International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Evaluating Privacy-Preserving Strategies via Perturbation based Data Mining Using Diverse Noise Techniques

[1]Ranjeet Kumar Rai, [2]Dr. Manish Varshney

Submitted: 13/12/2023 **Revised**: 28/01/2024 **Accepted**: 02/02/2024

Abstract: Knowledge discovery from data, commonly referred to as data mining. it involves the extraction of significant information, which may be previously unknown, concealed, or relevant, from extensive data sets or databases through the utilization of statistical methodologies. With the introduction of enhanced hardware technologies, there has been a proliferation in the storage and recording of personal data pertaining to individuals. Sophisticated organizations employ data mining algorithms to uncover hidden patterns or insights within data. Data mining techniques find application in diverse fields such as marketing, medical diagnosis, forecasting system, and national security. However, in scenarios where data privacy is paramount, mining certain types of data without violating the privacy of data owners presents a formidable challenge, sparking growing concerns among privacy advocates. To address these concerns, it is imperative to advance data mining procedures that are complex to individual privacy considerations. Perturbation of data plays a pivotal role in Privacy-Preserving Data Mining (PPDM). Additive data safeguard data privacy. In contrast, multiplicative data perturbation involves a series of transformations, including rotation, translation, and the addition of noise components to the perturbed data copy.

Keywords: *Data mining, Forecasting, Machine Learning, Cryptographic, Dataset.*

1. Introduction

Knowledge discovery from data, commonly referred to as data mining. It is the procedure of obtaining important information., often previously concealed or relevant, from extensive data sets or databases through the application of statistical techniques. Advancements in hardware technology have led to the increased storage and recording of individuals' personal data, raising concerns about potential misuse for invasive or malicious purposes. To address these concerns, privacy-preserving data mining has emerged as a vital approach. It seeks to accomplish data mining goals without compromising people's privacy or disclosing their underlying data values. Within this realm, the field of Protecting Personal Data through Data Mining (PPDM) plays a pivotal role in safeguarding sensitive information from unintentional or informal disclosure.

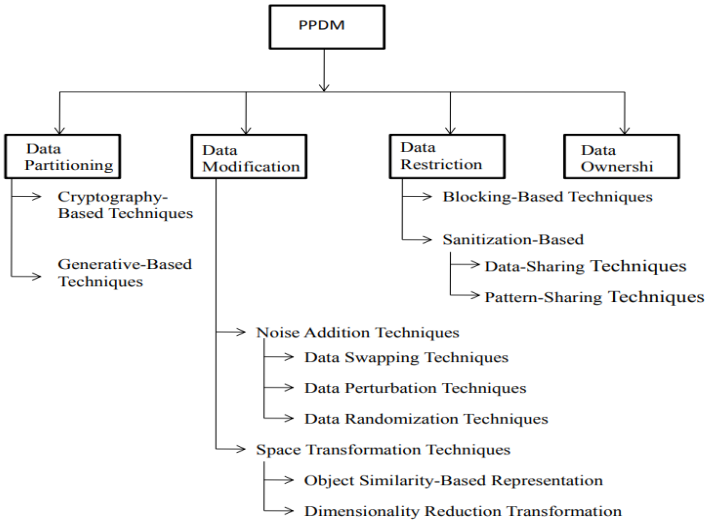


Fig 1: Privacy-preserving data mining technique

[1]*Research Scholar, Department of Computer Science & Engg. MUIT, Lucknow*
Enail Id-ranjeetrai2007@gmail.com
[2]*Prof. School of Engineering & Technology, MUIT, Lucknow*

2. Literature Survey

Exploring Privacy-Preserving Methods via Perturbation Data Mining employing diverse noise strategies encompasses a comprehensive investigation into safeguarding sensitive information during data analysis. This research delves into the realm of data privacy by employing a wide array of noise-based techniques, including additive Gaussian noise, multiplicative perturbations, and various other noise addition methods. By adopting this holistic approach, the study aims to strike a balance between preserving individual privacy and maintaining the utility of the data, offering a multifaceted perspective on the effectiveness and practicality of different privacy-preserving strategies. Through this exploration, it seeks to provide valuable insights into the nuanced challenges and opportunities in the area of data mining that protects privacy, catering to the evolving demands of data security and analytics in an increasingly data-centric world. Table 1 provides Comparative analysis previous work done.

