Data Privacy Preservation using Differential Privacy and Re- Identification Attacks

G. Sathish Kumar
Center for Computational Imaging and Machine Vision,
Assistant professor
Department of Artificial Intelligence and Data Science
Sri Eshwar College of Engineering Coimbatore, India
saathhish@gmail.com

Preethie K
Artificial Intelligence and Data Science
Sri Eshwar College of Engineering Coimbatore, India
kpreethie@gmail.com

Sushma R
Artificial Intelligence and Data Science
Sri Eshwar College of Engineering Coimbatore, India
sushma.r2020aids@sece.ac.in

Madhumitha S
Artificial Intelligence and Data Science
Sri Eshwar College of Engineering Coimbatore, India
madhumitha.s2020aids@sece.ac.in

Abstract— In today's data-driven world, machine learning holds immense potential for innovation and discovery, but it comes at a cost—individual privacy. The need to balance insights of data with the protection of privacy information rised to the privacy preserving methods. This paper explores powerful techniques, Differential Privacy and Re-identicatification attack mechanisms, as means to safeguard privacy while enabling meaningful data analysis. Differential Privacy quantifies and controls privacy risks in data analysis,. We discuss the strengths and limitations of both techniques and propose a hybrid approach to leverage their synergies. This hybrid solution is particularly valuable in contexts where strong privacy guarantees are paramount, such as medical research, finance, and confidential data analysis. As data privacy concerns grow, the integration of Differential Privacy and re-identification methods generates a promising pathway to unlocking the full potential of machine learning while preserving individual privacy.

Keywords: *Differential privacy, Data privacy, Machine Learning, Re-identification attacks, Data-driven, Medical data, Re-construction.*

I. INTRODUCTION

In the era of data-driven machine learning, where insights and knowledge are gleaned from vast and diverse datasets, the issue of individual privacy stands as an ever-looming concern. The use of personal data for model training and analysis has raised profound questions about data security and privacy. Organizations and researchers must navigate a delicate balance between harnessing the power of data and safeguarding the confidentiality of personal information. This challenge has rised to the domain of privacy-preserving models, which seeks to enable meaningful analysis while ensuring the privacy of individuals whose data contributes to these insights.

This research paper delves into two powerful strategies for addressing the privacy conundrum: Differential Privacy (DP) provides a formal standard mathematical approach for measuring and managing theft risks and the releases of the data, while it offers the intriguing ability to perform optimizations on open data without exposing its underlying information. Each approach has its unique strengths and limitations.

Moreover, this study proposes the combinational way that artfully strengths the Differential Privacy.. This synergy is particularly valuable in scenarios where maintaining a harmonious rational balance between a privacy and integral part of utility is paramount, in the fields of healthcare, finance, and confidential data analysis.

As we navigate the landscape that is evolving of data privacy and security, the integration of Differential Privacy represents a promising frontier for unleashing the full potential of machine learning while respecting the privacy rights of individuals. This paper sets the stage for a deeper exploration of these two crucial techniques and the innovative possibilities they bring to the forefront of data science and machine learning.

II. LITERATURE SURVEY

In this study[1] presents a Phase, Guarantee, and Utility (PGU) triangle based paradigm to comprehend and assist the assessment of different PPML solutions by breaking down their privacy-preserving functions. It also systematically evaluates and summarises existing methods to privacy preservation. It talks about the special qualities and

difficulties of PPML and suggests future areas of inquiry that draw from and advance several academic fields, including machine learning, distributed systems, security, and privacy. In this study[2], The basic ideas of Fully Homomorphic Encryption (FHE), real-world applications, cutting-edge techniques, constraints, benefits, drawbacks, possible uses, and neural network-focused development tools are all examined in this study. It lists available fixes, unresolved problems, difficulties, chances, and possible lines of inquiry.

