Grenze International Journal of Engineering and Technology, January Issue

AI-Driven Approaches to Data Anonymization: Techniques, Challenges, and Legal Perspectives

Dhana Lakshmi M P[1] and Bharath Singh Jebaraj[2]
[1]Research Scholar, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Srivilliputhur, India
Email: dhanalakshmimp@gmail.com
[2]Associate Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Srivilliputhur, India
Email: j.bharathsingh@gmail.com

***Abstract*— Anonymity of data is necessary to safeguard sensitive information while also promoting the use of analytics for data-driven exploration, intelligent decision-making, and sequestration regulations. This article extensively examines the latest anonymization techniques, such as data masking, pseudonymization and discrimination sequestration, and argues that they can safeguard data while still maintaining privacy. Additionally, this paper highlights how these methods enable both seclusion and data mileage protection. It also covers the legal materials that require anonymization, such as GDPR and HIPAA, and the practical obstacles that associations face in establishing robust sequestration-protection strategies. Additionally, this research examines the practical implementation of techniques across fields such as AI, healthcare and finance, revealing emerging patterns in data sequestration tools and fashionable practices that define the evolving data capture landscape. The outcomes signify the expansion of AI-driven anonymization strategies in minimizing identification glitches and safeguarding sensitive data. The paper concludes by providing guidance on improving anonymization methods, which would enable associations and experimenters to maintain sequestration while maximizing the benefits of participating datasets.**

***Index Terms*— Data Anonymization, artificial intelligence, data privacy, risk, privacy-preserving data publishing.**

I. INTRODUCTION

Humans have perfected the art and science of the integration of data into daily lives. There is an access to data in amounts to a value that lies beyond the boundaries of human conception, and at the same time, they are well-occupied with the tools and techniques to manipulate the data in whichever way is required. Ubiquitous computing, higher bandwidth, and extended connectivity have facilitated a form of data collection and processing that is hitherto undreamt of.

Data owners like banks, hospitals, insurance companies, social networks, and other tech giants collect huge amounts of digital data from their users for a variety of purposes. Information on the users' activities, demographics, finances, interests, preferences, location, political and religious beliefs, online communities, medical information, and opinions is frequently included in this data collection.

Grenze ID: 01.GIJET.12.1.223

Organizations, ventures, and indeed the logical community are exceptionally much inquisitive about such datasets but are incapable to lay their hands on these since security directions avoid sharing of such delicate information with third-party organizations. This shapes the requirement for an arrangement that makes a difference to share such datasets without compromising the user's security. The solution must also be able to strike the perfect balance where the privacy of users is not compromised still preserving the utility of data. One of the methods to accomplish this is by data anonymization.

A suite of models, tools, and approaches meant to reduce privacy risks connected with data releases by data miners or analysts, Privacy-Preserving Data Publishing (PPDP) offers to solve this issue and protect data against the rising global privacy violations, efficient anonymizing techniques are becoming more and more needed.

Data anonymization is the process of changing personal data such that, whether working with another party or the data controller alone, a data subject cannot be either directly or indirectly identified. Protecting user privacy, stopping inadvertent disclosures, and shielding against adversarial attacks is its main objectives. International data protection and privacy regulations, such as "EU-GDPR, HIPAA, and PCI-DSS," mandate data anonymization, influencing Indian regulations like the "Personal Data Protection Bill, 2019" (also termed as the Data Protection Bill, 2021) and the DISHA Act for healthcare data. Under GDPR, businesses can collect, use, and retain anonymized data without requiring user consent, as long as all identifiers are removed.

However, the absence of a universally accepted definition of a fully anonymized dataset, along with the challenge of developing robust anonymization algorithms that maintain data utility while ensuring privacy, presents a significant challenge for the technical community.

Anonymization, according to ISO/TS 25237:2008, is the process of severing the link between a data subject and an identifiable dataset. Often known as de-identification, it includes deleting, hiding, redacting, or delinking all personally identifiable data from an individual's digital health records. This indicates that the owner's identity cannot be unintentionally exposed while yet allowing, if needed, the data to be re-linked to the owner. [2].

