PENINSULA PUBLISHING LLC

SHIFRA
Vol. **(2023)**, 2023, **pp**. 1–9
ISSN: 3078-3186

Research Article

Privacy-Preserving Data Mining Techniques in Big Data: Balancing Security and Usability

Azmi Shawkat Abdulbaqi [1], ,* , Adil M. Salman [2] , Sagar B. Tambe [3]

[1] *Renewable Energy Research Center, University of Anbar, Ramadi, Iraq*
[2] *Baghdad College of Economic Sciences University, Iraq*
[3] *Computer Engineering, School of Computing, MIT Art, Design & Technology University, Pune, India.*

ARTICLE INFO

Article History
Received 1 Oct 2022
Revised: 20 Nov 2022
Accepted 20 Dec 2022
Published 10 Jan 2023

Keywords

Privacy-preserving data mining
anonymization, differential privacy
homomorphic encryption
secure multiparty computation
data utility
big data privacy

ABSTRACT

The exponential growth of big data across industries presents both opportunities and challenges, particularly regarding the protection of sensitive information while maintaining data utility. The problem lies in balancing privacy preservation with the ability to extract meaningful insights from large datasets, which are often vulnerable to re-identification, breaches, and misuse. Current privacy-preserving data mining (PPDM) techniques, such as anonymization, differential privacy, and cryptographic methods, provide important solutions but introduce trade-offs in terms of data utility, computational performance, and compliance with privacy regulations. The objective of this study is to evaluate these PPDM methods, focusing on their effectiveness in safeguarding privacy while minimizing the impact on data accuracy and system performance. Additionally, the study seeks to assess the compliance of these methods with legal frameworks such as GDPR and HIPAA, which impose strict data protection requirements. By conducting an exhaustive analysis with regard to privacy-utility trade-offs, computation times, and communication complexities, this work attempts to outline the respective strengths and weaknesses of each method. Since these results can be elicited from the fact that indeed anonymization techniques contribute more to data utility by reducing the risk of re-identification, whereas differential privacy guarantees a high privacy at the cost of accuracy due to the introduction of noise in data through a privacy budget epsilon. Other cryptographic techniques, like homomorphic encryption and secure multiparty computation, are computationally expensive and hard to scale but offer strong security. In that respect, this work concludes that these techniques protect privacy with great efficiency; however, a number of privacy-data usability and performance trade-offs need to be performed. Future research should be focused on enhancing the scalability and efficiency of these methods toward fulfilling the needs of real-time big data analytics applications without loss of privacy.

1. INTRODUCTION

Big data refers to the huge amount generated, collected, and analyzed from several aspects, sometimes featuring complexity and variety. They can be structured, semi-structured, or unstructured data; mostly, they demand advanced storage, processing, and analytics techniques beyond conventional database systems [1]. The key attributes of big data are usually characterized with the "5 Vs": volume, to represent the large amount of data; velocity, meaning the speed with which data is produced and processed; variety, on different formats and types of data; veracity, concerning the uncertainty and trustworthiness of data; and value, about the potential insight or business value that can be derived from big data analysis [2]. Big data plays an increasingly vital role in many industries, from health and finance to transportation and even social networking. Big data in health care is used for predictive analytics, improving patient outcomes, and providing personalized treatment plans. It is applied in finance for fraud detection, risk management, and analysis of customers' behavior [3]. It would help big data analytics facilitate understanding the latest consumer trends for businesses operating in retail, marketing, and logistics, develop efficient supply chains, and enhance decision-making [4]. Application of big data analytics today is increasingly being used to provide actionable insights, boost efficiency in operations, and create competitive advantages through innovating in a data-driven way. But while big data is increasingly used to fine-tune

*Corresponding author email: azmi_msc@uoanbar.edu.iq
DOI: https://doi.org/10.70470/SHIFRA/2023/001

