated to ensure its reliability and robustness in real-world applications.

In conclusion, this literature survey has provided a comprehensive overview of existing approaches for privacy-preserving Generative Adversarial Networks (GANs). Through Literature Survey, we have identified common challenges and limitations faced by these methodologies, ranging from balancing privacy preservation with data utility to scalability issues and computational complexity.

While each approach offers unique contributions and advancements in privacy preservation, they also exhibit inherent drawbacks that hinder their practical applicability in real-world scenarios. These limitations include challenges in handling diverse datasets, scalability issues, and concerns regarding overfitting and computational overhead.

However, by synthesizing insights from these papers, we have identified opportunities for improvement and addressed these challenges in our proposed model. By integrating Layer-wise Relevance Propagation (LRP) with differential privacy (DP) GANs, our model aims to enhance privacy preservation while maintaining data utility and scalability. Leveraging LRP for feature relevance analysis and explainability, we mitigate the risk of overfitting and ensure robustness across diverse datasets and use cases.

Overall, this literature survey underscores the importance of advancing privacy-preserving GAN methodologies to address the evolving needs of data-driven applications. By addressing the limitations identified in existing approaches, our proposed model offers a promising solution for synthetic data generation while ensuring privacy compliance and data utility. Further research and experimentation are warranted to validate the efficacy and robustness of our proposed model in real-world scenarios.

3. SYSTEM REQUIREMENTS

3.1 SOFTWARE REQUIREMENT SPECIFICATION

Various Software Requirements for SVS Platform are listed below

- Software:
    - ➢ Python
    - ➢ TensorFlow
    - ➢ TensorFlow Privacy
    - ➢ Captum 0.7
    - ➢ Google colab
- Operating System: Windows 7 or higher

3.2 SOFTWARE DESCRIPTION

3.2.1 PYTHON 3.10

Python 3.10 represents a significant milestone in the evolution of the Python programming language, continuing its legacy of simplicity, readability, and versatility. With each new release, Python aims to address the needs and feedback of its vast user base, further solidifying its position as one of the most popular languages for web development, data analysis, machine learning, automation, and more. Python 3.10 builds upon the success of previous versions, introducing innovative features and optimizations that empower developers to write cleaner, more expressive code with improved performance and reliability.

3.2.1.1 Features of PYTHON 3.10

- Python 3.10 introduces structural pattern matching, inspired by similar features in functional programming languages. This powerful new syntax allows developers to perform complex pattern matching operations on data structures such as lists, tuples, and dictionaries, enabling more concise and elegant code for tasks like parsing, data validation, and transformation.

- Python 3.10 enhances its support for type hints, enabling developers to specify more precise type annotations for variables, function parameters, and return values. These improvements facilitate better static analysis and type checking, helping to catch errors and improve code quality early in the development process.

- Python 3.10 includes various security enhancements, such as updates to cryptographic libraries and improvements to the handling of security-sensitive operations. These enhancements bolster the security of Python applications, reducing the risk of vulnerabilities and ensuring the integrity and confidentiality of sensitive data.

3.2.1.2 Advantages of Python 3.10

- With its intuitive syntax and extensive standard library, Python 3.10 enables developers to write code more quickly and efficiently, reducing development time and effort.

- Python 3.10 maintains backward compatibility with previous versions of Python, ensuring that existing codebases can be easily migrated and integrated with the latest features and improvements.

- As an open-source language with a large and active community, Python offers extensive documentation, tutorials, and third-party libraries, providing developers with resources and support to tackle a wide range of projects and challenges.

3.2.2 TensorFlow

TensorFlow 2.15 stands as a pinnacle of innovation in the realm of machine learning frameworks. Developed by Google, it embodies the culmination of years of research and refinement, offering a robust platform for building, training, and deploying state-of-the-art machine learning models. At its core, TensorFlow 2.15 is engineered to provide developers with a seamless and intuitive experience, empowering them to harness the power of artificial intelligence across a diverse range of applications. With its comprehensive suite of tools and libraries, TensorFlow 2.15 facilitates every stage of the machine learning workflow, from data preprocessing to model evaluation, enabling developers to tackle complex challenges with confidence and efficiency.

3.2.2.1 Features of TensorFlow

- TensorFlow 2.15 introduces optimizations and enhancements to improve the speed and efficiency of training and inference, allowing for faster iteration and deployment of models.
- With support for GPUs, TPUs, and other accelerators, TensorFlow 2.15 enables developers to leverage cutting-edge hardware to accelerate computation and handle larger datasets.
- TensorFlow Serving and TensorFlow Lite integration allow for seamless deployment of models in production environments and on resource-constrained devices, ensuring that AI solutions can be deployed at scale.

3.2.2.2 Advantages of TensorFlow

- TensorFlow 2.15 offers a flexible and extensible architecture that supports a wide range of machine learning tasks, from traditional supervised learning to advanced reinforcement learning algorithms.

- With support for interoperability with other popular machine learning libraries such as PyTorch and scikit-learn, TensorFlow 2.15 enables developers to leverage existing tools and workflows, facilitating collaboration and integration with existing systems.

- As an open-source project with a vibrant community of contributors, TensorFlow 2.15 benefits from ongoing development and innovation, ensuring that developers have access to the latest advancements in machine learning research and technology.

3.2.3 TensorFlow Privacy

TensorFlow Privacy 0.9 is an advanced extension of the TensorFlow machine learning framework, dedicated to fortifying privacy protection within model development. This iteration introduces differential privacy techniques seamlessly integrated into the training process, ensuring that sensitive data remains confidential. Its features include privacy-preserving optimization algorithms and customizable privacy budget management, empowering developers to strike a balance between privacy and model utility. With TensorFlow Privacy 0.9, organizations can confidently deploy machine learning models in privacy-sensitive domains, complying with regulations and fostering trust while innovating in privacy-preserving AI applications.