In [1] Nathiya's study (2015), the focus was on enhancing data privacy in data mining. The research presented a multi-security approach to privacy preservation, aiming to protect sensitive information while allowing for meaningful data analysis. This work contributes to the field by addressing the challenges of privacy-preserving data mining through a comprehensive multi-security framework. In [2] Kalaivani and Subbiah (2014) The study provides valuable insights into the application of noise-based techniques in preserving data privacy.

In [3] Kamaleswari (2014) proposed techniques to protect data from sophisticated attacks, contributing to the development of robust privacy-preserving data mining methods that can withstand various types of threats. The exploration of noise-based techniques for protecting sensitive data while ensuring meaningful data analysis.

In [4] Liu's research (2012) introduced two techniques for adding noise to data mining while maintaining privacy. This work examined different noise injection strategies and their effectiveness in maintaining privacy while allowing for meaningful data analysis. The study offers valuable insights into the practical implementation of privacy-preserving techniques.

In [5] Patil's study (2012) focused on reliability based on perturbations and data authentication maintenance mining. The research investigated methods to ensure data integrity and authentication while preserving privacy. This work contributes to the broader understanding of data security in the context of data mining.

In [6] Li, Chen, Li, and Zhang (2011) presented research on enabling multi-level trust in privacy-preserving data mining. Their work aimed to enhance the trustworthiness of privacy-preserving methods, addressing the challenge of balancing data privacy and utility in multi-level trust scenarios.

In [7] Chen and Liu (2008) conducted a comprehensive survey of multiplicative perturbation techniques for privacy-preserving data mining. This survey provides an overview of various perturbation methods and their applicability, serving as a valuable resource for researchers and practitioners in the field.

In [8] Sriharsha and Parthasarathy (2015) explored perturbing sensitive data using additive noise. Their research investigated the impact of additive noise on data privacy and utility, offering insights into an essential privacy-preserving technique.

In [9] Yang, Huo, Yang, Wang, Hu, and Liu (2012) introduced two noise addition methods for privacy-preserving data mining. This research examined noise-based approaches for data privacy protection, providing valuable insights into the practical implementation of privacy-preserving techniques.

In [10]Wilson and Rosen (2003) investigated data protection through perturbation techniques and its impact on knowledge discovery in databases. Their study shed light on the trade-offs between data privacy and data utility in the context of perturbation-based privacy preservation.

In [11] Al-Ahmadi, Rosen, and Wilson (2008) focused on data mining performance on perturbed databases and its influence on classification accuracy. This research examined the practical implications mining algorithms, contributing to our understanding of the trade-offs involved in privacy-preserving data mining.

Table I: Comparative analysis of previous research done by Authors

Authors	Techniques Used	Key Features	Limitations
Nathiya (2015)	- Secure data perturbation - Privacy preservation techniques	Multi-security in privacy-preserving data mining	- May not address all types of attacks - Scalability issues - Computational overhead