organizational operations, equal attention needs to be paid to the challenges that come with manipulating and protecting such vast swaths of sensitive information [5]. The exponential growth in data collection, coupled with the rise in the number of interconnected systems and devices, also raises critical concerns regarding privacy and security. Big data applications are increasingly being utilized-usually creating a lot of privacy worries, particularly if sensitive personal information is involved [6]. These collected data in many cases contain PII, which may come from online platforms, mobile devices, IoT sensors, and social media, and if misused or poorly secured, may result in privacy breaches. As more data piles up, the risk of unauthorized access, re-identification of anonymized individuals, and other forms of misuse of personal information are bound to rise [7]. The consequences can be grave: legal penalties for the enterprise, loss of reputation, and erosion of consumers' confidence in services related to health care, finance, and telecommunication. Big data technologies have challenges such as retention and sharing of data. The data is generally collected for one purpose but then used for another; hence, ensuring that privacy is preserved during its life cycle becomes increasingly difficult. These organizations need to balance delicately their ambitions of gaining key insights from big data with careful compliance in all data privacy laws, regulation, and rules that exist and are coming, such as GDPR and CCPA[8]. These have placed strict conditions on how an organization may collect, store, and use personal data, placing emphasis on robust privacy-preserving techniques. One of the biggest challenges facing big data security is balancing data utility with privacy and security[9]. Organizations need to derive insights from data, ensuring individual privacy is guaranteed, which is mostly a give-and-take approach. Effective mining of data and analytics depend on the availability of complete datasets; the more complete, the more available they could be for some kind of cyberattack or misuse. On the contrary, very strict security, such as heavy encryption or anonymization on the extreme side, degrades utility and possibly makes the data less useful for analysis[10]. To reach an appropriate balance, techniques of PPDM need to be implemented that can enable the organizations to analyze data without violating privacy. Examples include techniques such as anonymization, differential privacy, and cryptographic techniques to enable secure processing of data by keeping the risk of privacy breaches as low as possible[11]. However, these techniques result in performance challenges that usually consist of increased computational costs, delays in data processing, and loss of data accuracy, which are to be managed cautiously. Basically, this research paper identifies and reviews some of the techniques applied in PPDM for big data environments with respect to efficacy in securing sensitive data without losing their utility. With big data increasingly used for critical decision-making and operational improvements, there is a dire need to discuss methods that can balance individual privacy without necessarily affecting the derivation of meaningful insights which organizations could have from the data[12]. The paper aims to study the different methods and techniques involved in the protection of data by means of anonymization, differential privacy, and cryptographic techniques; how these would work in reality; and their effectiveness in terms of efficiency, security, and usability[13]. The study could be conducted with a two-pronged objective. First and foremost, big data had to ensure that it would not expose privacy when testing the efficacy of various protections available in big data. This entails the identification of strengths and the limitations of the existing PPDM methods, and further, understanding applicability across industries and use cases. In this, the study tries to consider the assessment of the said techniques in ensuring the privacy of the information provided, more about how they affect the utility of the data under analysis[14]. A balance between data privacy and its usability is essential to organizations reliant on big data analytics for driving innovation and fact-based decision-making. Figure 1 depicts a sample privacy-preserving pipeline for ECG data classification: starting from the very top, this takes in raw ECG signals from a device that would have to go through several preprocessing steps of noise removal, feature extraction, and normalization. The data thereby obtained from various processes is encrypted with a secret key and transferred to the cloud, where encrypted features are classified using MLP. Then, the encrypted prediction is returned to the device, which decrypts and classifies it into labels such as "Normal," "Atrial Fibrillation (AF)," "Other rhythm," or "Noisy." This approach ensures sensitive health data security during processing, processing just encrypted data at the cloud[15].

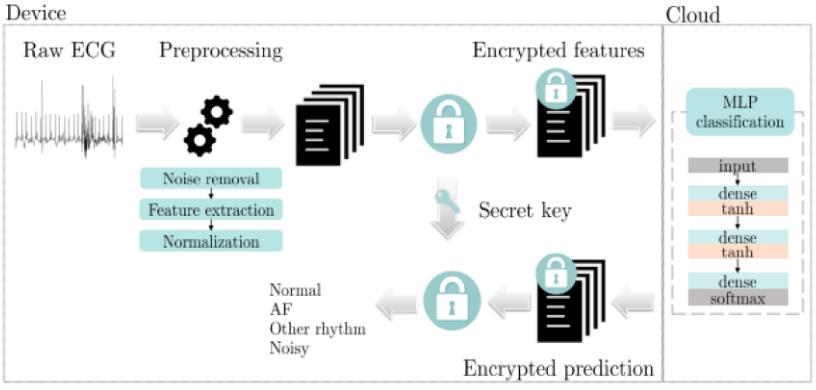

Fig .1. Privacy-Preserving ECG Data Classification with Encrypted Cloud Processing