Anonymization is achieved by carefully balancing privacy protection and minimizing information loss. The dataset must retain its utility while ensuring the removal of all PII and other details that could lead to privacy violations. However, the risk of re-identification remains a challenge. A properly anonymized dataset should prevent adversaries from de- identifying individuals. Privacy-preserving data mining and publishing methods have attracted significant research interest, resulting in numerous studies on the topic [3].

Personally Identifiable Information (PII) refers to any data used to locate, contact, or identify an individual uniquely from others, or that can be combined with other information to achieve such identification [2]. A more detailed list of PIIs is provided in the appendix.

De-anonymization is the process of reversing anonymization by cross-referencing de-identified data with other sources to re-identify individuals. This can occur if the de-identification process was never initiated or was not implemented effectively, leaving gaps that allow re-identification. According to the information provided, a dataset consists of four different kinds of attributes from a privacy perspective:

- Identifiers: These include specific details, such as name and Aadhar number, which directly identify the owners of the records
- Quasi-identifiers, or QIDs, may be able to identify the proprietors of records (e.g., age, postal code).
- Sensitive attributes: incorporate information about a specific individual that is sensitive (e.g., salary, sickness).
- Non-sensitive attributes: any characteristics that do not fall under one of the three aforementioned categories are included in this group.

One of the main strategies to solve these problems is data anonymizing, which transforms data into a format that preserves personal identities while yet allowing its use for artificial intelligence and machine learning applications. Consider the healthcare sector, which is progressively using AI-driven diagnostic tools. Using these tools calls for a lot of patient information training. Still, there are privacy issues along the way toward using artificial intelligence for improved patient outcomes. One main challenge that might impede progress is the possibility of revealing private patient data. Data anonymization is therefore quite important since it converts private information into a sanitized form devoid of identifiable information while maintaining its natural value. By means of training AI models, this conversion helps the healthcare industry to progress without violating people's privacy.

The use of anonymized data is a crucial impediment to AI/ML model training, safeguarding individual identities and facilitating the learning and improvement of models. Although the identities are not revealed, the AI/ML algorithms can still extract valuable and important information from the anonymized data.

Nevertheless, the technical aspects of analysing data in AI/ML are not clear-cut. By employing anonymization techniques like k-anonymity, l-diversity, and t-closeness, AI/ML models can efficiently utilize relevant data

without jeopardizing privacy. Every method presents an alternative means to secure data without decreasing its utility of it, which is essential for implementing AI/ML applications.
It is a high-stakes game for the financial services sector where fraud detection powers heavily rely on AI and ML technologies. This is a two-fold concern: to protect consumers from fraud while also safeguarding their privacy. The use of data anonymization makes it possible to construct fraud detection models using datasets that conceal sensitive information but still expose fraudulent patterns. Social media platforms utilize AI and machine learning techniques to scrutinize the vast amounts of user-generated content and filter out harmful content. This is a similar situation. While user-generated data is a valuable source of information, it also raises privacy concerns.
Anonymity in data provides a means of safeguarding privacy, enabling AI/ML models to learn and analyse content without any risk. The role of data anonymization in promoting ethical behaviour in AI/ML is becoming more apparent as we explore the different fields where AI has a significant impact. The story of progress in AI/ML is closely tied to the narrative of privacy protection, with data anonymization being the key factor.