Kalaivani & Subbiah (2014)	- Additive noise injection - Multi-level trust privacy preservation	Additive Gaussian noise-based data perturbation	- Sensitivity to noise parameters - Impact on data utility - Complexity of trust levels
Kamaleswari S (2014)	- Non-linear attack detection - Trust-based privacy preservation	Handling non-linear attacks preserving data mining	- Limited focus on non-linear attacks - Lack of real-world data validation
Likun Liu (2012)	- Noise addition techniques - Privacy preservation mechanisms	Two noise addition methods for privacy-preserving data mining	- Trade-off between privacy and data utility - Lack of robustness to certain attacks
Patil Dnyanesh (2012)	- Data perturbation for reliability and authentication	Perturbation-based reliability and data mining authentication	- May not address advanced privacy threats - Limited focus on scalability
Yaping Li et al. (2011)	- Multi-level trust model - Privacy preservation methods	Reliability based on perturbations and preserving authenticity in data mining	- Complexity in trust modelling - Potential trust-based conflicts
Chen & Liu (2008)	- Multiplicative noise techniques - Privacy preservation overview	An investigation of multiplicative perturbation for data mining with privacy protection	- Limited detail on specific techniques - Aging techniques over time
A V Sriharsha et al. (2015)	- Additive noise perturbation - Privacy-preserving methods	Perturbing sensitive data using additive noise	- Potential sensitivity to noise levels - Computational overhead
Yang et al. (2012)	- Noise addition techniques - Privacy preservation mechanisms	Two techniques for adding noise to data mining while maintaining privacy	- Potential trade-off between privacy and data utility - Lack of robustness
Wilson & Rosen (2003)	- Data perturbation for privacy preservation	Protecting data through perturbation techniques	- Limited focus on specific privacy threats - Data utility concerns
Al-Ahmadi et al. (2008)	- Impact analysis of data perturbation	Performance of data mining on disturbed databases	- Limited focus on privacy preservation techniques - Dependency on specific data mining algorithms

3. Effects on the Practice of Data Mining:

However, safeguarding the integrity of deeper insights within a perturbed database requires more than just preserving basic statistical metrics such as means, variances, and covariances [13]. Data mining tools, meticulously engineered to unveil concealed patterns within databases, hold a pivotal role in furnishing decision-makers with fresh perspectives on the database and its underlying content. Regrettably, there exists a conspicuous dearth of systematic research that explores how data protection methods impact these tools' ability to uncover such invaluable knowledge [14]. This represents a noteworthy void in the realm of data security literature.

Prior to 2011, only a solitary study had ventured into assessing the ramifications of data perturbation on data mining tools [15]. This study, conducted using the renowned IRIS and BUPA Liver datasets, yielded inconclusive results regarding its influence on classification accuracy. Notably, the study underscored the contrast between the two datasets employed: the IRIS dataset, celebrated for its linear separability, and the complex-to-classify BUPA Liver dataset. Most importantly, the study brought to light a significant discovery [16].

4. Ensuring Data Privacy Through the Utilization of Data Perturbation Methods

Today, organizations are amassing extensive data on various aspects of their operations, including organizational structure, human resources, and workflow. However, these organizations often struggle to unlock the full potential of this data in terms of extracting valuable information or knowledge [17]. To address this challenge, sophisticated organizations employ data mining algorithms designed to unearth hidden patterns and insights within their data. While these algorithms offer substantial benefits, they also raise concerns, particularly in cases where they could potentially be used to access confidential database [18].

As a result, database administrators are tasked with the crucial responsibility of safeguarding individuals' confidential information stored within the organizational database to prevent unauthorized disclosure. To address these privacy concerns, several techniques have emerged as commonly used applies for shielding privacy in the cloud [19].

A. Reconstruction

Recovery serves as a widely utilized technique within Privacy-Preserving Data Mining, offering several advantages. To obfuscate or transform sensitive data, supplementary data must be incorporated alongside the initial dataset.

B. Anonymization:

In the Anonymization Method detailed below, we employ suppression and generalization techniques to obfuscate the distinct characteristics of individual records. For this purpose, the widely adopted K-anonymity algorithm is employed, effectively concealing the records typically used as unique identifiers in the data [20].

C. Cryptographic:

This approach finds its primary application when multiple parties are simultaneously engaged in data mining on the same dataset. In such scenarios, safeguarding the privacy of each party's data mining activities becomes imperative. Various algorithms have been devised to address these privacy concerns in the context of account the sensitivity of privacy concerns [21].