In this paper [3], The basic ideas of Fully Homomorphic Encryption (FHE), real-world applications, cutting-edge techniques, constraints, benefits, drawbacks, possible uses, and neural network-focused development tools are all examined in this study. It lists available fixes, unresolved problems, difficulties, chances, and possible lines of inquiry. In this study [4], In this work, the standardized ResNet-20 is built using the bootstrapping mechanisms with RNS-CKKS FHE, and its implementation is validated using the plaintext model parameters and the CIFAR-10 dataset. It confirms quantitatively that, when using non-encrypted data, the suggested model performs 98.43%. The suggested model's classification accuracy is 92.

This study [5], investigates the relationship between privacy preserving methods and machine learning. It talks about the special qualities and difficulties of privacy-preserving machine learning (PPML) and suggests potential paths of inquiry. In this study [6],through careful subcategory organisation, this page offers a comprehensive review of fresh perspectives and methodical interpretations of existing material. The core ideas of the current privacy-preserving data mining techniques, together with their benefits and drawbacks, are presented.

This study [7], addresses the significance of data mining privacy preservation. They point out that the reason privacy-preserving data mining has gained popularity is because it makes it possible to share data that is sensitive to privacy for analytical purposes34. Additionally, they examine the benefits and drawbacks of several privacy-preserving data mining technologies. In this study [8], This study presents a Phase, Guarantee, and Utility (PGU) triangle based paradigm to comprehend and guide the assessment of different PPML solutions by breaking down their privacy-preserving functionalities1. It also thoroughly examines and summarises existing privacy-preserving techniques.

In this study [9], HE and FHE systems are the subject of this survey. It covers the fundamentals of HE as well as the specifics of the well-known, which are crucial building blocks for reaching FHE23. This study [10], address the significance of data mining privacy preservation. They point out that the popularity of privacy-preserving data mining has grown since it permits the sharing of data that is sensitive to privacy for analytical purposes. Additionally, they examine the benefits and drawbacks of several privacy-preserving data mining technologies.

In this study [11], We provide a method based on local differential privacy (LDP) to safeguard the confidentiality of energy usage information. Using the datasets, the utility measurements, some LDP tools, we assess the performance of our LDP solution. The outcomes demonstrate that, with strong privacy assurances, our method can achieve high accuracy and minimal estimation error. In this study [12], We present a re-identification risk metric in this study that we refer to directly prevented by design in the PIE. The adversary's lowest re-identification error probability, or Bayes error probability, is lower-bound by it. We examine the relationship between LDP and the PIE as well as the value of the PIE in distribution estimates for the two obfuscation techniques that produce LDP. By means of experiments, we demonstrate that LDP can lead to excessive obfuscation and ruin the usefulness when we take re-identification as a privacy issue. We subsequently demonstrate that the PIE may be utilised to provide low re-identification risks with great value for the local obfuscation methods.

In this study [13], Although sharing data is necessary, data custodians have a moral and legal duty to protect confidentiality, which includes protecting the data theft. For releasing sanitised reports, this study suggests a sanitization of data techniquewhich fulfils ε-differential methodology. Additionally, the suggested technique lowers the chance that the cleaned data will be re-identified. Two distinct data sets were used in the implementation and testing of the suggested technique. The suggested algorithm's performance is promising when compared to other previous efforts. In this study [14], Maintaining patient privacy appears to be a critical task for the healthcare centre. To protect privacy, a variety of methods are employed, including cryptography, anonymization, and disturbance. One well-known and useful solution to this issue is anonymization. Many techniques for anonymization have been put forth by scholars. This study proposes an enhanced method based on differential privacy and k-anonymity techniques. The suggested method's goal is to use generalisation and suppression approaches to more successfully guard against linking attempts that might re-identify the dataset.