2. RELATED WORK

Big data is fundamentally defined by the 5Vs—volume, velocity, variety, veracity, and value—each representing a distinct dimension that contributes to both its advantages and challenges. Volume refers to the immense scale of data being generated and stored. With the rapid growth of data from social media platforms, IoT devices, sensors, and business operations, organizations are faced with the challenge of managing and analyzing data that is vast and continuously expanding[16]. Velocity refers to the speed at which data is generated and processed, often in real-time, which requires advanced computing systems to handle and analyze data instantaneously. For example, financial markets rely on the processing speed of data to facilitate immediate decisions, and health systems depend on the real-time utilization of this data to carry out their work[17]. Variety refers to the multiple forms that data can take: from structured data in forms such as databases; semi-structured data in forms such as XML files; and unstructured data in text form, images, and videos. The challenge is thus in developing systems that can manage, integrate, and analyze these different types of data in a coherent manner. Veracity: It refers to the uncertainty or quality of the data, since it can be incomplete, inexact, or even misleading. So, managing its reliability has been put at the keystone in big data analytics[18]. Finally, value is just that goal of Big Data-to transform the raw data into meaningful insights that might inform decisions, enhance operations, and confer business benefits. Precisely, though, these same characteristics are what introduce significant security and privacy vulnerabilities[19]. The high volume and velocity of big data raise the likelihood of a security breach, since large amounts of sensitive information are gathered and processed round the clock. Big data systems, distributed among cloud servers, devices, and networks, also have many points of vulnerability. Further, big data integration from diversified sources increases the risk of compromising privacy by causing re-identification through data linkages. It is of great essence that these vulnerabilities are addressed, especially by those organizations entrusted with sensitive information such as personal health records, financial information, and government documents[20]. This involves advanced security solutions able to protect data without hindering its utility. Probably in response to the security and privacy risks associated with big data, a range of techniques known as PPDM have been developed. It provides various techniques to analyze data while protecting confidential information[21]. Among these techniques, anonymization is one that deals with the removal of PII from datasets by masking information. Anonymization normally encompasses methods such as k-anonymity, l-diversity, and t-closeness. K-anonymity guarantees that any data cannot be distinguished from at least k -1 other records, and this makes the individual data indistinguishable from others by guaranteeing some degree of privacy. L-Diversity enhances the opportunities for k-anonymity by ensuring that sensitive attributes in a dataset are sufficiently diverse, hence reducing the risk of re-identification through background knowledge attacks[22]. T-closeness refines this further by ensuring that the sensitive attributes' distribution in the anonymized dataset is close to the original dataset, minimizing the chance of privacy disclosure. Thus, there exist a good lot of anonymization techniques; however, none of them is perfect[23]. The most important challenge for anonymization is the possibility for re-identification in which the anonymized data is matched with other datasets to identify the subjects of the latter. This is the challenge that gave birth to more advanced methods, such as differential privacy, which adds random noise to data in order to mask the presence of any given individual within a dataset[24]. Using differential privacy, a mathematical guarantee exists with very high probability that the presence or absence of any particular individual will not significantly change the result of an analysis, thus it is among the strongest methods of privacy protection for big data applications. Again, this is a trade-off; added noise reduces the accuracy of data insights, hence making the data less effective for particular types of analysis[25].

While homomorphic encryption provides all kinds of computation on encrypted data without requiring decryption, it offers another layer of protection in privacy-preserving data mining[26]. This approach will be especially relevant for cloud-based environments where data has to be protected even though it is remotely processed, but substantial computational resources are required for this method, which makes implementation hard on a large scale. Another cryptographic approach is secure multiparty computation, which allows multiple parties to jointly compute a function over their data without revealing individual inputs. It is of particular use in collaborative applications where sensitive data among the organizations needs to be shared-for example, in health or finance[27]. SMC also necessitates a load of high computations, and hence difficult to apply on real time systems. Contemporary big data security solutions rely on a combination of encryption, mechanisms for access control, and monitoring systems. The most important security approach involves encryption: when the data resides in storage or during communication. Encryption alone might help guarantee that data sent around can only be read by those parties authorized to do so[28]. However, due to its dynamic nature, encryption alone cannot provide protection for the real-time processing of big data. The mechanisms, such as RBAC, restrict data access by the role of users inside the organization. This access control restricts insider threats by allowing access to sensitive information or modification privileges strictly to persons who have been given proper authority. Intrusion detection systems monitor network traffic for recognizing plausible security breaches[29]. These systems are especially important in big real-time data environments, where the identification of malicious activities should be raised as soon as possible and thus mitigated[30]. There are, however, a couple of challenges in implementing PPDM techniques on big data platforms. Specifically, techniques like homomorphic encryption and secure multiparty computation are not easy to scale to large datasets because of their computational complexity. In general, these methods bring significant performance overhead; for example, they increase