II. Literature Survey

Data privacy is a significant issue in the face of big data and widespread digital transactions. Organizations collect vast amounts of data, making anonymization techniques essential for safeguarding sensitive information while preserving data utility. These techniques aim to protect personally identifiable information (PII) while enabling meaningful analysis. The early 2000s saw the widespread adoption of k-anonymity, a technique that guarantees that every person included in n data is identical to at least some point between those others (k-1). The limitations of k- anonymity included the avoidance of repeated identification and disclosure of attributes. The use of l-diversity and t- closeness was one of the more advanced methods used to tackle these problems. Afterward, the implementation of Generative Adversarial Networks (GAN)-based techniques and differential privacy further enhanced privacy protection by offering more robust protection against data breaches.
To standardize and regulate the collection, storage capacity, transmission, authorization, dissemination of, and utilization of digital health data, the Act [2] establishes National and State eHealth Authorities and Health Information Exchanges. The primary objective is to maintain the trustworthiness, privacy, and security of such information while dealing with related and incidental issues. Furthermore, the Act specifies particular conditions for data anonymization and furnishes a generalized record of Personally Identifiable Information (PII)). The work done by Abdul Majeed and Sungchang Lee [3] has served to be the baseline of this study. Several concepts and ideas from the aforementioned study were incorporated into this work. While their research addressed both relational and structural datasets, this project focused solely on relational data structures. The authors conducted an extensive literature review, referencing over 300 papers and distilling key insights into their study. In addition to the differential privacy model, different types of statistical privacy models, including k-anonymity, l-diversity for proximity, and t- closeness, were examined. The need to share and even publicly release health information is increasing. However, such disclosures raise significant privacy concerns. To address these issues, data can be anonymized before dissemination. The study [4] evaluates the re-identification risks associated with k-anonymization, assessing its effectiveness in protecting sensitive health data. This work [5] focuses on the GDPR requirement of data anonymization and how it differs from pseudo- anonymization. The work studies the naive anonymization operations like suppression, substitution, shuffling (permutation), noise addition, and generalization while specifying the risk associated with each. It also looks into various anonymization tools like ARX, μ-Argus, SD Micro, and Privacy Analytics Eclipse.
In this study [6], a test is evaluated using a real-world health dataset provided by INEBIR's Test-Driven Anonymization (TDA) method. TDA conducts a systematic assessment and utilization of anonymization techniques to attain an ideal equilibrium between non-functional privacy quality (NDQ) and functional data usability (QCM). The findings indicate that k- anonymity with a k value of 16 provides the best results for this dataset, ensuring anonymity while preserving sufficient data utility to prevent significant errors in AI model training. Their research [7] covered a range of advancements and algorithms suggested in the realm of PPDP. They examined the strengths and weaknesses of 11 distinct algorithms proposed by different authors over the past decade, despite the fact that the main objective is to reformulate the initial data into a form that prevents the record owner's confidential data from being deduced. Most of the proposed algorithms provided security only until the initial data release or for the first recipient, as they were designed with a single data publication in mind. In the paper [8], they also discuss the k-anonymity model as a means of protecting privacy by placing an emphasis on the data owner's duty to ensure that the information shared with external parties cannot be redistributed. Re-identification attacks are also examined on the released data adhering to anonymity. Methods of mitigation discussed are used to prevent the data release from being re- identified. One possible method is by

altering the data and linking it to maximum data so that pointing to single data is completely impossible. The work by L. Sweeny [9] is regarded as one of the core baseline and a fundamental concept on the field of data anonymization. Out of a population of 248 million Americans, 87% reported having characteristics that could be used to differentiate them based solely on their ZIP/PIN/Area code, gender, and date of birth.

The paper [10] explores key privacy models, namely l- diversity and t-closeness:

- l-diversity: Protects privacy by ensuring that sensitive attributes in each equivalence class are represented in a heterogeneous manner, thereby decreasing the likelihood of inference attacks.
- t-Closeness: Strives to keep the distribution of sensitive data in the anonymized data roughly equivalent to that in its original dataset, while preventing the disclosure of such attributes.

This work in the paper [11] describes the de- anonymization of the famous Netflix Prize dataset by MIT researchers, which was then believed to be an anonymized dataset published by Netflix. They successfully attacked the Netflix Prize dataset through membership by comparing it to publicly available Amazon reviews. This was possible despite the removal of personally identifiable information and minor modifications to the data, highlighting vulnerabilities in anonymization techniques. The document [12] comprehensively overviews various anonymization methods, such as randomization, noise addition, permutation, differential privacy, generalizations, k- anonymity, and l-diversity. Additionally, the anonymous anonymizer is not restricted to specific categories. It also examines the guarantees, limitations, and potential shortcomings of each approach. By replacing individual data values with broad categories, generalization enhances privacy protection by safeguarding the data's utility. The paper [13] reviews different anonymization techniques, with a section dedicated to the generalization method. It emphasizes the need for hierarchical generalization in achieving strong privacy guarantees without sacrificing data utility.