In this study [15], Maintaining patient privacy appears to be a critical concern for the healthcare centre. To protect privacy, a variety of methods are employed, including cryptography, anonymization, and disruption. One well-known and useful solution to this issue is anonymization. The suggested method's goal is to use generalisation and suppression approaches to more successfully guard against linking attempts that might re-identify the dataset. In this study [16], the potential impact of privacy-preserving algorithms on the creation of mechanisms for strategic agents, which need to incentivize players to frank report information by preventing the disclosure of particular participant data. Specifically, we demonstrate how the novel concept of differentiating privacy may guarantee that participants have little influence over the mechanism's conclusion and, hence, little motivation to lie, in addition to its inherent virtue. We introduce and examine a generalisation of prior privacy work that takes into account the high sensitivity in the auction setting, where a single participant may significantly alter the optimal fixed price and a small change in the offered price may take the revenue. This is a crucial step before creating an auction mechanism that maintains privacy.

In this study [17], The method is different from most (but not all) of the relevant literature in the fields of theory, cryptography, databases, and statistics in that it defines a formal and ad omnia privacy guarantee and presents data analysis approaches that are rigorously shown to meet it. Differential privacy has emerged as the primary privacy assurance. In general, this guarantees that entering a statistical database has (almost, and quantifiably) no risk. We review the definition of differentiating privacy in this survey, along with two fundamental methods for obtaining it. We next demonstrate several intriguing uses of these methods, introducing three general results on differentially private learning as well as algorithms for three particular tasks.

In this study [18], The vast amount of data generated from the many sources may be processed and analysed to aid in decision-making. However, data analytics are susceptible to privacy violations. Recommendation systems are a popular type of data analytics framework that e-commerce companies like Flip Kart and Amazon use to suggest products to customers based on their past purchases and other characteristics. This study offers a variety of privacy conservation strategies that are currently being used by researchers, including data anonymization, randomness, generalisation, permutation, and more. We also examine the discrepancy between different procedures and privacy-preserving techniques, and we demonstrate how to resolve such problems with novel and creative approaches. The investigation concludes with a summary of the overall literature's results.

In this study [19], Proposing a perturbation-based privacy-preserving method is the aim. Principal component analysis and random projection are used here to modify the data. This is primarily because feature selection in conjunction with dimension reduction would result in more effective perturbation of the records. By selecting pertinent features, reducing the dimensionality of the data, and shortening the training period, the hybrid technique improves classification performance as shown by metrics such as mean absolute error, kappa statistics, and accuracy. In terms of classification accuracy, which rises from 63 percent to 68 percent, the suggested method beats all other methods, demonstrating its efficacy in identifying cardiovascular disease.

In this study [20], Techniques for anonymization or de-identification allow persons in sensitive data sets to have their privacy protected while the data sets' usefulness is maintained. As it is simpler to analyse enormous data sets, there have been recurrent attacks on the effectiveness of these technologies. Numerous academics have demonstrated that anonymised data may be reidentified using techniques like "linking" to discover the identity of the data participants. In order to minimise privacy infractions, we outline the issues that still need to be solved in this research, which surveys the anonymity landscape of options for tackling re-identification. We also examine a number of legal guidelines for disclosing personal information as well as the instruments used to carry them out.

III. IDENTIFICATION OF DATA THEFT IN SOCIAL MEDIA AND MEDICAL EHR RECORDS

In the age of digitization, where vast amounts of personal and sensitive data are stored electronically, the risks of data theft and breaches have become increasingly prominent. Two significant domains where data theft is of particular concern are social media and medical Electronic Health Records (EHR) systems. In this article, we delve into the challenges and strategies for identifying data theft in these critical areas.

A. Social Media: A Breeding Ground for Data Theft

Social media platforms offer opportunities for connection and information sharing, they also present a fertile ground for data thieves. Users willingly share a wealth of personal information, ranging from their locations and preferences to family details and even financial data. Malicious actors exploit this openness to steal, misuse, or sell users' data for various purposes, including identity theft and cyberattacks.