5. Data Mining with Privacy Preserving Techniques

Data mining techniques find applications across various domains, including marketing, medical diagnosis, forecasting, and national security, enabling knowledge extraction from datasets. However, safeguarding individual privacy while mining certain data types remains a complex challenge. Many organizations collect personal information for their purposes, necessitating precautions to avoid privacy infringements or the exposure of sensitive business data [22]. This calls for the implementation of Privacy-Preserving Data Mining (PPDM) techniques, which have been a focus of research in both public and private sectors. PPDM encompasses data mining strategies designed to shield sensitive information from unauthorized or unintended disclosure. Traditional data mining methods statistically analyze and model datasets collectively, whereas privacy preservation aims to prevent the exposure of individual data records by third parties. The clear distinction between these domains underscores the technical feasibility of PPDM.

Recent concerns have emerged regarding privacy implications tied to data collection and monitoring through data mining technology, particularly in commercial and security applications. PPDM algorithms are developed to extract valuable insights from vast datasets while simultaneously safeguarding sensitive data. Establishing evaluation criteria and creating relevant benchmarks play pivotal roles in designing such algorithms.

6. Data Perturbation Method Based on Different Noise

Numerous data-mining applications involve handling personally identifiable information, such as financial transactions, healthcare records, and network communication data, leading to heightened privacy concerns. Consequently, there has been a push to develop privacy-aware data-mining techniques. These advanced algorithms aim to extract meaningful patterns for the mining process doesn't reveal enough information to reconstruct the sensitive information. Patterns can be computed at individual nodes, with minimal information exchange among participating nodes, avoiding the transmission of raw data. This has given rise to various privacy-sensitive distributed algorithms. For instance, the JAM structure, grounded in Meta knowledge, was developed for mining sensitive data distributed across multiple parties, such as in economic scam discovery.

Additionally, approaches like Fourier spectrum-based decision tree representation and collective hierarchical clustering can be adapted for privacy-preserving mining in distributed data settings. Recent research has introduced several distributed techniques for mining multiparty data, including privacy-preserving decision tree construction, multiparty secure computation frameworks.

7. Laplace and Gaussian Noise

Differential privacy is a fundamental concept that places constraints on the amount of information that can be exposed through an individual's participation in a database. It is characterized by two key parameters, epsilon (ε) and delta (δ), which define these bounds. Specifically, we focus on the multiplicative bound represented by exp(ε), where ε approximates the potential amount of information an analyst could learn about an individual. It's important to note that our information measure employs natural logarithms, which means it's in nats rather than bits.

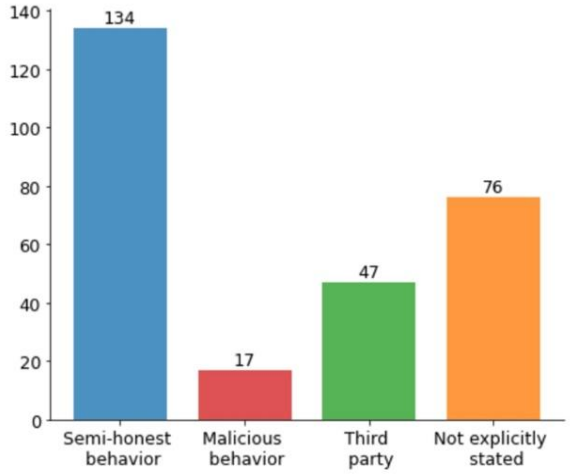
To achieve differential privacy, we introduce the term δ into the multiplicative bound equation. Ideally, δ would be zero, reflecting a preference for (ε, 0)-differential privacy. However, in practical situations, we may need to settle for (ε, δ)-differential privacy. Typically, δ accounts for the possibility that a small subset of individuals might experience greater privacy loss than the majority, making the multiplicative bound not universally applicable. When δ is extremely small, the associated risk becomes negligible.

The Laplace distribution, often referred to as the double exponential distribution, stands out due to its distribution function resembling the exponential distribution but mirrored about the y-axis. Now, why is this distribution noteworthy in our context? Well, considering our focus on multiplicative bounds, it's no surprise that exponential distributions offer valuable insights for our calculations. The exponential scale, especially when it comes to multiplicative scaling, aligns with our objectives.

