IJMRT: Volume (6), Issue 8, 2024
International Journal of Multidisciplinary Research Transactions
(*A Peer Reviewed Journal*)
www.ijmrt.in

ISSN (Print):2663-2381
ISSN (Online):2663-4007

A Review on Privacy Preserving Data Mining Techniques and Applications

Dr. Shashidhar V[1*], Sutej Kulkarni[2], Suraj B Karia[3], Pranya Shetty[4], Samyak M H[5]

[1]*Assistant Professor, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*
[2]*Student, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*
[3]*Student, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*
[4]*Student, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*
[5]*Student, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*

[*]Corresponding author

DoI: https://doi.org/10.5281/zenodo.13371837

Abstract

In the realm of data mining, techniques such as clustering, association rule mining, and classification serve as powerful tools to unveil concealed insights within datasets. However, the proliferation of sensitive personal information in data sources has sparked heightened privacy concerns. In response, a myriad of privacy-protection solutions have emerged. This paper aims to scrutinize the latest advancements in data mining privacy protection and furnish a comprehensive review of anonymization approaches. By delving into the intricacies of safeguarding personal data, the study contributes to the ongoing discourse on data privacy and security in the era of big data. It navigates the delicate balance between the promises of data mining and the imperative to shield personal information.

Keywords: Data mining, Privacy preservation, Anonymization.

1. Introduction

In the landscape of data mining, where the extraction of meaningful patterns and insights is paramount, the surge in sensitive personal information within datasets has given rise to profound privacy concerns. As we delve deeper into the age of big data, the juxtaposition of leveraging data mining's potential and safeguarding personal data becomes increasingly intricate. The very essence of data mining lies in unraveling patterns, making sense of intricate relationships, and deriving valuable knowledge from vast and complex datasets. However, this pursuit encounters a critical challenge – the imperative to preserve the privacy of individuals whose data is woven into the fabric of these expansive datasets.

The ubiquity of personal information in contemporary data sources necessitates a vigilant exploration of privacy-protection solutions. The promise of data mining, encapsulating techniques ranging from clustering to classification, must be harmonized with the ethical responsibility of shielding sensitive details from unauthorized exposure. Against this backdrop, the focus of this paper is to meticulously examine the latest innovations in the domain of data mining privacy protection. It aspires to cast a discerning eye on the myriad anonymization approaches that have evolved in response to the escalating concerns about data privacy and security.

As we embark on this exploration, it becomes evident that the intersection of data mining and privacy preservation is not a mere technical challenge but a profound societal consideration. Striking the right balance is not only crucial for the successful application of data mining techniques but is fundamentally tied to upholding the rights and confidentiality of individuals. The contemporary data-driven landscape demands a nuanced understanding of privacy preservation mechanisms, navigating the evolving terrain of legal, ethical, and technological

dimensions. In essence, this study aims to contribute to the ongoing dialogue by addressing the intricate issues of safeguarding personal data while harnessing the vast potential embedded in the expansive realm of data mining.

I. Micro-Aggregation

Additive microaggregation is a privacy-preserving technique used in data mining and statistical disclosure control. It involves grouping similar records and perturbing their values to achieve a certain level of anonymity while minimizing information loss. The goal is to protect sensitive information within the dataset, making it challenging for adversaries to identify individuals or disclose confidential details. This method is often employed in scenarios where preserving privacy is crucial, such as open government data, collaborative filtering in recommender systems, and dynamic data release in the context of big data.

Balkis Abidi proposed a novel microaggregation method addressing the limitations of conventional techniques like k-Anonymity. HM-PFSOM utilizes fuzzy possibilistic clustering in a hybrid approach, applying the anonymization process to distinct blocks of similar data. By dynamically determining the privacy parameter k, the method ensures diversity in confidential attributes within the anonymized microdata, reducing the risk of disclosing sensitive information. The approach demonstrates sensitivity to real-world datasets, showcasing advancements in privacy preservation for effective data mining.[1]

Jae-Seong Lee's study focuses on privacy-preserving data mining on open government data (OGD), addressing challenges related to privacy concerns, lack of identifiers, and variations in data sharing methods. The proposed micro-type PPDM algorithm integrates de-identified

OGDs using micro-aggregation and distance-based record linkage, allowing users to adjust the privacy threshold. The hybrid model aligns with the objectives of the International Open Data Charter, emphasizing in-depth analysis and privacy-preserving data mining. [2]

YanYan addresses privacy concerns in the age of big data, proposing a privacy-preserving dynamic data publishing method based on microaggregation. The method incorporates indicators to evaluate synonymous linkages between non-numerical sensitive values, enhancing clustering effects. A dynamic update program is introduced to handle the challenges of dynamic data release and prevent privacy leakage. The paper reviews existing privacy-preserving data publishing methods, emphasizing the limitations of traditional approaches like K-anonymity. The proposed dynamic data release algorithm (DRASL) stands out as a novel contribution to privacy preservation.[3]

Fran Casino, addresses privacy concerns in recommender systems, proposing a novel Privacy-Preserving Collaborative Filtering (PPCF) method based on variable-group-size microaggregation. The method ensures k-anonymity for users while outperforming existing methods in recommendation accuracy. The study emphasizes the significance of user privacy in recommender systems and introduces new evaluation metrics. The proposed PPCF approach offers a promising balance between privacy preservation and recommendation quality, contributing to the discourse on achieving a balance between personalized recommendations and user privacy protection.[4]

2. Data Swapping

Data mining, while a powerful tool for extracting valuable insights, poses significant challenges when it comes to preserving individual privacy. In this context, the technique of data swapping has emerged as a pivotal method, allowing for meaningful analysis while safeguarding sensitive information. This essay provides an in-depth exploration of data swapping, drawing insights from five influential papers that contribute significantly to this evolving field.