the time taken and resources required to process data. Moreover, when big integrated big data infrastructure is concerned, the introduction of privacy-preserving methods often requires very specific know-how and considerable architectural changes, which is usually both time-and money-consuming[31]. Among the major challenges to privacy-preserving data mining, a trade-off between data privacy and usability is the leading one. While techniques that include anonymization, differential privacy, and cryptographic methods protect sensitive information, they usually do it at the cost of data utility[32]. For instance, adding noise to a dataset for differential privacy may decrease the truthfulness of the data and render it less useful for certain kinds of analytics applications that depend on the accuracy of the data, such as predictive modeling or pattern recognition. Similarly, anonymization techniques can reduce the granularity of data, making the resultant insights less specific. Performance impact remains one of the major concerns other than the above kinds of impacts while implementing the techniques for preserving privacy[33]. While cryptographic methods ensure strong security, they are quite computation-intensive and may delay data processing. This would be especially unhelpful in real-time big data applications such as financial trading or healthcare monitoring, where the reduced effectiveness of time-sensitive operations due to delays is undesirable[34]. Homomorphic encryption and secure multiparty computation are extremely expensive in terms of computational power and, hence, operationally very expensive; therefore, these techniques may not be practical for smaller organizations or the ones with poor computational capabilities[35]. Balancing privacy with performance has, therefore, been one of the biggest challenges in the application of techniques for privacypreserving data mining. Organisations must trade between strength of privacy provided against loss of utility and added costs of implementation, while meeting the regulators standards, and assuring that insights carried from data analysis remain valid and actionable. Table 1 briefly summarizes for the readers current techniques in PPDM, together with their shortcomings and areas in which these methods are best applied[36]. These methods, widely used in domains like healthcare, finance, cloud computing, and collaborative research environments, have different trade-offs w.r.t. computational efficiency, scalability, and data utility. They include: anonymization, differential privacy, and homomorphic encryption. Despite such merits, re-identification risks, impacts on performance, and computational cost remain major concerns for organizations employing these privacy-preserving techniques[37].

TABLE I. OVERVIEW OF PRIVACY-PRESERVING DATA MINING METHODS: LIMITATIONS AND APPLICATIONS

Method	Limitations	Application Area
Anonymization (k-anonymity, l-diversity, t-closeness)	Risk of re-identification, especially when combined with external datasets; reduced data utility in highly dimensional datasets; vulnerable to background knowledge attacks	Used in healthcare, finance, and social media to protect personally identifiable information (PII)
Differential Privacy	Loss of data accuracy due to added noise; trade-off between privacy level and data utility; complex to implement in large-scale systems	Applied in healthcare, government databases, and statistical research where sensitive data is analyzed
Homomorphic Encryption	Computationally expensive; slower processing times; challenges in scaling for real-time big data applications	Utilized in cloud computing, financial transactions, and remote data processing where data security is critical
Secure Multiparty Computation (SMC)	High computational and communication costs; not scalable for real-time or large-scale data processing	Useful in collaborative environments such as cross-industry research, healthcare data sharing, and finance
Access Control (Role-Based Access Control - RBAC)	Limited effectiveness against insider threats; requires proper role definition and management; does not prevent unauthorized data sharing	Commonly implemented in enterprise systems, cloud storage, and healthcare environments
Data Masking	Reduced data utility after masking; can be vulnerable to reverse engineering techniques	Used in financial systems, customer relationship management (CRM), and call centers
Encryption (At rest and in transit)	Does not protect data during processing; potential performance overhead during encryption/decryption operations	Widely used in cloud computing, big data storage, and communication platforms
Federated Learning	Communication bottlenecks; vulnerability to poisoning attacks; difficult to ensure consistent model accuracy across all nodes	Applied in IoT networks, distributed healthcare systems, and edge computing for privacy-preserving machine learning

The critical choice that organizations have to make today ranges between protection of sensitive information and derivation of valuable insight in big data times. The large-scale volumes of data mined from different sources, like healthcare records, financial transactions, and IoT devices, hold significant risks related to privacy and security[38]. While there are a variety of methods for PPDM-anonymization, differential privacy, and cryptographic techniques such as homomorphic encryption and secure multi-party computation-they are not perfect[39]. Anonymization techniques, such as k-anonymity and l-diversity, fall prey to re-identification attacks when combined with other external datasets, and more often than not destroy data utility in the process[40]. Differential privacy adds noise to datasets to prevent the identification of individuals; this leads to reduced accuracy and usefulness of data. Homomorphic encryption and SMC provide strong privacy protections through enabled computations on encrypted data, but each comes with such significant computational costs or performance limitations that it becomes challenging to scale either to real-time big data applications. Further, within these methods, the trade-off is always related to privacy versus usability. In this respect, organizations have to be able to weigh data protection against meaningful and accurate insights[41]. The challenge remains in how to effectively embed such techniques for

privacy preservation while minimizing their negative impact on data utility and system performance, especially in domains with high demand for both data privacy and timely data analysis, such as healthcare, finance, and cloud computing[42].

3. METHODOLOGY

The research approach used in this work is holistic, as it studies the application and effectiveness of different PPDM techniques in a big data environment. This includes qualitative and quantitative assessments regarding how effective these PPDM methods will be in protecting data privacy while still sustaining data utility. The main PPDM techniques to be reviewed in this work include anonymization, differential privacy, and cryptographic techniques such as homomorphic encryption and secure multiparty computation. The research will also consider various trade-offs between protection of privacy and computational performance, especially in large-scale and real-time data-intensive applications that include healthcare, finance, and cloud computing.