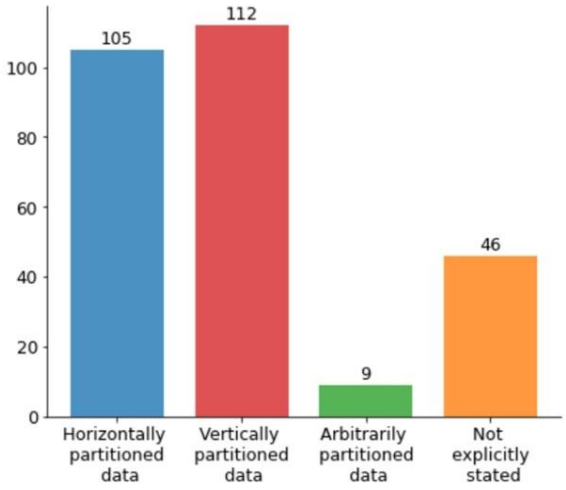
In the realm of privacy, the Laplace mechanism plays a pivotal role by introducing noise that follows a Laplacian distribution. By adding Laplace noise with scales $\Delta f/\varepsilon$, we achieve the preservation of (ε, 0)-differential privacy, where Δ represents the sensitivity of a function To be precise, we denote this sensitivity as fΔ, which corresponds to the l1 sensitivity. It's worth noting that because results with Gaussian noise are l2 sensitive, this distinction becomes essential

8. Ten Criteria Influencing Data Mining

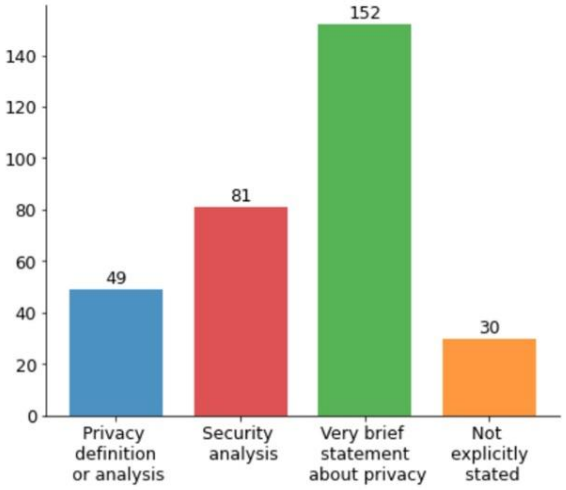
Using the 10 assessment criteria we previously established, we present a summary of the review findings for 231 publications in Fig. 5. The data repository at https://doi.org/10.6084/m9.figshare.14239937.v4 (DOI: 10.6084/m9.figshare.14239937) contains the complete review findings for 231 publications. The review findings for each element are further discussed in the part that follows.



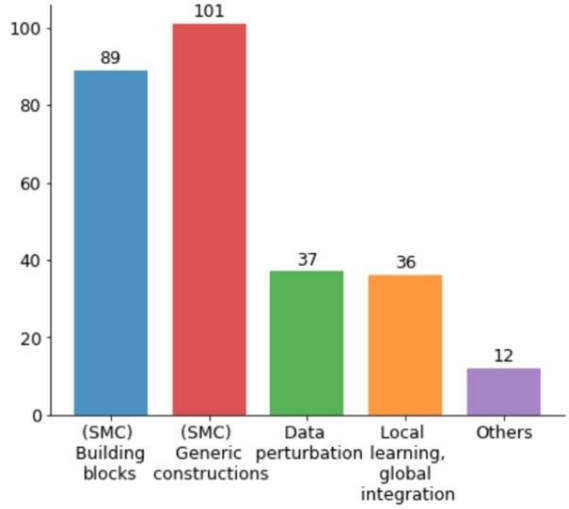
(a) Adversarial behavior of data parties