3.2.3.1 Features of TensorFlow Privacy

- TensorFlow Privacy 0.9 integrates differential privacy techniques into the training process, ensuring that sensitive information about individual data points is not inadvertently leaked by the trained model. It makes it easier for us to control DOM (Document Object Model) elements.

- The framework offers privacy-preserving variants of popular optimization algorithms, such as stochastic gradient descent (SGD) and Adam, which enable efficient training of machine learning models while maintaining privacy guarantees.

- Developers have fine-grained control over the privacy budget allocation during training, allowing them to balance between privacy protection and model utility according to their specific requirements.

3.2.3.2 Advantages of TensorFlow Privacy

- By leveraging differential privacy, TensorFlow Privacy 0.9 provides strong privacy guarantees, enabling organizations to build machine learning models that comply with stringent privacy regulations and protect sensitive user data.

- As an extension of the TensorFlow framework, TensorFlow Privacy 0.9 seamlessly integrates with existing TensorFlow workflows, enabling developers to leverage familiar APIs and tools while incorporating privacy protection into their machine learning pipelines.

- With TensorFlow Privacy 0.9, developers can confidently deploy machine learning models in privacy-sensitive domains, such as healthcare and finance, without compromising on privacy or model performance, thus unlocking new opportunities for privacy-preserving AI innovation.

3.2.4 Captum

Captum 0.7 is a cutting-edge interpretability library designed to provide deep insights into machine learning models' decision-making processes. Developed with a focus on transparency and comprehensibility, Captum 0.7 offers a suite of tools and techniques to facilitate the understanding and analysis of complex neural networks. With its intuitive interface and powerful visualization capabilities, this library empowers researchers and practitioners to uncover the inner workings of their models, identify influential features, and debug potential issues with ease.

3.2.4.1 Features of Captum

- Captum 0.7 offers a wide range of attribution methods, including Integrated Gradients, DeepLIFT, and Layer-wise Relevance Propagation (LRP), enabling users to compute feature importances and understand the contribution of individual features to model predictions.
- The library provides rich visualization tools for interpreting model behavior, such as feature attribution heatmaps, gradient visualizations, and saliency maps, facilitating intuitive exploration and analysis of model decisions.
- Captum 0.7 is designed to be model-agnostic, supporting various deep learning frameworks, including PyTorch and TensorFlow, allowing users to interpret models built with different architectures and frameworks seamlessly.

3.2.4.2 Advantages of Captum

- By providing insights into model decision-making processes, Captum 0.7 enhances the transparency and trustworthiness of machine learning models, enabling stakeholders to understand model behavior and make informed decisions.

- With its powerful debugging capabilities, including the ability to identify influential features and diagnose model biases, Captum 0.7 empowers users to debug and improve model performance effectively.
- Captum 0.7 accelerates the pace of research in the field of interpretability by offering a comprehensive set of tools and techniques for analyzing and interpreting deep learning models, fostering innovation and discovery in machine learning.

3.2.5 Google Colab

Google Colab is a cloud-based platform provided by Google that allows users to write and execute Python code in a collaborative environment. With Google Colab, users can access and run code through their web browser without the need for any special setup or installation. It provides a Jupyter notebook interface, making it easy to write and execute code in a step-by-step manner. Users can leverage powerful libraries and frameworks such as TensorFlow, PyTorch, and scikit-learn for machine learning and data analysis tasks. Google Colab offers integration with Google Drive, allowing seamless access to files and datasets stored in the cloud.

3.2.5.1 Features of Google Colab

- Multiple users can work on the same Colab notebook simultaneously, enabling collaborative coding and real-time collaboration.
- Google Colab provides free access to GPU (Graphics Processing Unit) and TPU (Tensor Processing Unit) resources, allowing users to accelerate computations for machine learning tasks.

- Google Colab comes with pre-installed popular libraries such as TensorFlow, PyTorch, and NumPy, eliminating the need for users to install them manually.

3.2.5.2 Advantages of Google Colab

- Google Colab can be accessed from any device with an internet connection and a web browser, making it convenient for users to work on their projects from anywhere.
- Google Colab is free to use, providing users with access to powerful computing resources without the need to invest in expensive hardware or infrastructure.
- Google Colab seamlessly integrates with other Google services such as Google Drive, allowing users to easily import and export data from their cloud storage.

4. PROPOSED SYSTEM

4.1 WORKFLOW

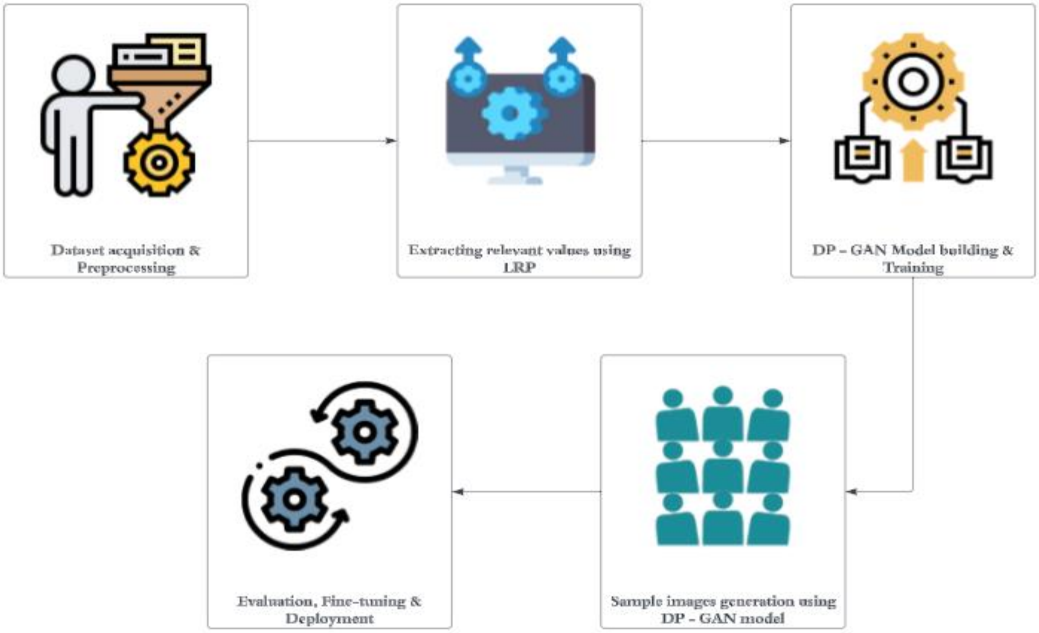

Dataset acquisition &
Preprocessing
Extracting relevant values using
LRP
DP – GAN Model building &
Training
Sample images generation using
DP – GAN model
Evaluation, Fine-tuning &
Deployment