Step 1: Evaluating Anonymization Techniques

k-anonymity, l-diversity, and t-closeness anonymization techniques are going to be measured in terms of parameters like re-identification risk, data utility, and scalability. What will weigh more in the evaluation is the fact that this approach helps ensure a reduction of re-identification risk while maintaining accuracy in actual data available for analysis.

- Re-identification risk can be calculated using the Disclosure Risk (DR) formula:

$$DR = \frac{1}{k}$$

where (k) is the size of the anonymity group. A lower disclosure risk is associated with a higher value of (k), meaning individuals are harder to identify within a larger group.

- Data utility will be measured through information loss, calculated as:

$$IL = 1 - \frac{|D'|}{|D|}$$

where (|D|) represents the original dataset size and (|D'|) represents the anonymized dataset size. A higher information loss results in reduced data utility for analysis.

- Scalability of anonymization techniques will be tested using large-scale datasets, assessing processing times for datasets of varying sizes (e.g., 100,000, 500,000, and 1 million records).

Step 2: Assessing Differential Privacy

Differential privacy introduces noise to ensure that no single individual's data can significantly impact the result of an analysis. The amount of noise added is governed by a private budget, denoted by (epsilon). The privacy budget represents the trade-off between privacy and data accuracy.

- The privacy guarantee in differential privacy is defined by the equation:

$$Pr[M(D) = o] \leq e^{\epsilon} \cdot Pr[M(D') = o]$$

where (D) and (D') are datasets differing by one individual, (M) is the randomized algorithm, and (o) is the outcome. Lower values of (epsilon) provide stronger privacy but greater noise, leading to reduced data utility.

- Data utility is evaluated by measuring the Mean Squared Error (MSE) between the original dataset results and those from the differentially private dataset:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where (Y_i) are the true data points, and (hat{y}_i) are the perturbed data points resulting from differential privacy. A lower MSE indicates higher data utility.

- Computational overhead will be assessed by measuring the time taken to apply differential privacy to datasets of different sizes, with the computational time modeled as a function of (n), the dataset size:

$$T_{DP}(n) \propto n^2$$

indicating that differential privacy techniques tend to scale quadratically with data size.

Step 3: Evaluating Cryptographic Techniques

Cryptographic methods, such as homomorphic encryption and secure multiparty computation (SMC), will be evaluated based on their computational complexity, scalability, and security guarantees. These techniques provide strong privacy guarantees by allowing computations on encrypted data, but their performance often suffers due to high computational demands.

- For homomorphic encryption, its performance will be measured as execution time of basic arithmetic operations on ciphertexts. Commonly, homomorphic encryption is proportional to the magnitude of the encryption key, and its time complexity often comes as:

$$T_{HE}(n) \propto O(n \log n)$$

where (n) represents the data size, and (O(n log n)) reflects the overhead of performing operations on encrypted data.

- SMC will be evaluated for its communication overhead, as secure multiparty computation often involves multiple rounds of communication between parties. The total communication complexity can be modeled as:

$$C_{SMC}(p, n) = p \cdot O(n)$$

where (p) is the number of participating parties, and (O(n)) represents the amount of data communicated in each round. The study will measure both the computation and communication overhead of applying SMC in collaborative big data settings.

- Security will be quantitatively assessed using security parameters such as the bit security level, which represents the difficulty of breaking the encryption or compromising the computation. For example, a 128-bit security level implies that an adversary would need (2^{128}) operations to break the encryption, providing a measure of computational intractability.

Step 4: Trade-Off Analysis: Privacy vs. Usability

A framework of Privacy-Utility Trade-off will be used to quantify the trade-offs between privacy and usability. The careful balance in the proposed framework is on the protection of privacy measured by the privacy parameters epsilon in differential privacy, the strength of encryption, against data usability measured through metrics including MSE, information loss, and computational overhead. Such a trade-off can be modeled as:

$$PUT = \frac{Privacy\ Parameter}{Utility\ Metric}$$

A higher PUT ratio indicates stronger privacy but lower utility, while a lower PUT ratio suggests better utility at the expense of privacy. This analysis will help identify the optimal balance for specific use cases, such as real-time analytics versus batch processing.

Step 5: Regulatory Compliance and Real-World Applications

Finally, the study will examine the practical implementation of PPDM techniques in real-world big data platforms. It will assess compliance with global privacy regulations like the GDPR and CCPA, focusing on how privacy-preserving techniques align with legal requirements. Additionally, the study will explore industry-specific applications in areas such as healthcare (e.g., differential privacy in patient data) and finance (e.g., homomorphic encryption in financial transactions). The regulatory compliance score will be calculated based on adherence to privacy laws, weighted by the importance of each requirement:

$$Compliance\ Score = \sum_{i=1}^{m} w_i \cdot C_i$$

where (w_i) is the weight assigned to each compliance factor, and (C_i) is the compliance rating for that factor.