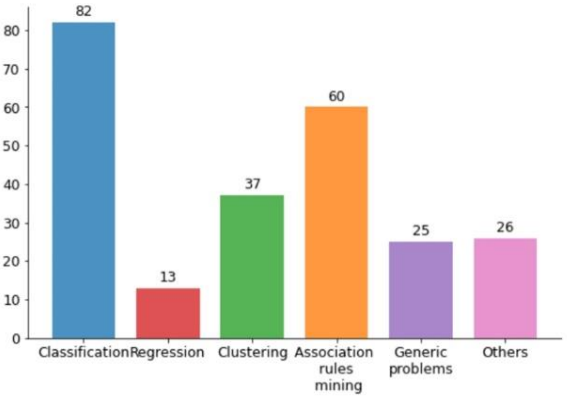
(b) Data partitioning



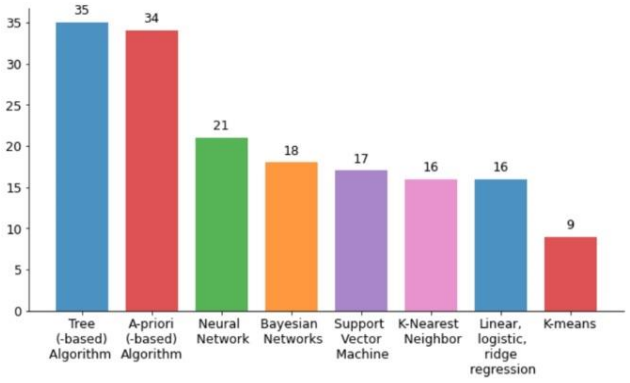
(c) Privacy definition or analysis



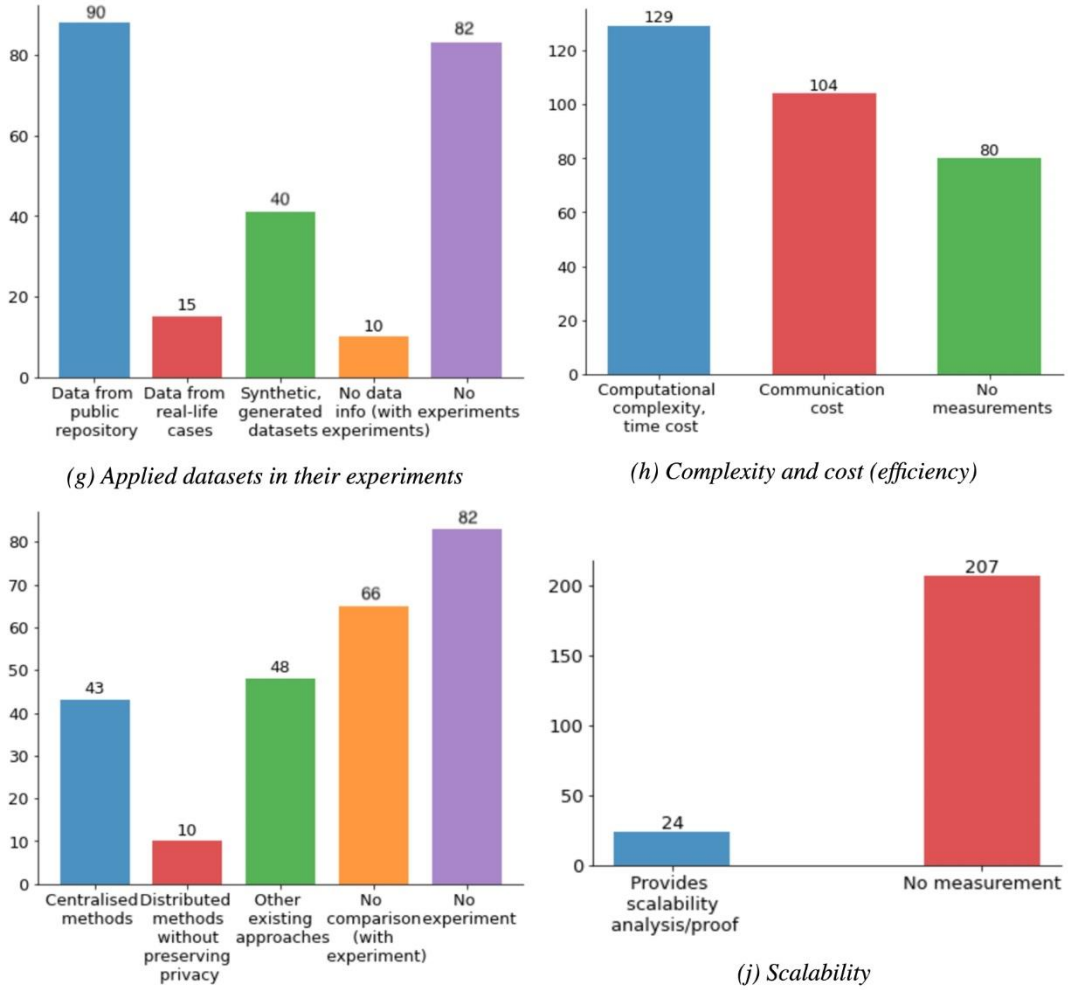
(d) Privacy-preserving methods



(e) Types of data problems



(f) Data mining algorithm



(g) Applied datasets in their experiments

(h) Complexity and cost (efficiency)

(i) Performance measures (compare with)

(j) Scalability

Fig 4: Using bar charts, review findings are shown along with the 10-factor assessment criteria. With the exception of privacy definition/analysis and scalability, papers may address one or more of the variables.

9. Conclusion

Data perturbation plays a vital role in Privacy-Preserving Data Mining (PPDM). It ensures data privacy by introducing noise into sensitive data, and there are two primary methods: additive and multiplicative data perturbation. In additive data perturbation, noise is incorporated into the sensitive data to protect privacy. This article delves into the examination of privacy and data mining utility using various noise techniques with perturbed data copies. Furthermore, the shortcomings of previous research pushed us to perform a full evaluation of the challenges associated with scattered and public data for sharing and mining. As a result, in the context of PPDM, the cost of computing for global mining, preserving the privacy of growing data, and maintaining the integrity of mining outcomes, data usefulness, scalability, and overhead performance are explored. To address these difficulties, it is critical to design a solid, efficient, and