B. Identifying Data Theft in Social Media:

- User Activity Monitoring: Monitoring user activities and access patterns can help identify unusual behavior that might indicate data theft. For instance, if a user's account shows a sudden increase in login attempts or access from unfamiliar locations, it could be a sign of unauthorized access.
- Content Analysis: Advanced content analysis tools can detect abnormal posting behavior or the presence of suspicious links and malware in posts and messages. These tools can also identify instances of account hijacking and impersonation.
- User Anomaly Detection: Anomaly detection algorithms can help identify unusual user behavior by analyzing historical data. Sudden changes in posting frequency, content, or connections can raise red flags.
- Medical EHR Records: A Goldmine for Data Thieves: Electronic Health Records (EHR) contain a patient's comprehensive medical history, diagnoses, treatments, and personal information. The highly sensitive nature of this data makes medical EHR systems an attractive target for data theft. Unauthorized access or leakage of EHR data can lead to identity data theft, fraudulent claim of insurance, or even threatening of life consequences for patients.

C. Identifying Data Theft in Medical EHR Records:

- Access Logs and Auditing: Regularly auditing access logs and monitoring who accesses EHR records can help identify unauthorized access. Any suspicious activity should trigger immediate investigation.
- Data Encryption: Implementing strong encryption measures for EHR data can deter data theft. Encryption refers that, if data is being stolen, then it remains unreadable without decrypting the data.

- Behavioral Analysis: Employ behavioral analysis tools that monitor user behavior within EHR systems. Deviations from established patterns can indicate potential data theft or misuse.

In conclusion, data theft in social media and medical EHR records poses significant threats to individuals and organizations. To address this issue, proactive monitoring, robust access control, encryption, and user authentication are crucial components of an effective data theft identification strategy. By staying vigilant and employing these strategies, both social media platforms and healthcare organizations can take steps to safeguard the protection of data and the privacy of the data entrusted n the userend.

IV. BACKGROUND

A. Differential Privacy:

Data mining algorithms have been suggested to operate over databases containing sensitive data using differential privacy [4], a robust privacy guarantee. Individuals' data can be safely taken into a database, and the privacy leakage ratio is determined by a ρ parameter [4, 5, 6]. According to the theory of differential privacy, a function's output is not totally dependent on any database instance. It asserts that the likelihood of getting the same result is high.

Case 1: Adjacent database Two neighbour databases, D and D', are different from one another by a one instance.

Case 2:Distinctive privacy (ρ-class) Similar to the Laplace mechanism, a randomly generated method A is ρdifferentially secure if each subset S of the outputs of algorithm A for every nearby database D' and D satisfy the following condition: Furthermore, S ⊆ Range (A). where reflects the range of outcomes that can be produced by the randomised process A and is used to ascertain the degree to which a malevolent client can discriminate among the record sets D' and D. The chance that $A(D)$ will be an element of S is given by Pr[$A(D) \in S$], and the results of the randomization technique A for the records are $A(D')$ and $F(D)$. SENSORY, DEFINED - 3 Assume that f(x) is a function that transforms x from a database into real numbers.

The greatest extent to which, in the worst scenario, the record of a single person may change the value of a function f is known as its sensitivity. Stated differently, a function's sensitivity gives rise to a maximum restriction on the amount of perturbation its output must undergo in order to offer differential privacy [4, 5, 6]. Laplace mechanism

Case 3: The randomized procedure A represents a Laplace process that generates the function and provides the following answer for $f(D)$: If and only if $A \geq \Delta f/\epsilon$, then algorithm A is ϵ-differentially private. When V is an independent, uniformly distributed random factor chosen from $LaP(\gamma)$, we get $A(f(D))$ => $f(D) + V$ (5).

Moreover, each predictor for all of the classes. Writing a count query as a function requires consideration of the differential privacy ideas discussed above. The differentially confidential outcome is δ plus b if the actual query result is δ. The Laplace with means 0 and deviation by standard deviation Δb ϵ is used to extract b in this case. Here, the degree of sensitivity of the count query is represented by Δf, and the privacy parameter, lower values of which signal significantly greater privacy.

Differential privacy is a mathematical framework developed to protect individuals' privacy when analyzing and sharing statistical information about data collections. The core idea behind differential privacy is to introduce a certain amount of randomness (or noise) to the results of queries on databases containing sensitive information, such that the one record doesn't significantly affect the outcome, hence ensuring an individual's privacy.