The foundational aspects of data swapping are introduced in Dedi Gunawan's "Classification of Privacy Preserving Data Mining Algorithms: A Review" (December 2020). Gunawan emphasizes the versatility of rank swapping, specifically designed for numerical and ordinal data. The paper showcases its effectiveness in set-valued databases, introducing innovative variations like partSwap and fullSwap. Gunawan's work lays the groundwork for understanding the nuanced applications of data swapping in various contexts.[5]

Anastasiia Pika's "Towards Privacy-Preserving Process Mining in Healthcare" (January 2020) takes the application of data swapping a step further by proposing a comprehensive plan. This plan integrates encryption, noise addition, and legal frameworks like GDPR, offering a strategic approach to balancing privacy and data utility in healthcare scenarios. Pika's study underlines the intricate application of data swapping, ensuring robust privacy protection while facilitating meaningful data analysis in the healthcare domain.[6]

In "Navigating the diverse landscape of privacy-preserving data mining methods" by Negar Nasiri, data swapping is a versatile tool crucial for privacy preservation. The paper provides a comprehensive review of privacy methods, highlighting the adaptability of data swapping in

safeguarding privacy during data collection, sharing, and study. Nasiri and Keyvanpour underscore the pivotal role of data swapping in diverse data mining scenarios.

Mercedes Rodriguez-Garcia, contributed significantly to the exploration of data swapping with their paper titled "Utility-preserving privacy protection of nominal data sets via semantic rank swapping" (May 2020). Introducing semantic rank swapping, the study leverages ontologies to protect privacy while preserving utility, particularly in nominal data sets. The experiments, notably in medical records, highlight the effectiveness of these techniques in maintaining data meaning, showcasing data swapping's adaptability.[7]

Finally, Majid Rafiei's "Privacy-Preserving Data Publishing in Process Mining" (January 2021) focuses specifically on data swapping in the context of process mining data publishing. The paper introduces techniques like rank swapping and data swapping, emphasizing the importance of privacy metadata. It provides formal models, an XES standard extension, and Event Log Abstraction (ELA), supporting the practical implementation of data swapping for privacy preservation in the domain of process mining.[8]

In conclusion, data swapping emerges as a cornerstone in privacy-preserving data mining, ensuring the delicate balance between extracting valuable insights and preserving individual privacy. These papers collectively contribute to our understanding of data swapping's role in achieving this balance. As the field continues to evolve, data swapping remains a pivotal tool in the arsenal of privacy-preserving methodologies, ensuring that valuable insights can be gleaned from data without compromising individual privacy. Exploring data swapping across these diverse studies showcases its adaptability and underscores its indispensable role in the evolving landscape of privacy-preserving data mining.

3. Randomization Noise Method

In the ever-evolving landscape of data mining, where the quest for valuable insights coexists with the imperative to protect individual privacy, the utilization of innovative methods becomes paramount. One such groundbreaking approach is the integration of randomization noise methods, a pioneering technique that introduces a layer of uncertainty into the data to safeguard sensitive information.

The fundamental challenge in data mining lies in striking a delicate balance between extracting meaningful patterns and preserving the privacy of individuals whose data is being analyzed. The conventional methods of data anonymization, encryption, and obfuscation have been instrumental in privacy preservation, but they often encounter limitations in providing comprehensive protection without compromising the utility of the data.

Randomization noise methods offer a novel paradigm, injecting a controlled level of randomness into the dataset to obscure the original values. The essence of this approach is akin to adding a protective veil over the data, making it challenging for any external entity to discern specific details while still allowing for meaningful analysis.

The first phase of this privacy-preserving technique is observed in the innovative Privacy-Preserving Distributed Machine Learning via Local Randomization and ADMM Perturbation framework proposed by Xin Wang et al. in 2020. In Phase 1, users locally randomize their labels by introducing randomization noise before transmitting data to a centralized server. This initial layer of protection ensures that sensitive labels are shielded from direct exposure, setting the stage for secure data collaboration.[9]

The second phase of randomization noise methods is exemplified in the work of M.A.P. Chamikara et al. with their PABIDOT method, published in 2019. PABIDOT introduces random noise to datasets by flipping, shifting, and rotating numbers, adding an element of

controlled chaos. What makes this approach innovative is its ability to maintain the usefulness of the data for analysis while obscuring its specifics. PABIDOT acts as a data security guard, ensuring that information remains safe from prying eyes while still serving its analytical purpose.[10]

The significance of randomization noise is further underscored in the paper titled "A Brief Study of Privacy-Preserving Practices (PPP) in Data Mining" by Dhinakaran D and Joe Prathap P.M in 2020. The paper acknowledges the trade-off involved in introducing random noise, emphasizing that an excessive amount can lead to information loss and reduced data mining effectiveness. It advocates for a nuanced approach, wherein the amount of noise added is carefully managed to strike the right balance between privacy and utility.[11]

Randomization noise methods, however, are not without their challenges. The delicate interplay between privacy preservation and data utility necessitates careful consideration. The amount of noise added must be calibrated to ensure that the data remains useful for analysis, without compromising the privacy it seeks to protect.

