OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21600799

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Review Article

Volume-06|Issue-04|2026

A Review on Privacy Preserving Data Analysis

Amitesh Bhaskar[1], Aditya Girdhar[2], Abhijith[3], Ankit Kumawat[4] & Dr. Shashidhar V[5]

[1,2,3,4]UG Students, Computer Science and Engineering, RV Institute of Technology and Management (RVITM), Bangalore, Karnataka, India
[5]Assistant Professor, Computer Science and Engineering, RV Institute of Technology and Management (RVITM), Bangalore, Karnataka, India

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Bhaskar, A., Girdhar, A., Abhijith., Kumawat, A. & Shashidhar, V. (2026). A Review on Privacy Preserving Data Analysis. *Indiana Journal of Multidisciplinary Research, 6*(4), 173-176.

Abstract: In today's data-driven environment, it has become a paramount challenge to ensure personal privacy while aggregating and mining data. With multiple industries healthcare, IoT networks, smart grids, and vehicular networks depending on the sharing and analysis of enormous amounts of sensitive information, the threat of privacy violations continues to increase. Conventional privacy preserving methods, especially those based on additive noise or data perturbation, tend to fail in guarding against infer- ence attacks and preserving data utility. This paper provides an in-depth study and deployment of cutting-edge privacy-preserving data mining and aggregation methods that aim to balance security and usability. An extensive literature review identifies the limitations of traditional approaches and describes the advancement of contemporary techniques in coping with high dimensionality, correlated data, and estimation vulnerabilities. The approach focuses on systematic data conversion and safe aggregation protocols that are resistant to reconstruction or linking attacks.

Keywords: Privacy Preserving, Data Privacy, Data Aggregation, Data Perturbation

INTRODUCTION

With the rapid growth of of cross-connecting technologies and the growing dependence upon information-based systems, privacy safeguards have become arguably the most prominent issue in today's computing. In areas ranging from intelligent health, vehicular fog computing, intelligent grids, and IoT domains, personal and confidential user data is continuously acquired, processed, and exchanged for facilitating predictive analytics, automation, and real-time decision-making [1][2].

To overcome this problem, Privacy-Preserving Data Mining (PPDM) and Privacy-Preserving Aggregation (PPAgg) methods have been proposed. These methods enable institutions and devices to derive meaningful insights from distributed data sets without disclosing raw, identifiable information [3][4]. Methods like anonymized noise addi- tion have proven to be effective in preserving data confidentiality. However, challenges remain—like finding optimal levels of noise to maintain both privacy and accuracy [5].

PRIVACY PRESERVING DATA ANALYSIS METHODS

Record Linkage

The IPUMS MLP project created a new two-step probabilistic record linkage approach. Step I identifies the linking of men with a large set of individual, household, and context variables using a trained logistic regression model, reaching a 46.3% linkage rate and 97.7% matching agreement with the FamilySearch database. Step II utilizes the initial links for matching remaining household members by reducing the search universe and loosening matching requirements, achieving over 150,000 total links with 98.0% accuracy [5].

Privacy-Preserving Record Linkage (PPRL) provides safe matching of records between databases without compromising sensitive personal identifiers. PPRL has, over time, migrated from simple exact matching to sophisticated, scalable approaches with Bloom filters, fuzzy matching, and differential privacy. Emerging solutions intend to integrate machine learning to enhance efficiency and accuracy [6].

The TIME study, a large clinical trial on the timing of antihypertensive medication, incorporated a technique known as record linkage. This method was essential for identifying and corroborating the study's primary composite endpoint (vascular death or hospitalization for non-fatal myocardial infarction or non-fatal stroke) and selected secondary outcomes [12].

Attribute Linkage

Differential Privacy (DP) is highly effective at protecting individual privacy by ensur- ing that the release of data does not reveal much about any single person. DP alone may not prevent certain types of attacks—such as the attribute linkage attack—where adver- saries use

***Corresponding Author:** Ankit Kumawat

background knowledge to infer private information. This study explores how such vulnerabilities persist even with DP in place and proposes a stronger privacy model called APL-Free-DP to address the issue. The authors also introduce an algo- rithm named APLKiller, which enforces this model by removing Attribute Privacy Leakages (APLs) while preserving the utility of the dataset [7].

A new way of seeing how product design relates to what customers really desire, through a quality-delivery system (QDS) based on the principles of means–end chain (MEC) theory. The QDS model assists marketers in determining which particular prod- uct attributes customers truly care about, and how these attributes relate to their under- lying values and preferences [8]. To better understand the relationship between product attributes and consumer values, this study adopts an enhanced Means-End Chain (MEC) framework, incorporating statistical techniques such as factor analysis and lin- ear regression to identify attribute–consequence–value (A-C-V) linkages [14].