Fig 4.1: Workflow

Figure 4.1 outlines a comprehensive workflow for the project, meticulously crafted to optimize efficiency, transparency, and overall effectiveness. This detailed analysis delineates each step of the project, from dataset acquisition to model evaluation, ensuring a systematic approach to handling data and implementing privacy-preserving techniques. By leveraging Layer-wise Relevance Propagation (LRP) and Differential Privacy Generative Adversarial Networks (DP-GANs), the project aims to extract relevant features, generate synthetic data, and evaluate model performance. This structured workflow not only facilitates the seamless integration of LRP and DP-GANs but also enables the project team to derive actionable insights from the generated synthetic data. Ultimately, this systematic approach empowers the project with the ability to preserve privacy while enhancing data utility, thereby supporting informed decision-making and advancing the field of privacy-preserving machine learning.

4.1.1 Dataset Acquisition

The dataset acquisition process involves leveraging the Kaggle API to access and integrate the large dataset seamlessly into the solution framework. By utilizing the Kaggle API, the solution can programmatically retrieve dataset directly from the Kaggle platform without the need for manual downloading. This streamlined approach not only simplifies the dataset acquisition process but also ensures access to up-to-date and high-quality datasets that may be too large or complex to be stored locally. Additionally, the integration of the Kaggle API enables automated data retrieval and updates, facilitating efficient experimentation and model training with fresh datasets.

4.1.2 Data Preprocessing

The data preprocessing step involves transforming raw data into a format that is suitable for analysis and model training. This process typically includes several key tasks such as data cleaning, feature selection, normalization, and encoding categorical variables. Data cleaning involves handling missing values, outliers, and inconsistencies in the dataset to ensure its integrity and reliability. Feature selection aims to identify and retain relevant features while removing redundant or irrelevant ones to reduce dimensionality and improve model performance. Normalization techniques are applied to scale numerical features to a common range, preventing certain features from dominating others during model training.

4.1.3 Extracting Relevant Features using LRP

This step involves analyzing the contribution of each feature to the model's output. LRP is a technique used to interpret the predictions of deep neural networks by attributing relevance scores to input features based on their impact on the model's decision-making process. In this process, LRP propagates relevance scores, assigning higher scores to features that have a greater influence on the final prediction. By examining these relevance scores, one can identify which features are most significant in driving the model's predictions and gaining insights into the underlying patterns and relationships in the data. This feature extraction step using LRP enables the selection of relevant features that are crucial for preserving privacy while maintaining the utility of the data. The extracted relevance scores serve as a basis for determining the amount of noise to be added during the privacy-preserving data generation process, ensuring that sensitive information is protected while generating synthetic data that retains key characteristics of the original dataset.

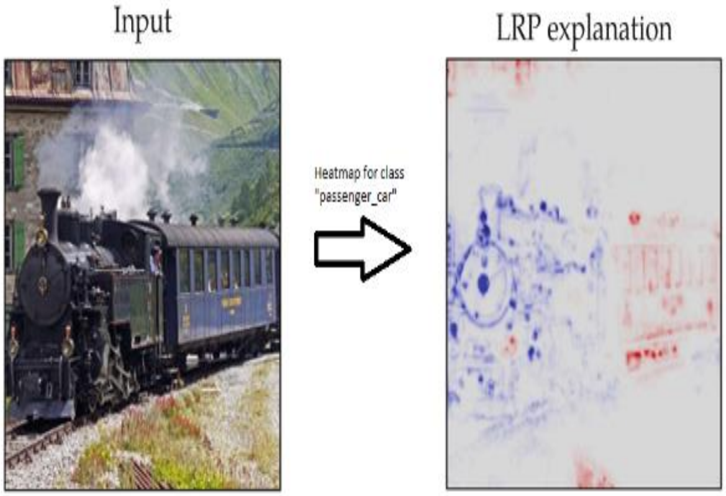


Fig 4.2: Working of LRP

4.1.4 DP – GAN Model

The construction and training of the Differential Privacy Generative Adversarial Network (DP-GAN) involves several key steps aimed at preserving privacy while generating synthetic data. Initially, the DP-GAN architecture, comprising a generator and a discriminator network, is established. The generator aims to produce synthetic data samples resembling the original dataset, while the discriminator is trained to distinguish between real and synthetic data. During training, the relevance scores obtained from Layer-wise Relevance Propagation (LRP) are utilized to determine the amount of noise to be added to the gradients of the discriminator and generator networks. This noise, modeled as Gaussian noise with a specific variance derived from the relevance scores, is incorporated into the gradient updates of the networks to ensure that sensitive information is protected while generating synthetic data. In this gaussian noise addition method, two important parameters epsilon and sensitivity are required. Totally seven epsilon values [0.01,0.1,0.5,1,5,10,15] are multiplied with the average of tensor values obtained from the saliency method from the Layer wise Relevance Propagation algorithm implemented to get the most contributing features of the images. The sensitivity is the square of the difference of the maximum tensor value and the minimum tensor value. For every epsilon value, the training takes place. Finally, the epsilon value that produces the lowest of the generator losses is chosen as the best parameter.