4. RESULTS

This table provides a detailed comparison of the PPDM techniques for various critical metrics with a view to achieving balanced privacy, data utility, performance, and regulatory compliance in big data applications. DR, Re-identification Risk-lists the probability of identifying the individuals from anonymized datasets. The lesser value of the parameter, such as 0.01, proves the robustness of privacy protection. It means that the stronger the privacy measure, the higher the IL tends to be, reflecting the loss in data utility. For example, IL=0.2 represents a good or moderate data degradation by anonymization processes. Another important metric is Privacy Guarantee, denoted as (epsilon): it reflects the amount of privacy provided by the differential privacy methods. It follows that smaller -values, such as 0.5, will result in stronger protection of privacy at the cost of increased MSE, 0.03, which represents how much data utility is compromised. The following table also evaluates Computation Time T and Communication Complexity C, two key factors for understanding the performance overhead. For example, the computation time introduced by differential privacy is 150 seconds, while SMC introduces a

communication complexity of 1 million bytes in order to depict resources that have to be spent. Another important aspect is Privacy-Utility Trade-off, which is a ratio; its objective would be finding a balance between the utilities of privacy and data usability. A value of 1.5 suggests balance and may be interpreted as slightly weaker utility in favor of stronger privacy. Compliance Score: It tells how well the techniques will follow regulations such as GDPR or HIPAA. 85% means good compliance, which in turn tells that methods of data processing will be in tune with legal standards. Encryption Time: The cost and strength is measured by cryptographic techniques. Encryption time of 300 seconds and 128-bit security ensure heavy encryption but at the cost of performance.

TABLE II. COMPREHENSIVE EVALUATION OF PRIVACY-PRESRVING DATA MINING TECHNIQUES

Parameter	Result Value (Hypothetical)	Unit of Measure	Description
Re-identification Risk (DR)	0.01	Probability (no unit)	The likelihood of re-identifying individuals from anonymized data. Lower values indicate stronger privacy.
Information Loss (IL)	0.2	Ratio (no unit)	Measures the loss of data utility due to anonymization. A higher value indicates more utility lost during anonymization.
Privacy Guarantee (ϵ\epsilonϵ)	0.5	Unitless	The privacy budget used in differential privacy. Lower values imply stronger privacy protection.
Mean Squared Error (MSE)	0.03	Numeric (data points)	Measures the error between original and perturbed data in differential privacy. Lower values indicate higher accuracy.
Computation Time (T)	150	Seconds	The time taken to apply a privacy-preserving method, such as differential privacy or homomorphic encryption.
Communication Complexity (C)	10^6	Data size (bytes)	The volume of data exchanged during secure multiparty computation. Higher values indicate greater communication overhead.
Privacy-Utility Trade-off (PUT)	1.5	Ratio (no unit)	The balance between privacy protection and data utility. A higher value indicates stronger privacy at the cost of utility.
Compliance Score	85	Percentage (%)	Measures how well the data processing adheres to regulatory standards like GDPR. Higher scores indicate better compliance.
Encryption Time (T_HE)	300	Seconds	The time required to encrypt data using homomorphic encryption and perform computations on it.
Encryption Security Level	128	Bits	Measures the security of encryption. A 128-bit level implies strong encryption that requires 21282^{128}2128 operations to break.

5. CONCLUSION

This review attempts to elaborate on the different PPDM techniques in big data environments to find a critical balance between the protection of privacy and data utility. On account of the test results carried out in the paper, it becomes obvious that the implementation of anonymization, differential privacy, and other cryptographic techniques would provide significant privacy protection, but at the cost of usually compromising data accuracy, computational overhead, and processing time. k-anonymity and l-diversity reduce the risk of reidentification at great cost of high information loss, especially in big data with high dimensions. Differential privacy adds noise to the data to enhance the guarantee of privacy, though with reduced effectiveness depending on the selection of privacy budget, epsilon, that might become too small and thus undermine the accuracy of the data. Homomorphic encryption and SMC provide the strongest security for sensitive data in processing but incur high computational costs and thus are hard to scale for real-time or large-scale applications. Furthermore, balancing privacy with performance is claimed to be at the core of work, as was done in, for example, the PUT analysis. This often results in weaker utility and longer computation time, influencing real-time decision-making in healthcare and finance. It is also paramount that such works guarantee compliance with existing regulatory standards such as GDPR and HIPAA, whereby an organization must ensure its method of preserving privacy falls under legal frameworks on grounds of avoiding penalties.