Privacy-Preserving Aggregation

Privacy-Preserving Federated Learning (PPFL) has been the focus of much attention because of growing issues on data privacy in Federated Learning (FL) frameworks. In FL, users train models locally, and a server aggregates them to create a global model. Yet, partial model knowledge may leak personal user information, which makes Pri- vacy-Preserving Aggregation (PPAgg) protocols necessary [9].

Privacy-Preserving Machine Learning (PPML) allows organizations to collectively en- hance the performance of ML while respecting data privacy laws. The scheme manages to cope efficiently with dropouts, decrease communication cost, and achieve up to 6.37x speedup compared to the state of the art. It maintains security against semi-honest as well as malicious adversaries based on signatures and consistency checks without any complexity to implementation and use in real-world environments [10].

Differential privacy

Differential Privacy (DP) is widely applied to secure sensitive information but tends to fail with too much noise from data correlation, particularly in multiparty environments. To overcome this, MP-CRDP (Multiparty Correlated Differential Privacy) proposes a new method that diminishes correlated sensitivity and dataset complexity in two steps, improving data utility. MP-CRDP strengthens privacy through the introduction of noise to machine learning model parameters and query outputs, with functionality for private data release and model training [11]. The sources address the rising uptake of deep learning (DL) and the escalating concerns over its vulnerabilities in privacy, security, robustness, and others. One of the main countermeasures is the introduction of random perturbation or calibrated artificial noise during DL training or prediction. Differential privacy (DP), which was initially a data publishing paradigm, has also become a rich tool to formalize defenses by measuring the required noise [12].

Privacy Preserving Data Publication

The privacy-preserving group data sharing scheme secures cloud-stored data using proxy re-encryption and an OCLT-ORAM structure to ensure confidentiality, fine- grained access control, and hidden access patterns. It features distinct phases for au- thentication, secure key exchange, and also user revocation. The system includes users, a trusted proxy, and a semi-trusted cloud server [19]. MedShare is a decentralized sys- tem for secure Electronic Health Record (EHR) sharing using blockchain and local en- crypted storage. It uses Attribute-Based Encryption for access control and supports ef- ficient, non-interactive multi-keyword searches via RSA-based tokens and smart con- tracts. Access patterns remain hidden, and security is validated under the random oracle model. Testing on Ethereum shows that it is scalable, low-latency, and cost-effective for healthcare data exchange [20]. This blockchain-based framework securely shares COVID-19 medical records using CP-ABE for fine-grained access control and verifia- ble user revocation. It ensures privacy, efficient decryption, and high throughput, with testing on Ethereum proving its practicality and scalability [21].

Privacy Preserving Data Generalization

By focusing on the bottom level of the generalization lattice and taking use of their structure, the Extended-OIGH algorithm enhances k-anonymity for IGH datasets. Com- pared to current techniques, it reduces node visits by up to 21% by avoiding superfluous nodes through an in-order depth-first search. By doing this, anonymization speed is increased without sacrificing accuracy [22]. The FDG framework enhances Federated Learning in medical imaging by addressing domain shifts using soft- and hard-thresh- olding in the Fourier domain. It modifies image frequency components to reduce noise and improve cross-domain consistency. Applied to polyp segmentation, it significantly boosts accuracy and generalization, making federated training more robust [23].

Privacy Preserving Perturbation

Privacy issues in data mining have resulted in many privacy-preserving methods, some of which are based on randomized data perturbation methods like additive noise for concealing sensitive data. These methods try to preserve privacy by altering original data values with random noise. Evidence shows that these patterns may be exploited by subjecting them to a random matrix-based spectral filtering technique to acquire the original data and in doing so undermine privacy. Theoretical and experimental studies show that additive noise is likely to provide no or minimal privacy protection [15].

Data perturbation is one generally used technique

***Corresponding Author:** Ankit Kumawat

towards this objective aimed at main- taining helpful patterns in the data while maintaining privacy. With the aim to enhance utility along with privacy, this study presents a kd-tree-based perturbation method. The algorithm involves recursively dividing the dataset into increasingly homogeneous sub- sets using a kd-tree data structure. The method preserves statistical dependencies be- tween attributes while safeguarding individual data [16].