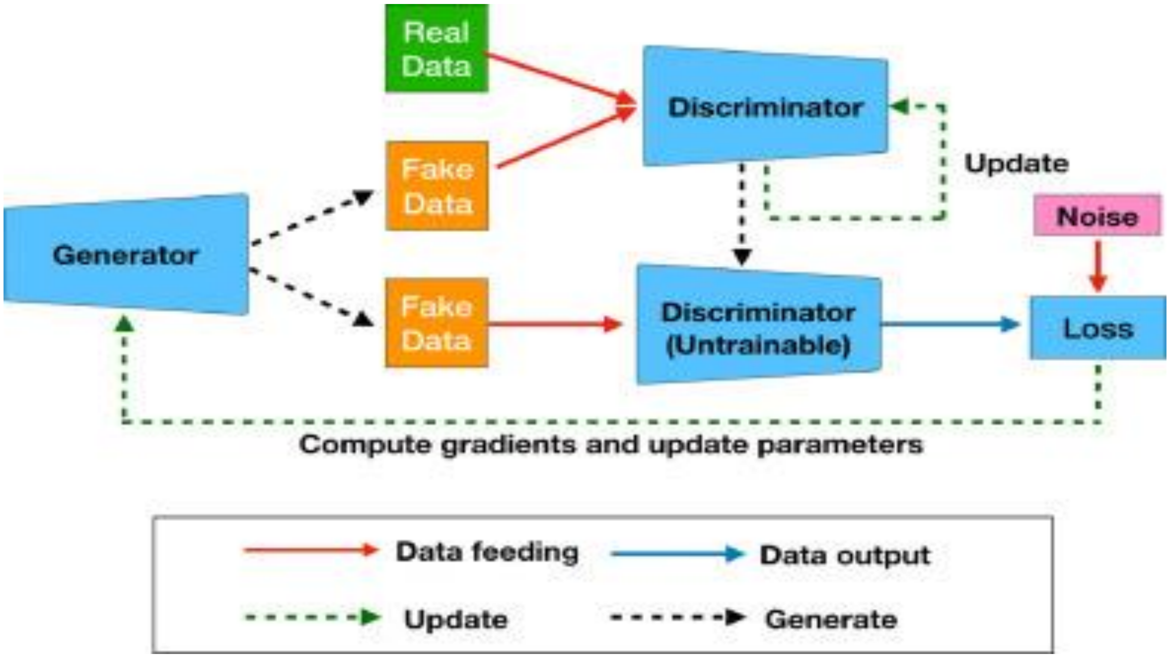


Fig 4.3: DP – GAN Architecture

4.1.5 Sample Generation

Through iterative training, the DP-GAN learns to generate synthetic data that closely resembles the original dataset in terms of statistical properties and key features, providing a privacy-preserving solution for generating realistic samples for various applications such as data augmentation, privacy-preserving data sharing, and synthetic data generation for machine learning models.

4.1.6 Evaluation and Fine Tuning

After training the DP-GAN model, the next crucial step involves evaluating its performance and fine-tuning parameters to optimize results. Evaluation typically entails assessing the quality of the generated synthetic images by comparing them with real data samples. Additionally, the DP-GAN model undergoes fine-tuning which involves experimenting with different settings, such as learning rates, batch sizes, and noise levels, to enhance the model's effectiveness in generating high-quality synthetic data while preserving privacy. Fine-tuning aims to strike a balance between privacy preservation and data utility, ensuring that the generated samples maintain sufficient realism and statistical properties for downstream tasks. Through rigorous evaluation and fine-tuning, the DP-GAN model can achieve improved performance, thus making it suitable for various privacy-preserving applications.

4.1.7 Algorithm

Input: Training dataset $D = \{x1, x2, . . . , xn\}$, privacy budget ϵ, relevance score β, the number of epochs E, discriminator loss function LD, generator loss function LG.

```
for $j \in [1, d]$ do
    Calculate $\beta$ of the jth feature using LRP.
end for
Initialize the discriminator D and generator G networks.
for epoch $\in [1, E]$ do
    for each sample xi in D do
        Compute the gradients $\nabla LD$, $\nabla LG$.
        Add Gaussian noise: $\nabla LD'=\nabla LD+Gaussian(0,\sigma^2)$,$\nabla LG'=\nabla LG+Gaussian(0,\sigma^2)$.
        Update D and G using the noisy gradients: $D \leftarrow D-\eta\nabla LD'$,$G \leftarrow G-\eta\nabla LG'$.
    end for
```

Output: Trained discriminator and generator networks D and G.

5. RESULTS

The average of the tensors obtained from the saliency method of Layer wise Relevance Propagation algorithm is passed as a noise multiplier in DP GAN algorithm which is a major factor for generating privacy. During the training of DP GANs, adaptive noise addition is done to the data wherein epsilon and sensitivity are two major parameters. The standard epsilon values chosen are 0.01, 0.1, 0.5, 1, 5, 10, 15, the product of these and the mean saliency value are used, and sensitivity is taken the square of the difference between the maximum and minimum tensors of the mean saliency.

The following images are the results for different epsilon values:

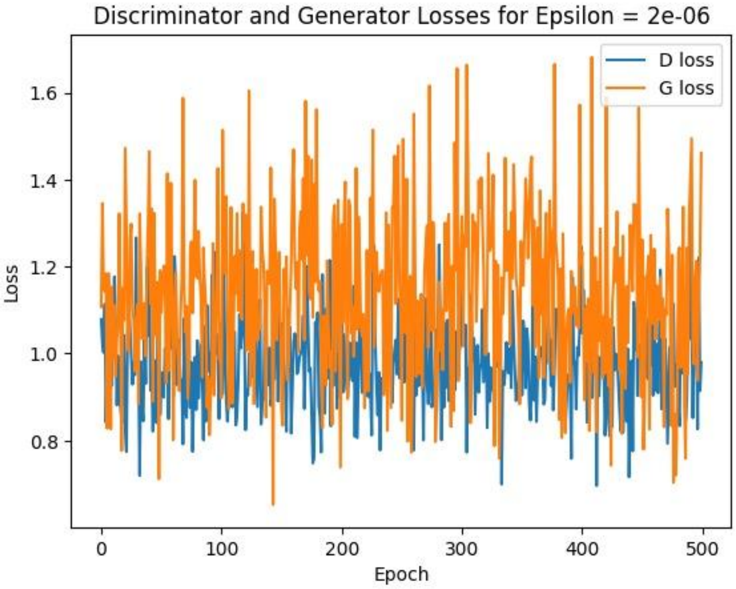


Fig 5.1 : Discriminator and Generator Losses for Epsilon = 2e-06

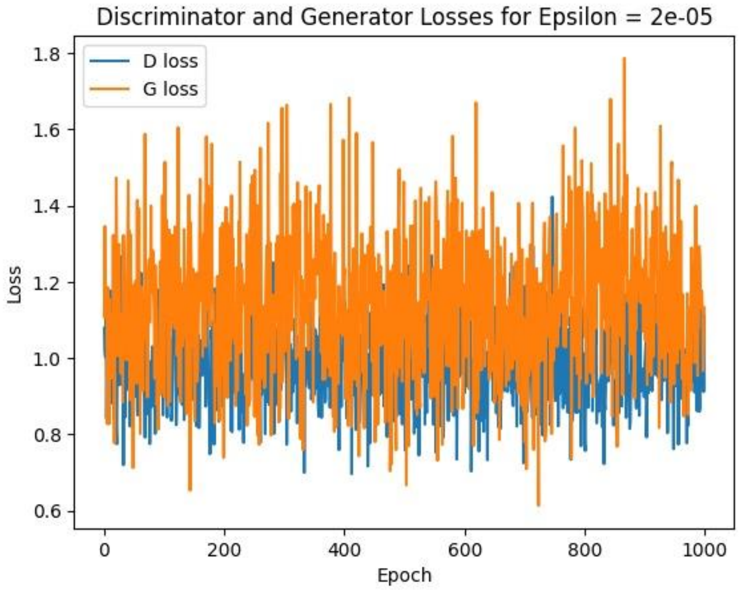


Fig 5.2 : Discriminator and Generator Losses for Epsilon = 2e-05

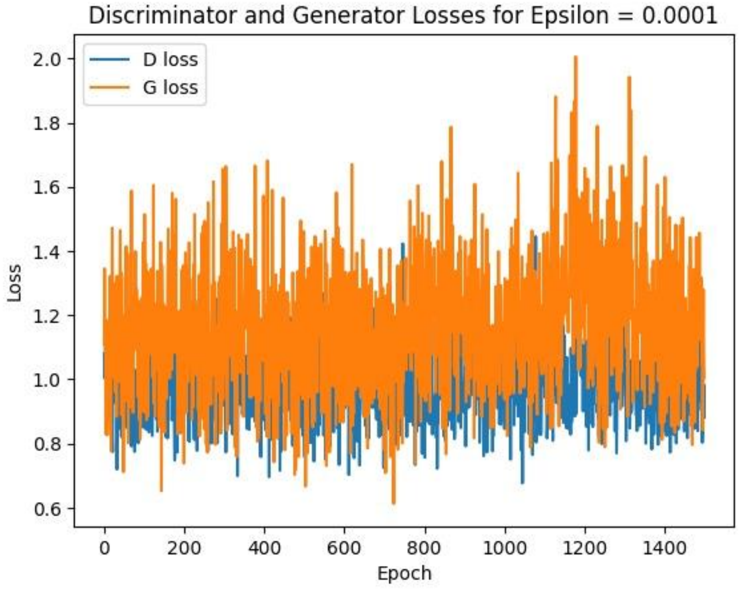


Fig 5.3 : Discriminator and Generator Losses for Epsilon = 0.0001

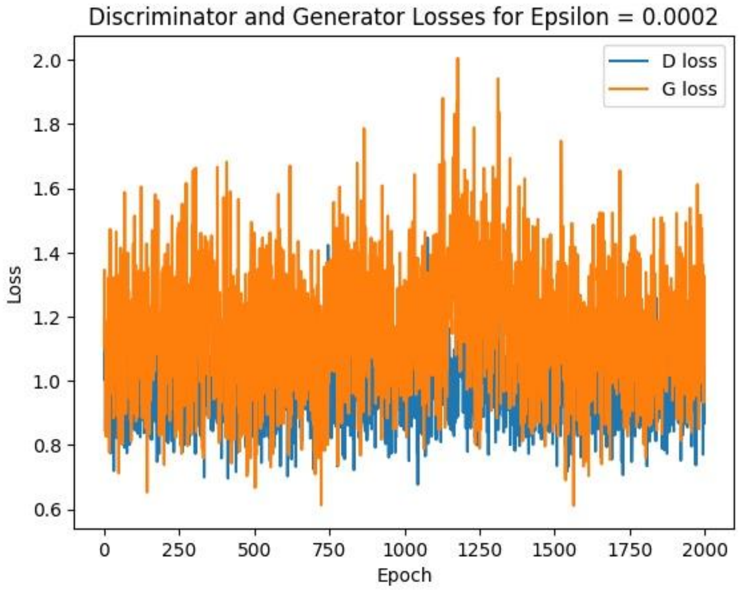


Fig 5.4 : Discriminator and Generator Losses for Epsilon = 0.0002

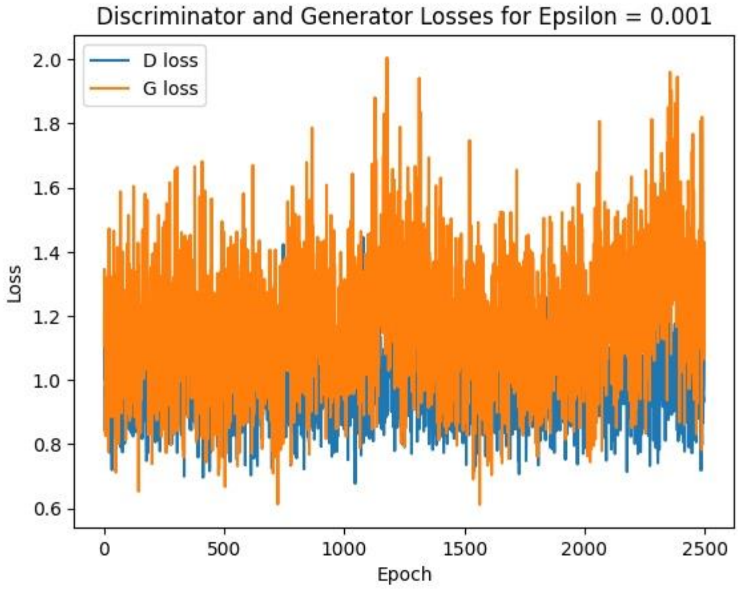


Fig 5.5 : Discriminator and Generator Losses for Epsilon = 0.001

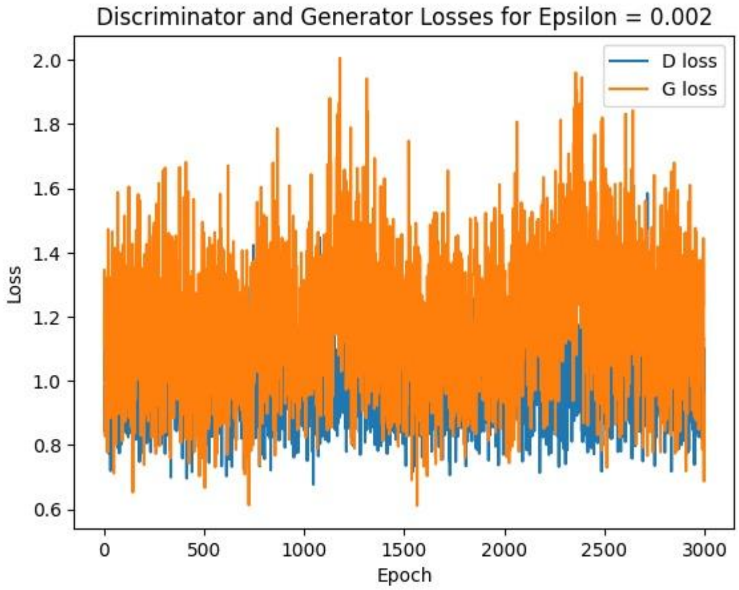


Fig 5.6 : Discriminator and Generator Losses for Epsilon = 0.002

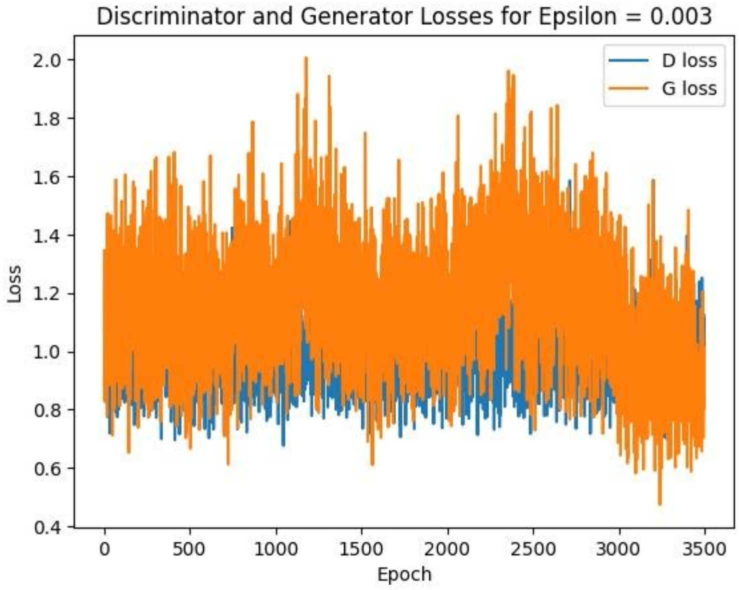


Fig 5.7 : Discriminator and Generator Losses for Epsilon = 0.003

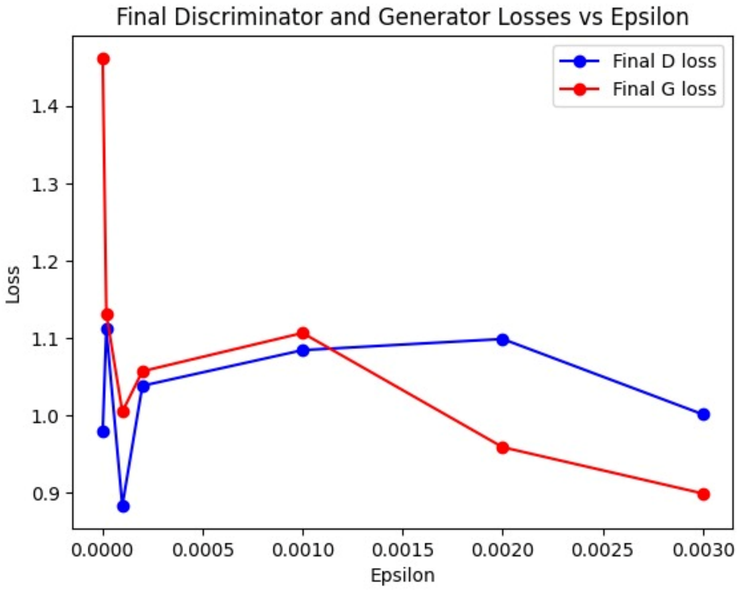


Fig 5.8 : Discriminator and Generator Losses for every Epsilon values

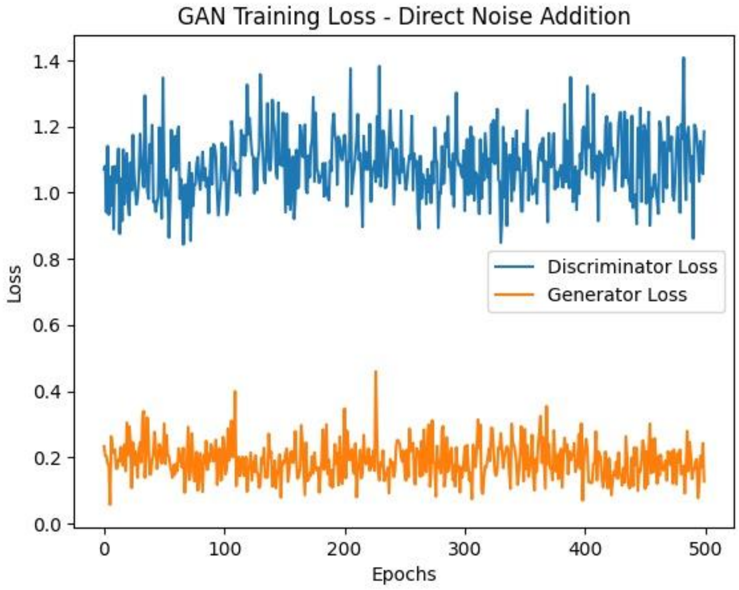


Fig 5.9 : GAN Training Losses for Direct Noise Addition

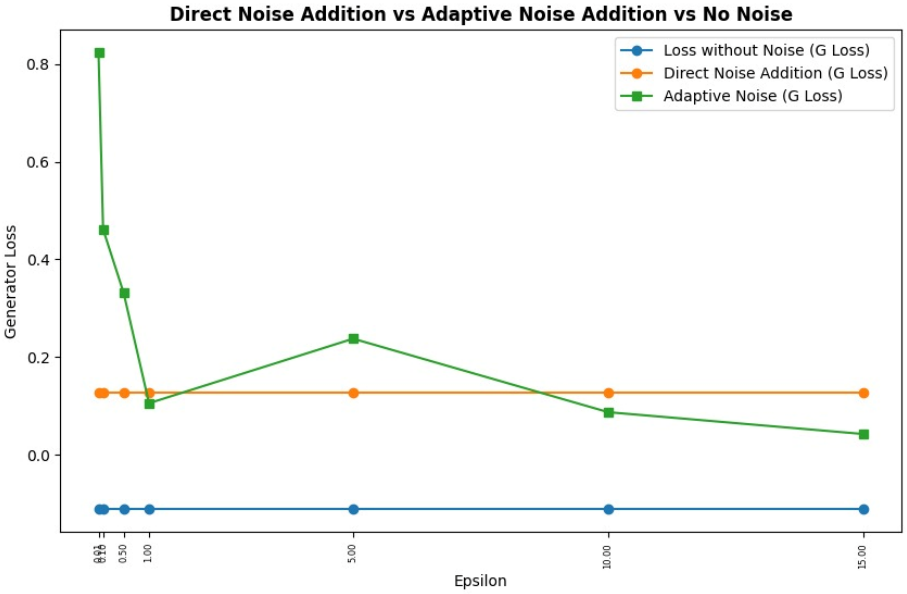


Fig 5.10 : Generator Loss during Direct Noise Addition, Adaptive Noise Addition, and No Noise

6. CONCLUSION

In conclusion, this project presents a robust framework for privacy-preserving machine learning, with a particular focus on mitigating membership inference attacks through the synthesis of synthetic data. By leveraging innovative techniques such as Layer-wise Relevance Propagation (LRP) and Differential