scalable model. In this regard, we found current literature gaps and shortcomings and assessed them for future major improvements, more effective privacy protection, and preservation. This comprehensive and informative review article is intended to act as a taxonomy for navigating and interpreting PPDM research advances.

References

[1] s.Nathiya (2015)" Providing Multi Security In Privacy Preserving Data Mining" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume – 4 Issue - 12 December, 2015 Page No. 15392-15396

[2] R, Kalaivani & Subbiah, Chidambaram. (2014). Additive Gaussian Noise Based Data Perturbation in Multi-Level Trust Privacy Preserving Data Mining. International Journal of Data Mining & Knowledge Management Process. 4. 21-29. 10.5121/ijdkp.2014.4303.

[3] Kamaleswari S (2014)" Handling Non-Linear Attacks in Multilevel Trust Privacy Preserving Data Mining" International Journal of Computer Science

and Information Technologies, Vol. 5 (2) , 2014, 1825-1827

[4] Likun Liu (2012)" Two Noise Addition Methods For Privacy-Preserving Data Mining" I.J. Wireless and Microwave Technologies, 2012, 3, 28-33

[5] Patil Dnyanesh (2012)" PERTURBATION BASED RELIABILITY AND MAINTAINING AUTHENTICATION IN DATA MINING" International Conference on Advances in Computer and Electrical Engineering (ICACEE'2012) Nov. 17-18

[6] Yaping Li, Minghua Chen, Qiwei Li and Wei Zhang (2011)" Enabling Multi-level Trust in Privacy Preserving Data Mining"

[7] Chen, Keke & Liu, Ling. (2008). A Survey of Multiplicative Perturbation for Privacy-Preserving Data Mining. 10.1007/978-0-387-70992-5_7.