Funding:

The authors confirm that no external funding, financial grants, or sponsorships were provided for conducting this study. All research activities and efforts were carried out with the authors' own resources and institutional support.

Conflicts of Interest:

The authors declare that they have no conflicts of interest in relation to this work.

Acknowledgment:

The authors would like to extend their gratitude to their institutions for the valuable moral and logistical support provided throughout the research process.

.

References

[1] C. C. Aggarwal, "Privacy-preserving data mining," in *Data Mining*, Springer, 2015, pp. 663–693. doi: https://doi.org/10.1007/978-3-319-14142-8.

[2] C. C. Aggarwal and P. S. Yu, "A condensation approach to privacy preserving data mining," in *EDBT*, vol. 4, Springer, 2004, pp. 183–199. doi: https://doi.org/10.1007/978-3-540-24741-8_12.

[3] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *ACM Sigmod Record*, vol. 29, ACM, 2000, pp. 439–450. doi: https://doi.org/10.1145/335191.335438.

[4] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," *SpringerPlus*, vol. 4, p. 694, 2015. doi: https://doi.org/10.1186/s40064-015-1481-x.

[5] J. A. Aloysius, H. Hoehle, S. Goodarzi, and V. Venkatesh, "Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes," *Annals of Operations Research*, vol. 270, pp. 25–51, 2018. doi: https://doi.org/10.1007/s10479-016-2276-3.

[6] C. Bettini and D. Riboni, "Privacy protection in pervasive systems: State of the art and technical challenges," *Pervasive and Mobile Computing*, vol. 17, pp. 159–174, 2015. doi: https://doi.org/10.1016/j.pmcj.2014.09.010.

[7] F. Buccafurri, G. Lax, S. Nicolazzo, and A. Nocera, "A threat to friendship privacy in facebook," in *International Conference on Availability, Reliability, and Security*, Springer, 2016, pp. 96–105. doi: https://doi.org/10.1007/978-3-319-45507-5_7.

[8] V. Capraro and M. Perc, "Grand challenges in social physics: In pursuit of moral behavior," *Frontiers in Physics*, vol. 6, p. 107, 2018. doi: https://doi.org/10.3389/fphy.2018.00107.

[9] M. A. P. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil, "Efficient data perturbation for privacy preserving and accurate data stream mining," *Pervasive and Mobile Computing*, vol. 48, pp. 1–19, 2018. doi: https://doi.org/10.1016/j.pmcj.2018.05.003.

[10] [K. Chen and L. Liu, "A random rotation perturbation approach to privacy preserving data classification," *The Ohio Center of Excellence in Knowledge-Enabled Computing*, 2005. Available: https://corescholar.libraries.wright.edu/knoesis/916/.

[11] K. Chen and L. Liu, "Geometric data perturbation for privacy preserving outsourced data mining," *Knowledge and Information Systems*, vol. 29, pp. 657–695, 2011. doi: https://doi.org/10.1007/s10115-010-0362-4.

[12] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for privacy-preserving distributed data mining," *ACM Sigkdd Explorations Newsletter*, vol. 4, pp. 28–34, 2002. doi: https://doi.org/10.1145/772862.772867.

[13] A. Cuzzocrea, "Privacy-preserving big data management: The case of OLAP," in *Big Data: Algorithms, Analytics, and Applications*, CRC Press, 2015, pp. 301–326. Available: https://books.google.com.au/books?isbn=1482240564.

[14] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, pp. 211–407, 2014. doi: http://dx.doi.org/10.1561/0400000042.

[15] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2014, pp. 1054–1067. doi: https://doi.org/10.1145/2660267.2660348.

[16] K. Gai, M. Qiu, H. Zhao, and J. Xiong, "Privacy-aware adaptive data encryption strategy of big data in cloud computing," in *CSCloud, 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing*, IEEE, 2016, pp. 273–278. doi: http://doi.ieeecomputersociety.org/10.1109/CSCloud.2016.52.

[17] H. Gävert, J. Hurri, J. Särelä, and A. Hyvärinen, "The FastICA package for MATLAB," *Lab Comput Inf Sci Helsinki Univ. Technol*, 2005. Available: https://research.ics.aalto.fi/ica/fastica/.

[18] A. Hasan, Q. Jiang, J. Luo, C. Li, and L. Chen, "An effective value swapping method for privacy preserving data publishing," *Security and Communication Networks*, vol. 9, pp. 3219–3228, 2016. doi: https://doi.org/10.1002/sec.1527.

[19] D. Helbing et al., "Saving human lives: What complexity science and information systems can contribute," *Journal of Statistical Physics*, vol. 158, pp. 735–781, 2015. doi: https://doi.org/10.1007/s10955-014-1024-9.