In conclusion, randomization noise methods represent a paradigm shift in privacy-preserving data mining. From local randomization to controlled chaos introduced by PABIDOT, these methods signify a promising trajectory in the quest for secure and meaningful data analysis. While challenges persist, the innovative integration of randomization noise stands as a testament to the dynamism of the field, propelling it towards a future where data can be mined for insights without sacrificing individual privacy.

4. Rounding

Rounding is a process that involves converting continuous data into a discrete set of values, typically by replacing specific values with more general representative ones. This technique is

often employed in statistical disclosure control to protect individual privacy when dealing with sensitive or confidential information. Rounding aims to coarsen the data, making it less granular and helps minimize the risk of disclosure while preserving the overall utility of the information.

Navoda Senavirathne along with Vicenç Torra explore the concept of rounding as a method for protecting privacy in statistical disclosure control for continuous data. The authors address three objectives: studying alternative methods for obtaining rounding values, empirically evaluating rounding as a data protection technique, and analyzing the impact of data rounding on machine learning models. They consider various rounding methods, emphasizing the importance of selecting appropriate rounding points to balance data protection and disclosure risk. The authors find that microaggregation-based techniques offer a fair trade-off between information loss and disclosure risk.[12]

Raffael BILD, focuses on implementing reliable data anonymization methods in biomedical research to address privacy concerns. The authors propose a computing framework based on fractional and interval arithmetic to improve the reliability of anonymization implementations. The study evaluates the impact of floating-point errors on privacy guarantees and suggests a reliable computing framework integrated into the ARX data anonymization tool. The authors emphasize the importance of privacy protection in the context of data-driven approaches in precision medicine.[13]

This literature mapping study authored by Zheming Zuo, explores data anonymization challenges in digital health care, emphasizing the importance of privacy protection amidst advanced health technologies. The authors review various aspects of data anonymization, including operations, privacy models, reidentification risk, and usability metrics. They identify

gaps in existing studies and stress the need for more research efforts to achieve a balance between privacy preservation and usability in health care.[14]

Baek Kyung Song's paper addresses data privacy and security concerns in cloud servers, focusing on privacy-preserving data mining. The authors propose a bitwise fully homomorphic encryption (FHE) scheme to process data in the encrypted domain without decryption. They design atomic operations using bitwise logical circuits, expanding the range of operations supported by FHE. The paper demonstrates the practicality of bitwise FHE schemes in various data mining techniques, highlighting the advantages of representing real-valued data in bits for generalized encryption.[15]

5. Additive Noise

Additive noise is a technique used in privacy-preserving data mining to protect individual entries in a database while responding to queries. By introducing random noise to the query responses, this method aims to achieve a balance between preserving privacy and maintaining data utility.

Benjamin Denham's paper addresses privacy concerns in data stream mining. The authors propose two innovative data perturbation methods utilizing random projection, random translation, and two types of additive noise: independently generated noise for each record (RPIN) and noise accumulating over the stream's lifetime (RPCN). The cumulative noise injection scheme demonstrated superior performance in terms of privacy guarantees and classification accuracy compared to other schemes, showcasing a notable trade-off between privacy and utility.[16]

This paper discusses challenges in privacy-preserving data mining for high-dimensional continuous data. The author Shashidhar Virupaksha introduce Anonymized Noise Addition in

Subspaces (ANAS), a method that incorporates random noise within subspace limits to enhance cluster identification and reduce data loss. ANAS outperformed existing techniques, demonstrating its effectiveness in cluster identification, data loss reduction, information preservation, and privacy enhancement.[17]

Shashidhar Virupaksha presented a study introducing Subspace-Based Noise Addition (SBNA) as a solution to privacy concerns in high-dimensional datasets. In contrast to conventional methods that add noise independently to each dimension, SBNA customizes noise addition to pertinent subspaces, reducing data loss and preserving information content, thereby optimizing cluster identification. Experimental findings on benchmark datasets demonstrate substantial enhancements compared to similar approaches, highlighting SBNA's superior performance in terms of data utility, cluster identification, and information measures.[18]

In the work by Jinzhao Shan the Range Noise Perturbation (RNP) method is presented as an approach for privacy-preserving data mining. While the paper lacks in-depth specifics, it highlights the deployment of the SVM machine-learning algorithm to achieve a harmonious equilibrium between data utility and security. Experimental outcomes indicate that RNP outperforms the NMF and NMFSVD algorithms, providing a pragmatic solution for accurate predictions while ensuring the safeguarding of sensitive information, particularly in the domain of big data and data analysis.

6. Sampling

Sampling is a fundamental technique in research and data analysis that involves selecting a subset of individuals or elements from a larger population to conclude the entire population. It is impractical or impossible to study an entire population directly, especially in cases where

the population is large or dynamic. Sampling allows researchers to make inferences about a population based on the analysis of a smaller, representative subset.