Key Concepts:

1. ε (Epsilon): Represents the privacy budget. A smaller ε indicates better privacy but might reduce data utility, while a larger ε compromises privacy but improves data accuracy.
2. Laplacian Noise: Commonly used method to introduce randomness in differential privacy.
3. Sensitivity: Measures the maximum amount a function's output can change by modifying one record in the input database.

Applications of differential privacy range from census data collection to machine learning model training. The aim is to allow organizations to share useful insights and data statistics without compromising individual privacy.

B. Rise of Big Data and Privacy Concerns

1.Explosive Growth of Data:
The advent of big data is characterized increase in the volume, the velocity, and variety by individuals, organizations, and devices.

2.Data Utilization Challenges:
The immense potential of big data for insights and innovation creates challenges in finding a balance between maximizing data utility for analysis and protecting individual privacy.

3.Re-Identification Risks:
Big data analytics increases the risk of re-identification, where seemingly anonymized data can be combined with other datasets to identify individuals, posing a significant threat to privacy.

4.Granular and Personalized Insights:
The granularity of big data allows for highly personalized insights into individual behaviors, preferences, and activities, making it imperative to establish robust privacy safeguards.

5.Ubiquity of Data Collection:
Pervasive data collection through online platforms, IoT devices, and sensors results in a constant stream of personal information, heightening the need for comprehensive privacy protection.

6.Privacy Concerns in Social Media:
Social media platforms contribute significantly to big data, raising concerns about the extensive profiling of individuals, potential misuse of personal information, and the erosion of privacy boundaries.

7.Challenges in Anonymization:
Traditional anonymization methods may prove insufficient in the face of big data analytics, as sophisticated algorithms can potentially reverse engineer anonymized datasets.

8.Emergence of Algorithmic Biases:
The use of bigdata and ML algorithms introduces risk in algorithmic biases, amplifying concerns about the unfair and discriminatory treatment of certain groups.

9.Need for Privacy by Design:
The integration of privacy considerations at the design stage of data systems becomes crucial to proactively address privacy concerns and mitigate risks associated with big data.

10.Regulatory Responses:
Governments and regulatory bodies are responding to the challenges posed by big data with the introduction of stringent privacy regulations, emphasizing the importance of individual rights and data protection.

11.Ethical Data Handling:
Organizations are increasingly recognizing the ethical responsibility of handling big data and are adopting ethical frameworks to guide the collection, storage, and use of personal information.

12.Public Awareness and Concerns:
The rise of big data has sparked increased public awareness regarding the potential privacy risks, leading to a demand for transparency, accountability, and stronger privacy measures from businesses and institutions.

C. Reidentification Attacks with Differential Privacy

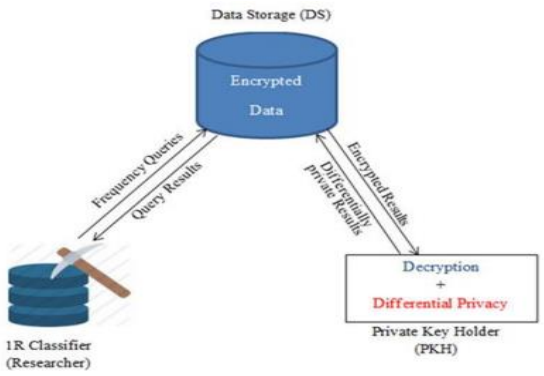

Fig 1.Scheme of Proposed privacy model

PROJECT DESCRIPTION:

In this part, we will see about the project domains in a detailed manner.

A. METHODOLOGY:

Three datasets were taken from the Kaggle which are Public medical dataset, which contains k- anonymized demographic and some medical sensitive information, second dataset is an attacker's data collection with basic information of demographics and the third dataset is a public medical dataset which are preprocessed for the MWEM synthesizer. The anonymized and attacker's dataset includes the Gender, Age, Zip which are the personal information of patients in a base manner which can be further synthesized.In addition, the aforementioned data sets have a unique id for every record, which is utilised to calculate the amount of recognised records following the attack. The attack is not carried out using this information.