The paper [14] focuses on the application of generalization in privacy-preserving data mining, analysing different generalization methods and their integration with other anonymization techniques. This article [15] investigates the role of suppression in privacy-preserving data mining. The authors explain how suppression works by removing sensitive attributes or values from datasets before they are mined for patterns or insights. This method is most commonly used when other anonymization techniques, such as generalization, may not provide sufficient privacy. The paper also discusses challenges related to selecting which data to suppress and how much data should be suppressed to achieve optimal privacy protection without overly compromising data utility. The article [16] explores ongoing challenges and future directions in suppression- based anonymization, highlighting the difficulty of balancing data privacy with utility. It introduces adaptive suppression techniques that adjust the level of suppression based on data characteristics, analysis needs, and privacy risks. Additionally, the authors examine machine learning-based methods to enhance suppression decisions dynamically, aiming to minimize data loss while ensuring robust privacy protection. The paper [17] examines the weaknesses of k-anonymity, particularly its susceptibility to homogeneity attacks, which allow attackers to infer sensitive information from anonymized data. In this work, to prevent the inference of sensitive information from anonymized data, l-diversity is considered to be the privacy model as there is an assurance that all record types with identical characteristics possess at least n distinct values for the sensitive attribute.

In order to overcome the limitations of k-anonymity and l-diversity, the paper [18] presents a new privacy model that ensures the distribution of sensitive attributes within each equivalence class is as close to the overall dataset distribution as possible. The proximity is measured using a distance metric and its closeness is determined by setting 't. This results in adversaries being unable to deduce much, or even, about an individual's sensitive nature from the class of equivalence. l-diversity is presented in [19] as an extension to the k-Anonymity privacy model. While k-anonymity ensures that every person in a dataset is at least identical to mk, it remains susceptible to attacks on homogeneity and background knowledge. To enhance privacy protection, l- diversity is necessary for each equivalence class to have at least a ll well- represented value for sensitive attributes. t- Closeness, a privacy model that improves k- Anonymity and l-diversity by decreasing the risk of sensitive data inference in large datasets, is explored in the paper [20]. However, it is also noted that the limitations of k-anonymity (which ensures only indistinguishability) and l-diversity (which may still permit some attribute disclosure), make for a more likely candidate t-closeness. The authors [21] proposed a method that protects sensitive QIDs by implementing two key algorithms. Anonymization Algorithm: It randomizes sensitive QID values by adding original and randomly selected values to create diverse sets for each record, ensuring that sensitive information is not easily inferred. Reconstruction Algorithm: This algorithm reduces the reconstruction error between the anonymized and original values, making it difficult for data analysers to accurately estimate sensitive QID values, even if they know most of the data.

In the paper [22], key privacy is demonstrated through the utilization of a differential privacy mechanism. It also discusses the way in which noise is added to the output of queries, depending on whether a function's sensitivity

and some privacy settings (Ɛ) dictate that each individual data point has little impact on the results. The privacy mechanism protects individual data even if the adversary has complete knowledge of other database entries, maintaining privacy regardless of external information. The mechanism also extends to groups, ensuring that the collective data of multiple participants does not leak information, bounding the probability of data leakage by a factor related to the group size. The paper [23] delves into the methodology of designing and testing algorithms that disclose information while maintaining the confidentiality of individuals in the dataset. Additionally, This method employs the k-anonymity model, which means that all quasi-identifiers in the dataset are identical to at least a few other records. Additionally, the article explores efficient algorithms for generating l-diverse tables, leveraging the property of monotonicity to enable practical data publishing while preserving privacy. Enhancing privacy while preserving data utility is the paper's goal [24]. One of the arguments for improving data privacy is the t-closeness principle, which seeks to limit observer learning by measuring the amount of individual-specific information available. Improved data privacy is achieved through t-closeness, which means that the distribution of sensitive attributes within equivalence classes closely mirrors and limits the possible information gain for an observer. It reduces the risk of disclosure by certain attributes and provides a better data utility than earlier privacy protections, including k-anonymity and l-diversity. In 2006, Cynthia Dwork and her colleagues introduced the concept of privacy guarantees in their influential paper "Differential Privacy," which established a mathematical model for quantifying privacy protections during data analysis. By incorporating or eliminating data from an individual's dataset, the outputs of this method are considered completely unidentifiable, rendering any significant information about them unavailable. GANs can be designed to use a generator-discriminator architecture, which results in data that closely mirrors real datasets. This provides an opportunity to utilize EHRs for research and analysis while maintaining privacy. Techniques integrated into the GAN framework for privacy preservation include the addition of noise and the inclusion of secure multiparty computation or homomorphic encryption [26].