In this study, Mehdi Gheisari, present the "Ontology-Based Privacy-Preserving" (OBPP) framework as a solution to the challenges posed by the management of data generated by Internet of Things (IoT) devices in smart cities. The three-module framework incorporates ontology-based data storage, semantic reasoning rules, and a privacy rules manager to effectively address issues related to device heterogeneity, improve service quality, and dynamically adapt privacy behaviors. Through simulations, the study demonstrates OBPP's superior performance, surpassing existing solutions by offering enhanced affordability and robustness against potential information leakages.[19]

Addressing privacy issues within data centers, Weibei Fan, Jing He, put forth a paper that introduces a local differential privacy-based classification algorithm. Recognizing the diversity of devices within data centers, the algorithm integrates a differential privacy protection mechanism, employing Laplace noise during the pattern mining process. The paper underscores the importance of privacy protection in the changing landscape of cloud computing and big data applications.[20]

M.A.P. Chamikara, P. Bertok, contribute to the discourse on privacy in the era of burgeoning data with their paper introducing the "Privacy Preservation Algorithm for Big Data Using Optimal Geometric Transformations" (PABIDOT). This algorithm is designed to efficiently and scalably perturb data in a nonreversible manner, aiming to strike a delicate balance between privacy and utility in the realm of big data. Positioned as a response to challenges posed by existing privacy-preserving methods, PABIDOT emerges as a solution addressing privacy concerns associated with the escalating volume of data.[10]

Celestine Iwendi presents the N-Sanitization framework in their paper as a response to privacy concerns stemming from the widespread integration of the Internet of Medical Things (IoMT). This framework offers a holistic solution for sanitizing sensitive terms within medical documents, incorporating semantic privacy through random sampling. Notably, N-Sanitization showcases improvements in both detection accuracy and data utility, positioning it as a promising solution for safeguarding sensitive medical information within the IoMT landscape.[21]

7. Data Perturbation

Data perturbation is a privacy-preserving technique employed in data mining to protect sensitive information during analysis. It involves intentionally introducing controlled noise or modifications to the original data, aiming to obfuscate individual records while preserving overall statistical properties. Perturbation methods, such as random noise addition or enhanced principal component analysis (PCA), help strike a balance between data privacy and utility. By altering the dataset in a controlled manner, data perturbation enables the extraction of valuable insights without compromising the confidentiality of individual records. This approach is crucial in scenarios where privacy concerns are paramount, emphasizing the importance of adopting such techniques in data analysis for improved security and operational efficiency.

In her paper, Ritu Ratra, explore hybrid perturbation methods like enhanced principal component analysis (PCA) and improved random projection in privacy-preserving data mining (PPDM). The study, conducted on cardiovascular and hypothyroid datasets, demonstrates that these perturbation techniques outperform original data in accuracy, TP/FP rates, F-measure, and runtime. Emphasizing the importance of safeguarding personal information in data mining, the research advocates for the adoption of privacy-preserving strategies to enhance both security and performance throughout the data analysis process. [22]

The paper by S. Singaravelan, P. provides an overview of privacy-preserving data mining (PPDM) techniques, with a focus on data perturbation, Trust Third Party, Secure Multiparty Computation (SMC), and game theoretic approaches. The emphasis is on modifying or anonymizing raw data to safeguard identifiers and maintain privacy during the mining process. Notably, Secure Multiparty Computation (SMC) is highlighted as a robust strategy within PPDM. SMC allows distrustful entities to collaboratively analyze data while keeping individual inputs and final outputs private. This ensures that participants can derive insights from collective data analysis without compromising the confidentiality of their inputs, exemplifying an effective balance between privacy and analytical value.[23]

In their systematic literature review (SLR), U. H. W. A. Hewage, extensively investigate privacy-preserving data mining (PPDM) and data stream mining (PPDSM) techniques. The review categorizes PPDM methods into perturbation and non-perturbation approaches, specifically exploring data perturbation, random projection, condensation, and fuzzy logic-based perturbation. The emphasis is on achieving a balance between accuracy and privacy, addressing challenges associated with applying these methods to data streams. Through an analysis of strengths, weaknesses, and applicability across various data mining tasks, the SLR underscores the importance of optimizing the accuracy-privacy trade-off. Additionally, it identifies gaps in adapting PPDM techniques to data streams, urging further research to enhance the applicability of privacy-preserving methods in the context of streaming data.[24]

The document authored by Vijaya Pinjarkar delves into the realm of data privacy within medical contexts, with a specific focus on mental health information. Highlighting the challenges associated with preserving privacy while sharing sensitive data for research purposes, the document introduces an innovative perturbation method that involves pre-processing, encoding, and mathematical operations such as RMS addition and sorting. The

primary objective is to safeguard data privacy before third-party sharing. By emphasizing the delicate balance between privacy protection and data quality, the document underscores the importance of maintaining accuracy while preserving privacy in the domain of privacy-preserving data mining. It concludes by proposing further enhancements through encryption techniques and stresses the ongoing need to carefully balance privacy preservation and information loss, particularly crucial when handling sensitive medical data for research.[25]

8. PRAM

PRAM, or Privacy-preserving Randomized Aggregatable Mechanisms, is a method employed in the context of differential privacy. It aims to protect individual privacy in datasets by introducing randomness during data aggregation processes. The main goal is to provide privacy guarantees by ensuring that the inclusion or exclusion of any individual's data does not significantly impact the overall results. However, the research on PRAM has identified limitations, including low privacy protection efficiency and challenges in maintaining data utility. Alternative perturbative methods, such as additive noise and data swapping, are considered for enhancing data privacy in scenarios like the Japanese Population Census.