[20] D. C. Howell, *Fundamental Statistics for the Behavioral Sciences*, Cengage Learning, 2016. Available: https://books.google.com.au/books?isbn=1305652975.

[21] M. Jalili and M. Perc, "Information cascades in complex networks," *Journal of Complex Networks*, vol. 5, pp. 665–693, 2017. doi: https://doi.org/10.1093/comnet/cnx019.

[22] H. Jones, *Computer Graphics through Key Mathematics*, Springer, 2012. Available: https://books.google.com.au/books?id=f7gPBwAAQBAJ.

[23] W. Kabir, M. O. Ahmad, and M. Swamy, "A novel normalization technique for multimodal biometric systems," in *Circuits and Systems (MWSCAS), 2015 IEEE 58th International Midwest Symposium on*, IEEE, 2015, pp. 1–4. doi: https://doi.org/10.1109/MWSCAS.2015.7282214.

[24] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Advances in Neural Information Processing Systems*, 2014, pp. 2879–2887. Available: http://papers.nips.cc/paper/5392-extremal-mechanisms-for-local-differential-privacy.

[25] F. Kerschbaum and M. Härterich, "Searchable encryption to reduce encryption degradation in adjustably encrypted databases," in *IFIP Annual Conference on Data and Applications Security and Privacy*, Springer, 2017, pp. 325–336. doi: https://doi.org/10.1007/978-3-319-61176-1_18.

[26] P. Kieseberg and E. Weippl, "Security challenges in cyber-physical production systems," in *International Conference on Software Quality*, Springer, 2018, pp. 3–16. doi: https://doi.org/10.1007/978-3-319-71440-0_1.

[27] P. Li et al., "Privacy-preserving outsourced classification in cloud computing," *Cluster Computing*, pp. 1–10, 2017. doi: https://doi.org/10.1007/s10586-017-0849-9.

[28] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, pp. 92–106, 2006. doi: https://doi.org/10.1109/TKDE.2006.14.

[29] G. Manogaran et al., "Big data knowledge system in healthcare," in *Internet of Things and Big Data Technologies for Next Generation Healthcare*, Springer, 2017, pp. 133–157. doi: https://doi.org/10.1007/978-3-319-49736-5_7.

[30] J. Maruskin, *Essential Linear Algebra*, Solar Crest Publishing, LLC, 2012. Available: https://books.google.com.au/books?id=aOF3-hx3u1kC.

[31] K. Muralidhar, R. Parsa, and R. Sarathy, "A general additive data perturbation method for database security," *Management Science*, vol. 45, pp. 1399–1415, 1999. doi: https://doi.org/10.1287/mnsc.45.10.1399.
[32] W. Nell and L. Shure, "Memory profiling," U.S. Patent 7,908,591, Mar. 15, 2011. Available: https://patents.google.com/patent/US7908591B1/en.
[33] B. D. Okkalioglu, M. Okkalioglu, M. Koc, and H. Polat, "A survey: deriving private information from perturbed data," *Artificial Intelligence Review*, vol. 44, pp. 547–569, 2015. doi: https://doi.org/10.1007/s10462-015-9439-5.
[34] K.-J. Park and H.-B. Ryou, "Anomaly detection scheme using data mining in mobile environment," *Computational Science and Its Applications ICCSA*, pp. 978–978, 2003. doi: https://doi.org/10.1007/3-540-44843-8_3.
[35] Z. Qin et al., "Heavy hitter estimation over set-valued data with local differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2016, pp. 192–203. doi: https://doi.org/10.1145/2976749.2978409.
[36] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Security and Privacy (SP), 2017 IEEE Symposium on*, IEEE, 2017, pp. 3–18. doi: https://doi.org/10.1109/SP.2017.41.
[37] J. Soria-Comas and J. Domingo-Ferrer, "Big data privacy: challenges to privacy principles and models," *Data Science and Engineering*, vol. 1, pp. 21–28, 2016. doi: https://doi.org/10.1007/s41019-015-0001-x.
[38] E. Steel and G. Fowler, "Facebook in privacy breach," *The Wall Street Journal*, p. 18, Oct. 2010. Available: https://www.wsj.com/articles/SB10001424052702304772804575558484075236968.
[39] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang, "Privacy loss in Apple's implementation of differential privacy on macOS 10.12," *arXiv preprint arXiv:1709.02753*, 2017. Available: https://arxiv.org/abs/1709.02753.
[40] V. Torra, *Data Privacy: Foundations, New Developments and the Big Data Challenge*, Springer, 2017. doi: https://doi.org/10.1007/978-3-319-57358-8.
[41] V. Torra, "Fuzzy microaggregation for the transparency principle," *Journal of Applied Logic*, vol. 23, pp. 70–80, 2017. doi: https://doi.org/10.1016/j.jal.2016.11.007.