TABLE I. COMPARISON OF DATA ANONYMIZATION TECHNIQUES

Sl No	Reference Paper	Results
1.	Anonymization Techniques for Privacy-Preserving Data Publishing: A Comprehensive Survey	Classifications of attribute types, privacy threats, and common anonymization operations. Various anonymization and deanonymization techniques are discussed. Anonymization is considered to be important to avoid any illicit incidents.
2.	Privacy- Preserving Data Publishing and Data Anonymization Approaches: A Review	It discusses the importance of data protection and how data owners ensure data security. Managing data privacy and maintaining its confidentiality is still a formidable undertaking.
3.	Test-Driven Anonymization in Health Data: A Case Study on Assistive Reproduction	The goal of data anonymization is to enhance privacy and minimize the risk of sensitive information being disclosed without warning. This involves manipulating the data.
4.	k-anonymity: A model for protecting privacy	It uses the k-anonymity model for data protection. Privacy should be the focus when deciding on the value of k.
5.	l-Diversity: Privacy Beyond k- Anonymity	The k-anonymity model, which is the least robust in sensitive attributes due to its lack of diversity, makes it easier for adversaries to deduce private information more easily than the l-diversity approach. This distinction helps counteract this flaw. The working of the anonymity theory is discussed.
6.	t-Closeness: Privacy Beyond k-Anonymity and l- Diversity	A new model named "t-closeness" is defined, which uses sensitive attributes in a relational model.
7.	Anonymization of Sensitive Quasi-Identifiers for l- Diversity and t- Closeness	Two algorithms are considered for effective data anonymization that can be used by data holders and data analyzers.
8.	Differential Privacy: A Survey of Results	By adding noise, data anonymization is enhanced in contrast to conventional anonymizing mechanisms like k- anonymity, l- diversity and t-closeness, which offer better privacy protection but lower the risk of being re-identified.
9.	Simple Demographics Often Identify People Uniquely	The set of attributes that could be considered as identifiers that point to individuals or the ones that do not point to individuals are discussed. The article demonstrates the dataset and the attributes that are used to identify the people uniquely.
10.	Who's Watching? De- anonymization of Netflix Reviews using Amazon Reviews	Removing personally identifiable information itself cannot completely anonymize the data. Data holders believe that removing PII from the dataset is enough and the dataset can be distributed for analysis purposes. But removal of PII alone cannot completely anonymize the data.

III. DATA ANONYMIZATION

Data anonymization as a topic is quite broad in the sense that it contains several problem statements in itself. Some of the major issues are described below:

- Data Protection (PPDP) techniques aim to maintain user privacy by publicly releasing information while maintaining the confidentiality of the data. Not all applications need the entire dataset to be published or shared with. Often, the requirement is just for some visualizations or insights made from the underlying dataset to understand the general trend. These visualizations and insights must also not violate the privacy of the users.
- News of data breaches and data theft from time to time is being heard everywhere. The leaked data could potentially harm the users' privacy if the attackers are capable of mapping the users' identity to the sensitive data property values. When stored and used, the problem of privacy being violated in the event of a data leak is eliminated.
- Data protection regulations such as the EU-GDPR, Personal Data Protection Bill (2019), and Digital Information Security Act (DISHA Act) (2018), have been enacted and others aim to protect the identity of the user and uphold the privacy of the individuals. Many of these regulations require companies to anonymize their data so that even in case of a data leak, the private information of the users will not be exposed. Also, there are guidelines preventing the datasets from being shared with third parties without proper consent from the users, and also it must ensure that the original content of the data shall not be revealed even if the researchers use any method for analysis. Many of the advertising and business analytics companies and also the research community are very much interested in these datasets, but these privacy regulations prevent sharing unanonymized data.
- There arises a need for a proper data anonymization algorithm that can adhere to the privacy regulations by protecting the privacy of the users without actually degrading the utility of the dataset. As we turn the knob to increase data privacy, this would in turn decrease the utility of the data available. Similarly, too much anonymization may end up generating a very generalized and random dataset, which is of no use to the consumer.
- As the legal framework evolves, it will become increasingly important to integrate data anonymization seamlessly into existing pipelines. Tools, APIs, and anonymization-as-a-service solutions can be effective approaches to address the problem