In their research, Shinsuke Ito, explore the application of PRAM and other differential privacy methods to protect data privacy in the Japanese Population Census. The study highlights challenges in selecting suitable mechanisms and reveals limitations of PRAM, including low privacy protection efficiency and data utility issues. Comparative analysis indicates PRAM's inferiority to alternative methods in privacy protection efficiency, emphasizing the need to consider the ε value for balancing privacy and utility. The research underscores the drawbacks of PRAM, leading to a discussion on the potential of alternative perturbative methods like additive noise and data swapping for ensuring data privacy in the census context. Overall, the

study provides critical insights into the efficacy, limitations, and considerations of PRAM and other differential privacy methods in safeguarding individual census data privacy. [26]

This paper, authored by S. Singaravelan, P. Gopalsamy, explores Privacy Preserving Data Mining (PPDM) techniques, with a specific focus on data perturbation methods. PPDM involves modifying or anonymizing sensitive raw data to safeguard identifiers and maintain privacy during mining operations. Various techniques like k-anonymization, data shuffling, and micro-aggregation are discussed, aiming to alter or eliminate identifiable information while preserving privacy. Notably, the paper emphasizes the robustness of Secure Multiparty Computation (SMC) among PPDM strategies. SMC allows entities with inherent distrust to collaboratively mine data while ensuring individual inputs and final outputs remain private. This approach secures computations at the individual level, enabling participants to glean insights from collective data analysis without exposing sensitive information, thus striking a balance between privacy and analytical value.[27]

Mohammad Anagreh, introduces a Parallel Privacy-Preserving Minimum Spanning Tree Algorithm designed for large graphs, with a focus on data privacy within the Parallel Random Access Machine (PRAM) model. Implemented on the Sharemind secure multiparty computation (SMC) platform, the algorithm utilizes Prim's algorithm and an oblivious reading subroutine to ensure privacy in computations. The research contributes to both the security and efficiency aspects of the proposed algorithm, providing a security theorem that guarantees the privacy of inputs and computation integrity. Benchmarking across various graph sizes highlights the practical applicability of the algorithm, emphasizing the significance of secure and efficient privacy-preserving solutions in data analysis and computation.[28]

In the study by Steven Verheyena a comparison is made between the Total-Set Pairwise Rating Method (PRaM) and Spatial Arrangement Method (SpAM) for obtaining similarity data in the conceptual domain. The research finds that while PRaM is more time-consuming, it provides more reliable and central data, making it preferable for non-perceptual stimuli with a manageable number. The study places a strong emphasis on ethical principles, ensuring informed consent and privacy protection for participants. It highlights the importance of balancing time efficiency and data reliability, particularly considering the impact on participants' privacy when choosing between PRaM and SpAM for data collection. [29]

Table.1. Summary Table

Authors	Title	Year	Summary
Balkis Abidi, Sadok Ben Yahia & Charith Perera [1]	Hybrid micro aggregation for privacy-preserving data mining	2018	HM-PFSOM introduces an innovative microaggregation method, combining fuzzy possibilistic clustering with a hybrid approach.
Jae-Seong Lee, Seung-Pyo Jun [2]	Privacy-preserving data mining for open government data from heterogeneous sources	2020	The study proposes a micro-type privacy-preserving algorithm for open government data (OGD) mining, aligning with the International Open Data Charter's objectives.
YanYan, Anselme Herman Eyeleko, Adnan Mahmood, Jing Li, Zhuoyue Dong & Fei Xu [3]	Privacy preserving dynamic data release against synonymous linkage based on microaggregation	2022	The paper proposes a novel microaggregation-based dynamic data publishing method for privacy concerns in big data, specifically addressing non-numerical sensitive information
Fran Casino, Constantinos Patsakis, Agusti Solanas [4]	Privacy-preserving collaborative filtering: A new approach based on variable-group-size microaggregation	2019	The article proposes a novel Privacy-Preserving Collaborative Filtering (PPCF) method using variable-group-size microaggregation for k-anonymity, outperforming established techniques like ALS and GNA in

			both user privacy enhancement and recommendation accuracy.
Dedi Gunawan [5]	Classification of Privacy Preserving Data Mining Algorithms: A Review	2020	Data swapping for disclosure protection uses rank swapping for numerical data; recent variations address limitations in set-valued databases, especially in sensitive domains like health records.
Anastasiia Pika [6]	Towards Privacy-Preserving Process Mining in Healthcare	2020	Paper focuses on safeguarding private healthcare information through encryption, noise addition, and legal frameworks like GDPR in process analysis, aiming for a balance between privacy and data utility.
Mercedes Rodriguez-Garciaa, Montserrat Batetb, David Sánchezа [7]	Utility-preserving privacy protection of nominal data sets via semantic rank swapping	2020	Emphasizes personal data privacy through rank swapping and data swapping, extends rank swapping to meaning-focused data using ontologies, and proposes solutions for secure and meaningful data analysis while preserving privacy.
Majid Rafiei and Wil M.P. van der Aalst [8]	Privacy-Preserving Data Publishing in Process Mining	2021	Paper enhances process mining data privacy using techniques like rank swapping and data swapping, introduces privacy metadata, extends XES standard, and provides a Python library for privacy tools.
Xin Wang, Hideaki Ishii, Linkang Du, Peng Cheng, and Jiming Chen [9]	Privacy-Preserving Distributed Machine Learning via Local Randomization and ADMM Perturbation	2020	Research introduces PDML framework for privacy-preserving distributed machine learning, safeguarding sensitive labels through local randomization and collaborative server efforts with noise addition, ensuring privacy and robustness.
M.A.P. Chamikara, P. Bertok, D. Liu, S. Camtepe, I. Khalil [10]	Efficient Privacy Preservation of Big Data for Accurate Data Mining	2019	PABIDOT, a privacy method for large datasets, introduces random noise via flipping and shifting, ensuring data privacy without sacrificing accuracy or speed, outperforming other methods and serving as a privacy superhero for data integrity and usefulness.
Dhinakaran D, Joe Prathap P.M	A Brief Study of Privacy-Preserving	2020	Explores privacy-preserving data mining, emphasizing the delicate