Re-identification Attack using Differential Privacy:

Sensitive personal data can be shielded from re-identification threats with differential privacy. If an attacker can link anonymized records of individuals from a public dataset with information about these individuals from many sources, then the identities of those persons may become public knowledge. Patient records are included in the public anonymised dataset used in this demo. The attacker uses basic demographic data, such as zip codes and age, to try to identify specific people. We demonstrate that even when the sensitive material is made public in an anonymised manner, effective reidentification attacks are still feasible. After safeguarding the sensitive data with a dataset synthesiser from the SmartNoise system, we launch a second attack.

- Importing a medical data set that has been anonymised and gathering data from the attacker
- Authentication Attack I: Using the anonymized data set to reveal identities
- Protecting the medical dataset via the multiple weights exponential mechanism (MWEM) and differential privacy
- Verifying the combined data set's suitability for statistical analysis
- Reidentification Attack II: Pursuing identity disclosure based on the medical data set's uniquely private version

	Gender	Age	Zip	Diagnosis	Treatment	Outcome
18865	M	50-59	212**	Arthritis	45	unchanged
21789	F	60-69	348**	Depression	50	unchanged
27460	F	40-49	355**	Cancer	45	recovered
29618	M	40-49	596**	Alzheimer	34	intensive care
11476	M	80-89	151**	Osteoporosis	49	unchanged
17750	M	60-69	143**	Stroke	34	recovered
20882	M	10-19	189**	COPD	40	intensive care
7375	M	30-39	555**	Arthritis	31	intensive care

Fig 2. Medical Dataset(Re-identification Attack I)

The above are the Anonymized dataset with sensitive medical information

	Name	Gender	Age	Zip
3900	Keith Rodriguez	M	71	11112
12425	Stacy Roberson	F	72	84085
2799	Katherine Baker	F	26	72910
13200	Linda Clayton	F	18	14439
20610	Charles Delgado	M	19	91140
26901	Michael Hall	M	12	8164
7877	Brenda Franco	F	30	82518
7754	Paul Morales	M	11	63792

Fig 3. Demographic Dataset(Re-identification Attack I)

Attacker`s data collection with basic demographic information

Sample of re-identified patients:

	Name	Gender	Age	Zip	Diagnosis	Treatment	Outcome	ID_Match
18104	Lacey Harper	F	72	33521	Diabetes	34	unchanged	True
17784	Kristen Hill	F	26	60720	Alzheimer	45	recovered	True
21726	Samantha Sutton	F	44	90645	Depression	30	unchanged	True
26295	Linda Wade	F	35	51198	Depression	40	recovered	True
6473	Ethan Miller	M	34	23345	High Blood Pressure	47	intensive care	True
22658	James Vargas	M	86	13661	Arthritis	43	unchanged	True
12894	John Ramos	M	78	71113	COPD	39	intensive care	True
8138	Jose Beck	M	33	70481	High Blood Pressure	44	recovered	True
23030	Marvin Nelson	M	25	84199	Cancer	21	recovered	True
21937	Thomas Jones	M	39	93549	COPD	38	unchanged	True

Fig 4. Sample of reidentified Patients

Encoding of data

	Gender_encoded	Age_encoded	Zip_encoded	Diagnosis_encoded
0	0	10	65418	9
1	0	14	65475	2
2	0	10	65484	9
3	0	30	27727	7
4	0	36	27772	3

Fig 5.Encoding of data

Reidentification Attack II

Medical Dataset:

	ID	Gender	Age	Zip	Diagnosis	Treatment	Outcome
4222	d07f281fdce643918fe6408d7d8eea19	M	10	26876	Osteoporosis	26	recovered
26109	d21e74fcde3b4ebba7893c7815b31118	F	44	35937	Diabetes	33	recovered
21055	234cbf3c006845b8880293ad9924c84f	M	53	20246	Alzheimer	41	recovered
29054	0f2c3685b8e54ce4b77f2f15bf85a4ee	M	65	69421	Alzheimer	37	unchanged
8186	db30880c0da94130a197dd5c821af4e0	M	31	43137	High Blood Pressure	30	recovered