Privacy Generative Adversarial Networks (DP-GANs), the project demonstrates the feasibility of enhancing data privacy while maintaining data utility. Through meticulous dataset acquisition, preprocessing, and model training, the project achieves a balance between privacy preservation and model performance, offering a scalable solution applicable to various domains. Among the experimented epsilon values [0.01,0.1,0.5,1,5,10,15], the epsilon value of 15 gives the lowest generator noise and all the respective generator loss values are lower than the generator loss through direct noise method. Hence adaptive noise addition (Gaussian noise addition) gives the best noise.

Furthermore, the project's systematic workflow and transparent methodology lay the groundwork for future advancements in privacy-preserving machine learning. By addressing the critical challenge of membership inference attacks, the project contributes to the development of more resilient and privacy-conscious machine learning systems. Moving forward, continued research and refinement of these techniques will be essential to further bolstering data privacy and security in the era of data-driven decision-making.

7. REFERENCES

[1] Liu, Y., Peng, J., James, J.Q. and Wu, Y., 2019, December. PPGAN: Privacy-preserving generative adversarial network. In 2019 IEEE 25Th international conference on parallel and distributed systems (ICPADS) (pp. 985-989). IEEE.

[2] Venugopal, R., Shafqat, N., Venugopal, I., Tillbury, B.M.J., Stafford, H.D. and Bourazeri, A., 2022. Privacy preserving generative adversarial networks to model electronic health records. Neural Networks, 153, pp.339-348.

[3] Chen, D., Cheung, S.C.S., Chuah, C.N. and Ozonoff, S., 2021, December. Differentially private generative adversarial networks with model inversion. In 2021 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-6). IEEE.

[4] Shateri, M., Messina, F., Labeau, F. and Piantanida, P., 2023. Preserving privacy in GANs against membership inference attack. IEEE Transactions on Information Forensics and Security.

[5] Jordon, J., Yoon, J. and Van Der Schaar, M., 2018, September. PATE-GAN: Generating synthetic data with differential privacy guarantees. In International conference on learning representations.

[6] Delaney, A.M., Brophy, E. and Ward, T.E., 2019. Synthesis of realistic ECG using generative adversarial networks. arXiv preprint arXiv:1909.09150.

[7] Ma, C., Li, J., Ding, M., Liu, B., Wei, K., Weng, J. and Poor, H.V., 2023. RDP-GAN: A Rényi-differential privacy based generative adversarial network. IEEE Transactions on Dependable and Secure Computing.

[8] Phan, N., Wu, X., Hu, H. and Dou, D., 2017, November. Adaptive laplace mechanism: Differential privacy preservation in deep learning. In 2017 IEEE international conference on data mining (ICDM) (pp. 385-394). IEEE.