[11]	Practices (PPP) in Data Mining		balance between privacy and utility. Investigates encryption, anonymization, and differential privacy, advocating for cryptographic-based methods while acknowledging ongoing research needs for effective privacy preservation.
Navoda Senavirathne, Vicenç Torra [12]	Rounding based continuous data discretization for statistical disclosure control	2019	The paper explores the use of rounding in statistical disclosure control for privacy protection with continuous data, focusing on coarsening to discrete values and integrating methods like microaggregation.
Raffael BILD, Klaus A. KUHN and Fabian PRASSER [13]	Better Safe than Sorry - Implementing Reliable Health Data Anonymization	2020	The study proposes a computing framework for reliable data anonymization in precision medicine datasets, integrating fractional and interval arithmetic into the ARX tool to address challenges in translating privacy models into software and ensure accuracy in complex mathematical expressions over real numbers.
Zheming Zuo; Matthew Watson; David Budgen; Robert Hall; Chris Kennelly; Noura Al Moubayed [14]	Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study	2020	This literature mapping study reviews challenges in data anonymization in digital healthcare, highlighting privacy concerns amid advanced health technologies and addressing legal aspects, including GDPR and ICO regulations.
Baek Kyung Song,Joon Soo Yoo,Miyeon Hong, and Ji Won Yoon [15]	A Bitwise Design and Implementation for Privacy-Preserving Data Mining: From Atomic Operations to Advanced Algorithms	2019	The paper proposes homomorphic encryption (HE) for cloud data privacy, introducing a practical bitwise fully homomorphic encryption (FHE) computation method for secure machine learning, aiming to bridge the gap between FHE and machine learning integration for broader applicability.
Benjamin Denham, Russel Pears, M. Asif Naeem [16]	Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream	2020	Paper introduces RPIN and RPCN techniques for privacy-preserving data stream mining. RPCN shows superior balance between privacy and accuracy, achieving 15% higher accuracy and

	mining		10% greater utility in high-dimensional set-valued data mining compared to existing schemes.
Shashidhar Virupaksha, Venkatesulu Dondeti [17]	Anonymized noise addition in subspaces for privacy preserved data mining in high dimensional continuous data	2021	Paper introduces ANAS, a privacy-preserving method for high-dimensional data mining, with improved performance and surpassing existing techniques in cluster identification, data loss reduction, and privacy enhancement.
Shashidhar Virupaksha, Venkatesulu Dondeti [18]	Subspace based noise addition for privacy preserved data mining on high dimensional continuous data	2020	Study introduces SBNA algorithm for privacy-preserving high-dimensional data mining, achieving 80% enhanced data utility and 90% improved cluster identification on benchmarks.
Mehdi Gheisari, Hamid Esmaeili Najafabadi, Jafar A. Alzubi, Jiechao Gao, Guojun Wang, Aaqif Afzaal Abbasi, Aniello Castiglione[19]	OBPP: An ontology-based framework for privacy-preserving in IoT-based smart city	2021	The paper introduces OBPP, an Ontology-Based Privacy-Preserving Framework for smart city IoT data, demonstrating superiority in simulations with affordability, robustness against information leakages, and the ability to address heterogeneity and privacy preservation in dynamic environments.
Weibei Fan, Jing He, Mengjiao Guo, Peng Li, Zhijie Han, Ruchuan Wang [20]	Privacy-preserving classification on local differential privacy in data centers	2019	Paper proposes a local differential privacy-based classification algorithm for data center privacy, addressing device heterogeneity and emphasizing privacy in cloud computing and big data applications.
Celestine Iwendi, Syed Atif Moqurrab, Adeel Anjum, Sangeen Khan, Senthilkumar Mohan, Gautam Srivastava [21]	N-Sanitization: A Semantic privacy-preserving Framework for Unstructured Medical Datasets	2020	Paper introduces N-Sanitization, a framework for IoMT privacy emphasizing semantic privacy, showcasing notable improvements in PHI detection accuracy and overall data utility through advanced processes and realistic evaluations.
Ritu Ratra,Preeti Gulia,Nasib Singh Gill [22]	Performance analysis of perturbation based privacy preserving techniques: an experimental perspective	2023	The research emphasizes the superior performance of hybrid perturbation methods in privacy-preserving data mining, showcasing their effectiveness in maintaining data privacy.