IV. PROPOSED METHOD

The proposed system Fig.1 takes as input a .csv file, which is pre-processed. Attributes are to be identified and classified. A suitable privacy model must be chosen for data anonymization. This requires a process that checks whether the information is still useful for intended purposes without infringing on security or otherwise. A synthetic dataset is generated by the protection algorithm through the use of differential privacy, which involves adding noise to the original data. Keeping out the essential characteristics of raw data while respecting individual privacy is achieved by this.

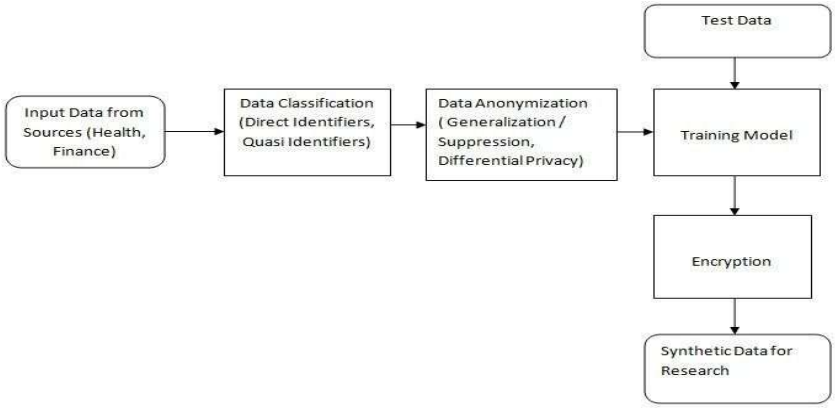

Fig. 1. Data Anonymization

The first step is to run a range query and synthesis on a raw data column to get appropriate synthetic data. Once the properties of the synthetic data have been verified against the original, the output of the anonymization model could be encrypted, and the data can be stored and shared among researchers. The third party can decrypt the shared data and utilize it for their research purposes. This ensures the secrecy and privacy properties of the data.

V. CONCLUSION

The suggested framework gives the best values for parameters like quasi-identifiers when adaptive anonymization is needed, and it gives transparency into the differential levels of privacy. The input data are under the influence of a statistical transformation done with the application of differential privacy, where privacy-preserving noise is injected in a systematic manner to preserve the individual records but leave the aggregate data utility intact. This work reviews some relational anonymization models that have been proposed for structured tabular data sets. The differential privacy method has been evaluated and shows an acquiescently strict mathematical derivation toward measurable privacy safeguards, and offers uniform safeguards for varying parameter configurations. This work suggests that the application of mathematically strict privacy methodologies greatly minimizes the likelihood of a privacy violation while safeguarding the integrity of the analytical result. Another feature is the capability to encrypt anonymized outputs for guaranteeing storage and transmission that is privacy breach risk-free, allowing anonymized data to be shared that can drive improved results for analytical or machine learning.

REFERENCES

[1] MeitY “Report of the Joint Committee on The Personal Data Protection Bill”. In:(2021).

[2] Ministry of Health & Family Welfare Government of India. “Digital Information Security in Healthcare, Act (DISHA)”. In: (2017).

[3] Abdul Majeed and Sungchang Lee. “Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey”. In: IEEE Access 9 (2021), pp. 8512–8545. doi: 10.1109/ACCESS.2020.3045700.

[4] Khaled El Emam and Fida Kamal Dankar. "Protecting privacy using kanonymity", Journal of the American Medical Informatics Association, 15(5):627–637, 2008.

[5] Joana Ferreira Marques and Jorge Bernardino. “Analysis of Data Anonymization Techniques”. In: (2020). doi: 10.5220/0010142302350241.