Fig 6.Medical Dataset(Re-identification Attack II)

Synthesized Demographic Dataset:

	Gender_encoded	Age_encoded	Zip_encoded	Diagnosis_encoded
14884	1	11	15433	9
6599	0	65	23989	9
2564	1	47	45478	9
21303	0	87	23591	5
972	1	62	83284	7

Fig 7. Synthesized Demographic Dataset(Re-identification Attack II)

Smartnoise:

A collection of technologies called SmartNoise is used to create dashboards, synthetic data releases, reports, and synopses that are distinctly private. It consists of a set of synthesisers and a SQL processing
layer that supports queries across Spark and well-known database engines.

z3 solver:
Microsoft Research's Z3 theorem prover can handle bitvectors, booleans, arrays, floating point integers, texts, and other data types.

Matplotlib:
A complete Python visualisation toolkit for static, animated, and interactive graphics is called Matplotlib. Matplotlib enables both difficult and easy tasks.

MWEM Synthesizer:
MWEM Synthesizer combines the methods of exponential mechanisms and multiplicative weights to provide differential privacy. It's a quite straightforward yet successful strategy. It requires shorter duration and less computational resources are needed.

V. RESULT

After importing the three datasets, Performing the re-indentification attack using the medical and demographic dataset. The results of the attack of the amount of potential matches and actual matches are found to be 9045 actual validated matches.

As a next step, To boost the level of security, we will synthesise the dataset in the following phase. For this, the Multiple Weights Exponential Mechanism (MWEM) synthesiser will be used to encode the diagnostic and
the demographic information.the analysis does not include the additional variables. Thereby synthesizing the

demographic data and relating the actual and synthesized part of data in the form of plotting up a histogram.

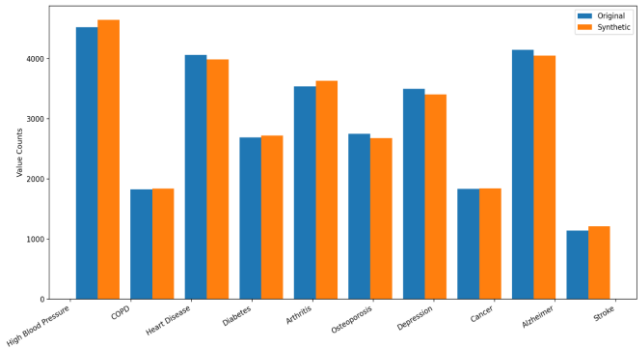

Fig 8. Comparing Original and Synthesized Data

Comparing using Metric:

For Synthesized Data:

```
Silhouette Score: 0.6384068715526127
```

For Original Data:

```
Silhouette Score: 0.6656184702063361
```

Cosine Similarity:

```
Mean Cosine Similarity: 0.9980150005494549
```

Comparison between the Actual and Synthesized data:

We'll use create_histogram function to show the distribution of diagnoses for both data sets below. The bars should ideally not deviate significantly from one another for any given diagnosis. Less information is lost during the synthetization process the more comparable the bars are for the corresponding disease.

Lastly, we use the try_reidentification_noise-function to attempt the re-identification attack on the synthesised data. We no longer deal with the raw/real data since, as previously said, the synthesised data set contains novel arrangements of demographic data. It is doubtful that we are dealing with a real match here, even though
it is possible that a prospective match is found.

After performing the re-indentification final attack agin with a combinational approach, the amount of actual and potential matches results in zero potential matches thereby preserving the data.

Below, we show the amount of potential and actual matches and provide a glance at the data.

```
Found 0 potential matches!
```

ID	Gender	Age	Zip	Diagnosis	Treatment	Outcome

Fig 9. Potential and actual matches

VI. REFERENCES

[1] Homomorphic encryption security standard: Albrecht, M., et al. November 2018 technical report, HomomorphicEncryption.org, Toronto, Canada.