[8] A V Sriharsha, A V Sriharsha & Parthasarathy, C.. (2015). Perturbing sensitive data using additive noise. International Journal of Applied Engineering Research. 10. 38296-38301.

[9] Yang, Kexin & Huo, Yanmei & Yang, Lei & Wang, Di & Hu, Liang & Liu, Likun. (2012). Two Noise Addition Methods For Privacy-Preserving Data Mining. International Journal of Wireless and Microwave Technologies. 2. 10.5815/ijwmt.2012.03.05.

[10] Wilson, Rick & Rosen, Peter. (2003). Protecting Data through Perturbation Techniques: The Impact on Knowledge Discovery in Databases.. J. Database Manag.. 14. 14-26. 10.4018/978-1-59140-471-2.ch003.

[11] Al-Ahmadi, Mohammad & Rosen, Peter & Wilson, Rick. (2008). Data mining performance on perturbed databases: important influences on classification accuracy. International Journal of Information and Computer Security. 2. 10.1504/IJICS.2008.016822.

[12] Mall, Pawan Kumar, et al. "Rank Based Two Stage Semi-Supervised Deep Learning Model for X-Ray Images Classification: AN APPROACH TOWARD TAGGING UNLABELED MEDICAL DATASET." Journal of Scientific & Industrial Research (JSIR) 82.08 (2023): 818-830.

[13] Narayan, Vipul, et al. "Enhance-Net: An Approach to Boost the Performance of Deep Learning Model Based on Real-Time Medical Images." Journal of Sensors 2023 (2023).

[14] Narayan, Vipul, A. K. Daniel, and Pooja Chaturvedi. "E-FEERP: Enhanced Fuzzy based Energy Efficient Routing Protocol for Wireless Sensor Network." Wireless Personal Communications (2023): 1-28.

[15] Mall, Pawan Kumar, et al. "FuzzyNet-Based Modelling Smart Traffic System in Smart Cities Using Deep Learning Models." Handbook of Research on Data-Driven Mathematical Modeling in Smart Cities. IGI Global, 2023. 76-95.

[16] Narayan, Vipul, and A. K. Daniel. "A novel approach for cluster head selection using trust function in WSN." Scalable Computing: Practice and Experience 22.1 (2021): 1-13.

[17] Narayan, Vipul, and A. K. Daniel. "Energy Efficient Protocol for Lifetime Prediction of Wireless Sensor Network using Multivariate Polynomial Regression Model." (2022).

[18] Narayan, Vipul, and A. K. Daniel. "CHHP: coverage optimization and hole healing protocol using sleep and wake-up concept for wireless sensor network." International Journal of System Assurance Engineering and Management 13.Suppl 1 (2022): 546-556.

[19] Roy, S., Roy, S., Sharma, M., Mishra, V., Rashtrapal, O., & Mall, P. K. A COMPARATIVE ANALYSIS of DEEP LEARNING MODELS FOR AIR QUALITY INDEX PREDICTION.

[20] Chourase, I., & Mall, P. K. Forest Fire and Smoke Detection Using Ensemble. Learning technique with Deep Learning neural Networks.

[21] Mall, P.K., Shukla, A., Singh, J. (2023). An Approach Towards Early Stage Detection of Lung Cancer Using Machine Learning. In: Singh, S.N., Mahanta, S., Singh, Y.J. (eds) Proceedings of the NIELIT's International Conference on Communication, Electronics and Digital Technology. NICE-DT 2023. Lecture Notes in Networks and Systems, vol 676. Springer, Singapore. https://doi.org/10.1007/978-981-99-1699-3_37

[22] Mall, P.K., Mishra, A., Sinha, A. (2023). Comparative Analysis of Anomaly-Based Intrusion Detection System on Artificial Intelligence. In: Singh, S.N., Mahanta, S., Singh, Y.J. (eds) Proceedings of the NIELIT's International Conference on Communication, Electronics and Digital Technology. NICE-DT 2023. Lecture Notes in Networks and Systems, vol 676. Springer, Singapore. https://doi.org/10.1007/978-981-99-1699-3_12