S.Singaravelan , P. Gopalsamy and S. Balaganesh [23]	Accumulation of data perturbation techniques for privacy preserving data classification	2020	Privacy Preserving Data Mining (PPDM) uses techniques like data perturbation and Secure Multiparty Computation (SMC) to modify or anonymize sensitive data.
U. H. W. A. Hewage, R. Sinha, M. Asif Naeem [24]	Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review	2023	The literature review categorizes privacy-preserving data mining methods, highlighting the importance of balancing accuracy and privacy.
Vijaya Pinjarkar, Amit Jain, Anand Bhaskar, Prateek Srivastava [25]	Pertinent Exploration of Privacy Preserving Perturbation Methods	2020	The document introduces a novel perturbation method for preserving privacy in mental health data sharing, emphasizing the balance between privacy protection and data quality.
Shinsuke Ito , Masayuki Terada , Shunsuke Kato [26]	The Potential of Differential Privacy Applied to Detailed Statistical Tables Created Using Microdata from the Japanese PopulationCensus	2023	The research examines the application of PRAM and differential privacy methods in the Japanese Population Census.
S.Singaravelan , P. Gopalsamy and S. Balaganesh [27]	Accumulation of data perturbation techniques for privacy preserving data classification	2020	Privacy Preserving Data Mining (PPDM) techniques, including Secure Multiparty Computation (SMC), modify sensitive raw data for privacy preservation.
Mohammad Anagreh1, Eero Vainikko and Peeter Laud [28]	Parallel Privacy-preserving Computation of Minimum Spanning Trees	2021	The paper presents a privacy-preserving minimum spanning tree algorithm for large graphs in the PRAM model, implemented on Sharemind SMC.
Steven Verheyena, Anne White, and Gert Storms [29]	A Comparison of the Spatial Arrangement Method and the Total-Set Pairwise Rating Method for Obtaining Similarity Data in the Conceptual Domain	2022	The study compares PRaM and SpAM for similarity measures, highlighting PRaM's reliability but time-consuming nature, making it preferable for manageable non-perceptual stimuli.

Farhad Farokhi, Henrik Sandberg [30]	Ensuring privacy with constrained additive noise by minimizing Fisher information	2018	Research introduces a method for individual entry privacy in database queries, using constrained additive noise optimization through a convex lower bound approach. Applied in smart meter privacy and dynamic estimation problems.
Yuncheng Wu, Shaofeng Cai, Xiaokui Xiao, Gang Chen, Beng Chin Ooi [31]	Privacy-Preserving Vertical Federated Learning for Tree-based Models	2020	Paper introduces Pivot, a solution for privacy in vertical federated learning, offering robust defense against semi-honest adversaries during decision tree training and prediction, applicable to tree ensemble models, and advancing privacy-preserving methodologies significantly.

9. Conclusion

The exploration of Privacy Preserving Data Mining Techniques and Applications through methods like noise addition, data swapping, PRAM, data perturbation, and rounding underscores a multifaceted approach to safeguarding information integrity. These techniques signify a concerted effort to reconcile the conflicting demands of data utility and privacy protection. Noise addition injects controlled randomness into datasets, obscuring individual details while maintaining overall patterns. Data swapping swaps attributes among records to preserve privacy while upholding the analytical value of the dataset. PRAM (Privacy-preserving Association Rule Mining) algorithms navigate the intricate balance between mining valuable patterns and safeguarding sensitive information. Perturbation techniques introduce controlled distortions to data, protecting individual identities without compromising the dataset's utility. Rounding involves discretizing numerical values to mitigate the risk of reidentification while retaining analytical value. Together, these methods establish a framework that reflects the ongoing quest for privacy-conscious data mining strategies.

However, despite their efficacy, these techniques aren't without limitations. Noise addition, while preserving privacy, can potentially impact the accuracy of analysis. Data swapping, although effective, may alter the dataset's underlying characteristics, impacting the fidelity of mining outcomes. PRAM algorithms often involve complex computations that can be resource-intensive, affecting scalability. Perturbation techniques must strike a delicate balance between privacy and utility, as excessive distortion can diminish the dataset's analytical value. Rounding, though simple, might not provide foolproof protection against all reidentification risks. Acknowledging these limitations underscores the need for continual innovation and refinement in privacy-preserving data mining approaches.

In essence, the review underscores the crucial strides made in privacy-preserving data mining, showcasing an array of methodologies designed to shield sensitive information without compromising the analytical potential of datasets. These techniques, while not devoid of limitations, form a robust foundation for future advancements in ensuring data privacy and utility coexist harmoniously. As technology evolves and data becomes more pervasive, ongoing research and innovation in this field will be pivotal in addressing emerging privacy challenges while harnessing the full potential of data-driven insights.

REFERENCES

[1]. Abidi, B., Ben Yahia, S. and Perera, C., 2020. Hybrid microaggregation for privacy preserving data mining. Journal of Ambient Intelligence and Humanized Computing, 11, pp.23-38.

[2]. Lee, J.S. and Jun, S.P., 2021. Privacy-preserving data mining for open government data from heterogeneous sources. Government Information Quarterly, 38(1), p.101544.

[3]. Yan, Y., Eyeleko, A.H., Mahmood, A., Li, J., Dong, Z. and Xu, F., 2022. Privacy preserving dynamic data release against synonymous linkage based on microaggregation. Scientific Reports, 12(1), p.2352.

[4]. Casino, F., Patsakis, C. and Solanas, A., 2019. Privacy-preserving collaborative filtering: A new approach based on variable-group-size microaggregation. Electronic Commerce Research and Applications, 38, p.100895.tal record linkage. Journal of Big Data, 9(1), 1-27.

[5]. Rodriguez-Garcia, M., Batet, M. and Sánchez, D., 2019. Utility-preserving privacy protection of nominal data sets via semantic rank swapping. Information Fusion, 45, pp.282-295.

[6]. Gunawan, D., 2020. Classification of privacy preserving data mining algorithms: a review. Jurnal Elektronika dan Telekomunikasi, 20(2), pp.36-46.