[2] Using the Rényi divergence instead of the statistical distance, Bai, S., Lepoint, T., Roux-Langlois, A., Sakzad, A., Stehlé, D., and Steinfeld, R. have improved security proofs in lattice-based cryptography. 2018; J. Cryptol. 31(2), 610–640

[3] Canonne, C.L.: A distribution testing survey: You have a lot of data. Is it blue, though? Theory of Computing (2020), 1–100 pp.

[4] Canonne, C.L., Steinke, T., and Kamath, G.: The discrete gaussian for differential privacy. Vol. 33, pp. 15676–15688, in Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.). Advances in Neural Information Processing Systems. In 2020, Curran Associates Inc.

[5] Cheon, J.H., and colleagues: Using real-time homomorphic authenticated encryption to fly drones in a secure manner.

[6] Cheon, J.H.; Han, K.; Kim, A.; Kim, M.; Song, Y.: Approximate homomorphic encryption via bootstrapping. LNCS, vol. 10820, pp. 360–384, in Nielsen, J.B., and Rijmen, V. (eds.), EUROCRYPT 2018, Part I. Cham Springer (2018).

[7] Cheon, J.H., Han, K., Kim, A., Kim, M., Song, Y.: An approximate homomorphic encryption variant using full RNS. In: Jacobson Jr., M. and Cid, C. (eds.) LNCS, vol. 11349, pp. 347–368; SAC 2018. Cham Springer (2018).

[8] Cheon, J.H.; Kim, A.; Kim, M.; Song, Y.: Homomorphic encryption for approximate number arithmetic. LNCS, vol. 10624, pp. 409–437; in: Takagi, T., Peyrin, T. (eds.) ASIACRYPT 2017, Part I. Cham Springer (2017).

[9] L. Fan, 2018. Pixelization of images with varying levels of privacy. In 32nd Annual IFIP WG on Data and Applications Security and Privacy XXXII

[10] Dankar & El Emam, 2013. Dankar, F.K. A review of differential privacy practices in healthcare. 35–67 in Trans. Data Priv., 6(1).

[11] Gursoy, M.E., and Z.S. Kaya, 2023. Regarding Local Differential Privacy-Based Solutions and Re-Identification Attacks' Efficacy for Smart Meter Data.

[12] Towards assessing re-identification risks in the local privacy model, Murakami, T., and Takahashi, K., 2020. preprint arXiv:2010.08238 at arXiv.

[13] In July 2017, Zaman, A.N.K., Obimbo, C., and Dara, R.A. An enhanced algorithm for differential privacy to safeguard data reidentification. International Humanitarian Technology Conference (IHTC), IEEE Canada, 2017, pp. 133–138. IEEE.

[14] Gill, N.S., Gulia, P., and Ratra, R. (2022). Assessment of Re-identification risk in healthcare through differential privacy and anonymization. International Journal of Advanced Applications in Computer Science, 13(2).

[15] Gill, N.S., Gulia, P., and Ratra, R. (2022). Assessment of Re-identification risk in healthcare through differential privacy and anonymization. International Journal of Advanced Applications in Computer Science, 13(2).

[16] October 2007, McSherry, F. and Talwar, K. Differential privacy in mechanism design. Found on pages 94–103 of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07). IEEE.

[17] Dwork, C. (April 2008). Differential privacy: A survey with findings. In International Conference on Computational Models: Theory and Applications (pp. 1-19). Springer Berlin Heidelberg, Berlin, Germany.

[18] In 2021, Nikam, R.R. and Shahapurkar, R. Techniques for Protecting Sensitive Data Privacy and Ensuring Security in Big Data.

[19] In 2022, Chatterjee, J.M., Gill, N.S., Gulia, P., and Ratra, R. Big Data Privacy Preservation Using Principal Component Analysis and Random Projection in Healthcare. Engineering Mathematical Problems, 2022.

[20] In 2014, Gosain, A. and Chugh, N. protecting privacy in large data sets. Computer Applications International, 100(17).