[6] Cristian Augusto et al. “Test-Driven Anonymization in Health Data: A Case Study on Assistive Reproduction”. In: (2020), pp. 81–82. doi: 10.1109/AITEST49225.2020.00019.

[7] Puneet Goswami and Suman Madan. “Privacy preserving data publishing and data anonymization approaches: A review”. In: (2017), pp. 139–142. doi: 10.1109/CCAA.2017.8229787.

[8] L. Sweeney, "k-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

[9] Latanya Sweeney. “Simple Demographics Often Identify People Uniquely”. In: Carnegie Mellon University, Data Privacy (2000). url: http : / / dataprivacylab . org /projects/identifiability/.

[10] J. Li et al., "A Privacy-Preserving Online Deep Learning Algorithm Based on Differential Privacy," 2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Rio de Janeiro, Brazil, 2023, pp. 559-564, doi: 10.1109/CSCWD57460.2023.10152847.

[11] Maryam Archie et al. “Who’s Watching ? De-anonymization of Netflix Reviews using Amazon Reviews”. In: (2018).

[12] Working Party European Union. “European Union Working Party on the protection of individuals with regard to the processing of personal data”. In: (2014).

[13] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. “Mondrian Multidimensional KAnonymity”.In: (2006), pp. 25–25. doi: 10.1109/ICDE.2006.101.

[14] Vanessa Ayala-Rivera et al. “A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners”. In: Transactions on Data Privacy 7 (Dec. 2014),pp. 337–370.

[15] Raj, H. W., & Balachandran, S. (2020). A Survey of Data Anonymization Techniques for Privacy- Preserving Mining in Bigdata. Journal of Computer Science, 16(2), 194-201

[16] S. V, N. Pant, N. B, S. I. Maaz and S. Luharuka, "Privacy Preserving Data Mining and its Applications: A Survey of Recent Developments," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 789-794, doi: 10.1109/ICAAIC56838.2023.10140450.

[17] Hewage, U.H.W.A., Sinha, R. & Naeem, M.A. Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review. Artif Intell Rev 56, 10427–10464 (2023).

[18] D. Gunawan, Y. S. Nugroho, F. Y. Al Irsyadi, I. C. Utomo, I. Andreansyah and S. Islam, "ℓρ-suppression: A Privacy Preserving Data Anonymization Method for Customer Transaction Data Publishing," 2022 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung, Indonesia, 2022, pp. 171-176, doi: 10.1109/ICITSI56531.2022.9970910.

[19] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k- anonymity," 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 2006, pp. 24-24, doi: 10.1109/ICDE.2006.1.

[20] Kajol Patel, 2019, A study on T-Closeness over K-anonymization Technique for Privacy Preserving in Big Data, International Journal of Engineering Research & Technology (IJERT) Volume 08, Issue 05 (May 2019),

[21] Wong, Kok-Seng & Anh, Tu & Bui, Dinh-Mao & Ooi, Shih Yin & Kim, Myung Ho. (2019). Privacy- Preserving Collaborative Data Anonymization with Sensitive Quasi-Identifiers. 1-6. 10.1109/CMI48017.2019.8962140.
[22] Wang J, Liu S, Li Y. A Review of Differential Privacy in Individual Data Release. International Journal of Distributed Sensor Networks. 2015;11(10). doi:10.1155/2015/259682
[23] Latanyasweeney,. (2012). Achieving K-Anonymity Privacy Protection Using Generalization And Suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 10. 10.1142/S021848850200165X.
[24] Kajol Patel, 2019, A study on T-Closeness over K-anonymization Technique for Privacy Preserving in Big Data, International Journal Of Engineering Research & Technology (IJERT) Volume 08, Issue 05 (May 2019)
[25] Dwork, C. (2008). Differential Privacy: A Survey of Results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds) Theory and Applications of Models of Computation. TAMC 2008. Lecture Notes in Computer Science, vol 4978. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79228-4_1
[26] Chang Sun, Johan van Soest, Michel Dumontier, Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy, Journal of Biomedical Informatics, Volume 143, 2023, 104404, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2023.104404.