[7]. Pika, A., Wynn, M.T., Budiono, S., ter Hofstede, A.H., van der Aalst, W.M. and Reijers, H.A., 2019. Towards privacy-preserving process mining in healthcare. In Business Process Management Workshops: BPM 2019 International Workshops, Vienna, Austria, September 1–6, 2019, Revised Selected Papers 17 (pp. 483-495). Springer International Publishing..

[8]. Rafiei, M. and van der Aalst, W.M., 2020. Privacy-preserving data publishing in process mining. In Business Process Management Forum: BPM Forum 2020, Seville, Spain, September 13–18, 2020, Proceedings 18 (pp. 122-138). Springer International Publishing.

[9]. Chamikara, M.A.P., Bertok, P., Liu, D., Camtepe, S. and Khalil, I., 2020. Efficient privacy preservation of big data for accurate data mining. Information Sciences, 527, pp.420-443.

[10]. Dhinakaran, D. and Joe Prathap, P.M., 2020. A brief study of privacy-preserving practices (PPP) in data mining. TEST Eng Manage, 82, pp.7611-7622.

[11]. Wang, X., Ishii, H., Du, L., Cheng, P. and Chen, J., 2020. Privacy-preserving distributed machine learning via local randomization and ADMM perturbation. IEEE Transactions on Signal Processing, 68, pp.4226-4241.

[12]. Senavirathne, N. and Torra, V., 2019. Rounding based continuous data discretization for statistical disclosure control. Journal of Ambient Intelligence and Humanized Computing, pp.1-19.

[13]. Bild, R., Kuhn, K.A. and Prasser, F., 2020. Better safe than sorry–implementing reliable health data anonymization. In Digital Personalized Health and Medicine (pp. 68-72). IOS Press.

[14]. Zuo, Z., Watson, M., Budgen, D., Hall, R., Kennelly, C. and Al Moubayed, N., 2021. Data anonymization for pervasive health care: systematic literature mapping study. JMIR medical informatics, 9(10), p.e29871.

[15]. Song, B.K., Yoo, J.S., Hong, M. and Yoon, J.W., 2019. A bitwise design and implementation for privacy-preserving data mining: from atomic operations to advanced algorithms. Security and Communication Networks, 2019, pp.1-14.

[16]. Farokhi, F. and Sandberg, H., 2019. Ensuring privacy with constrained additive noise by minimizing fisher information. Automatica, 99, pp.275-288.

[17]. Denham, B., Pears, R. and Naeem, M.A., 2020. Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining. Expert Systems with Applications, 152, p.113380.

[18]. Virupaksha, S. and Dondeti, V., 2021. Anonymized noise addition in subspaces for privacy preserved data mining in high dimensional continuous data. Peer-to-Peer Networking and Applications, 14(3), pp.1608-1628.

[19]. Fan, W., He, J., Guo, M., Li, P., Han, Z. and Wang, R., 2020. Privacy preserving classification on local differential privacy in data centers. Journal of Parallel and Distributed Computing, 135, pp.70-82.

[20]. Wu, Y., Cai, S., Xiao, X., Chen, G. and Ooi, B.C., 2020. Privacy preserving vertical federated learning for tree-based models. arXiv preprint arXiv:2008.06170.

[21]. Gheisari, M., Najafabadi, H.E., Alzubi, J.A., Gao, J., Wang, G., Abbasi, A.A. and Castiglione, A., 2021. OBPP: An ontology-based framework for privacy-preserving in IoT-based smart city. Future Generation Computer Systems, 123, pp.1-13.

[22]. Iwendi, C., Moqurrab, S.A., Anjum, A., Khan, S., Mohan, S. and Srivastava, G., 2020. N-sanitization: A semantic privacy-preserving framework for unstructured medical datasets. Computer Communications, 161, pp.160-171.

[23]. Ratra, R., Gulia, P. and Gill, N.S., 2023. Performance analysis of perturbation-based privacy preserving techniques: an experimental perspective. International Journal of Electrical & Computer Engineering (2088-8708), 13(5).

[24]. Singaravelan, S., Gopalsamy, P. and Balaganesh, S., 2021. Accumulation of data perturbation techniques for privacy preserving data classification. Asian Journal of Current Research, 6(1), pp.38-49.

[25]. Hewage, U.H.W.A., Sinha, R. and Naeem, M.A., 2023. Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review. Artificial Intelligence Review, pp.1-38.

[26]. Pinjarkar, V., Amit, J., Bhasker, B. and Srivastava, P., 1945. Pertinent Exploration of Privacy Preserving Perturbation Methods. international Journal of Recent Technology and Engineering, 8(6), p.2020.

[27]. Ito, S., Terada, M. and Kato, S., The Potential of Differential Privacy Applied to Detailed Statistical Tables Created Using Microdata from the Japanese Population Census.

[28]. Singaravelan, S., Gopalsamy, P. and Balaganesh, S., 2021. Accumulation of data perturbation techniques for privacy preserving data classification. Asian Journal of Current Research, 6(1), pp.38-49.

[29]. Anagreh, M., Vainikko, E. and Laud, P., 2021, February. Parallel Privacy-preserving Computation of Minimum Spanning Trees. In ICISSP (pp. 181-190).

[30]. Verheyen, S., White, A. and Storms, G., 2022. A comparison of the Spatial Arrangement Method and the Total-Set Pairwise Rating Method for obtaining similarity data in the conceptual domain. Multivariate Behavioral Research, 57(2-3), pp.356-384.