[9] Adesuyi, T.A. and Kim, B.M., 2020. A neuron noise-injection technique for privacy preserving deep neural networks. Open Computer Science, 10(1), pp.137-152.

[10] Adesuyi, T.A. and Kim, B.M., 2019, February. A layer-wise perturbation based privacy preserving deep neural networks. In 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) (pp. 389-394). IEEE.

[11] Ullah, I., Rios, A., Gala, V. and Mckeever, S., 2021. Explaining deep learning models for tabular data using layer-wise relevance propagation. Applied Sciences, 12(1), p.136.

[12] Wu, X., Qi, L., Gao, J., Ji, G. and Xu, X., 2022. An ensemble of random decision trees with local differential privacy in edge computing. Neurocomputing, 485, pp.181-195.

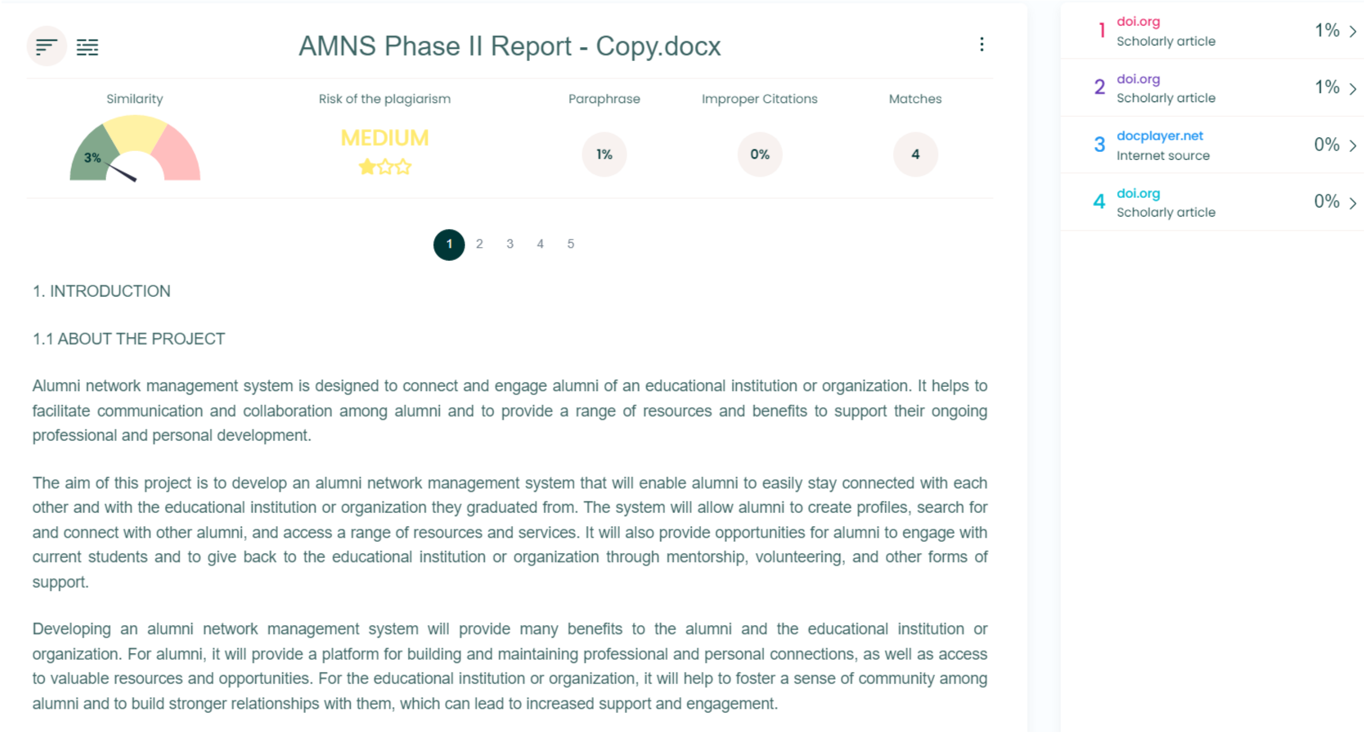
[13] Huang, H., Yan, Z., Tang, X., Xiao, F. and Li, Q., 2022. Differential privacy protection scheme based on community density aggregation and matrix perturbation. Information Sciences, 615, pp.167-190.

[14] Gong, M., Pan, K., Xie, Y., Qin, A.K. and Tang, Z., 2020. Preserving differential privacy in deep neural networks with relevance-based adaptive noise imposition. Neural Networks, 125, pp.131-141.

[15]https://colab.research.google.com/drive/1KLvn7xLBI4kdUXsotXsNPfi5_N61X5Y?usp=sharing

[16]https://colab.research.google.com/drive/1ltZvaBpixVmRkel3LpemsdEhWj75F8G?usp=sharing

8. PLAGIARISM REPORT



AMNS Phase II Report - Copy.docx

1	doi.org Scholarly article	1%
2	doi.org Scholarly article	1%
3	docplayer.net Internet source	0%
4	doi.org Scholarly article	0%

1. INTRODUCTION

1.1 ABOUT THE PROJECT

Alumni network management system is designed to connect and engage alumni of an educational institution or organization. It helps to facilitate communication and collaboration among alumni and to provide a range of resources and benefits to support their ongoing professional and personal development.

The aim of this project is to develop an alumni network management system that will enable alumni to easily stay connected with each other and with the educational institution or organization they graduated from. The system will allow alumni to create profiles, search for and connect with other alumni, and access a range of resources and services. It will also provide opportunities for alumni to engage with current students and to give back to the educational institution or organization through mentorship, volunteering, and other forms of support.

Developing an alumni network management system will provide many benefits to the alumni and the educational institution or organization. For alumni, it will provide a platform for building and maintaining professional and personal connections, as well as access to valuable resources and opportunities. For the educational institution or organization, it will help to foster a sense of community among alumni and to build stronger relationships with them, which can lead to increased support and engagement.