Privacy Preserving Noise addition

Network computing systems commonly employ distributed algorithms based on itera- tive exchange of local data among neighboring nodes to implement global computation objectives. During these exchanges, to ensure privacy, nodes commonly append ran- dom noise to their initial data. This paper proves the task of optimizing such an estima- tion and privacy risk analysis. A theoretical model known as optimal distributed esti- mation is introduced to find the optimal possible estimate from local information. The model is subsequently employed for the measurement of disclosure probability under optimal estimation. The authors extend their technique to the privacy-preserving aver- age consensus algorithm and determine the ideal noise needed to achieve maximum data privacy [17]. Data privacy is an important issue in data mining, especially for high-dimensional continuous data, which tends to cause massive data and information loss and makes clustering very challenging. To overcome such challenges, the paper pre- sents a new approach named Anonymized Noise Addition in Subspaces (ANAS). The proposed method minimizes data and information loss and increases both privacy and cluster detection [18].

FUTURE SCOPE

The continuous improvements in data-oriented technologies and privacy requirements emphasize the urgency for deeper studies in privacy-protecting aggregation methods. Upcoming studies need to establish more efficient, scalable, and light-weighted proto- cols that are suitable for actual deployments in resource-scarce scenarios like IoT net- works, mobiles, and edge computing setups. The adaptation of context-sensitive and adaptive differential privacy mechanisms might ensure stronger data utility while still supporting robust privacy protections. Also, adding support for machine learning-based intelligent threat detection and automated privacy risk assessment can make privacy- preserving systems more responsive and intelligent. As applications move into sensitive areas such as healthcare, transportation, and finance, strong privacy-protecting methods will be essential to winning user trust and regulatory recognition.

CONCLUSION

The increasing dependence on data-driven systems in healthcare, IoT, smart grids, and other sensitive areas has turned privacy preservation into an essential concern. This paper provides an extensive exploration and discussion of several privacy-preserving strategies employed in data mining and aggregation, covering techniques such as differential privacy, secure multiparty computation, anonymized perturbation, and state of-the-art record and attribute linkage models. By extensive review of literature and comparative analysis, we underscored the weaknesses of conventional methods—especially those that are based exclusively on additive noise or non-adaptive generalization—and illustrated how contemporary alternatives enhance privacy assurances and data utility. Furthermore, the use of privacy-preserving aggregation schemes in federated learning systems and distributed consensus protocols provides viable avenues towards secure collaborative computation even with dropout or adversarial scenarios.

REFERENCES

1. Liu, Z., Guo, J., Yang, W., et al. (2023). Privacy-preserving aggregation in federated learning: A survey. *IEEE Transactions on Big Data, 9*(2).
2. Yang, Y., Zhang, L., Zhao, Y., et al. (2021). Privacy-preserving aggregation-authentication scheme for safety warning system in fog-cloud based VANET. *IEEE Transactions on Information Forensics and Security, 16*(12).
3. Kong, Q., et al. (2022). Privacy-preserving aggregation for federated learning-based navigation in vehicular fog. *IEEE Transactions on Industrial Informatics, 18*(7).
4. Song, J., Wang, W., Gadekallu, T. R., et al. (2021). EPPDA: An efficient privacy-preserving data aggregation federated learning scheme. *IEEE Transactions on Network Science and Engineering, 8*(4).
5. Helgertz, J., Price, J., Wellington, J., Thompson, K. J., Ruggles, S., & Fitch, C. A. (2022). A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 55*(1), 12–29.
6. Gkoulalas-Divanis, A., Vatsalan, D., Karapiperis, D., & Kantarcioglu, M. (2021). Modern privacy-preserving record linkage techniques: An overview. *IEEE Transactions on Information Forensics and Security, 16*, 4966–4987. https://doi.org/10.1109/TIFS.2021.3114026
7. Wang, J., Li, Z., Lui, J. C. S., & Sun, M. (2022). Topology-theoretic approach to address attribute linkage attacks in differential privacy. *Computers & Security, 113*, 102552. https://doi.org/10.1016/j.cose.2021.102552
8. Lin, C. F. (2003). Quality-delivery system: A conceptual framework of attribute level value linkages. *Total Quality Management & Business Excellence, 14*(10), 1079–1092. https://doi.org/10.1080/1478336032000107663
9. Liu, Z., Guo, J., Yang, W., Fan, J., Lam, K. Y., & Zhao, J. (2023). Privacy-preserving aggregation in federated learning: A survey. *IEEE Transactions on*

***Corresponding Author:** Ankit Kumawat

Big Data. https://doi.org/10.1109/TBDATA.2022.3190835
10. Liu, Z., Guo, J., Lam, K. Y., & Zhao, J. (2023). Efficient dropout-resilient aggregation for privacy-preserving machine learning. *IEEE Transactions on Information Forensics and Security, 18*, 1839–1854. https://doi.org/10.1109/TIFS.2022.3163592
11. Zhao, J. Z., Wang, X. W., Mao, K. M., et al. (2022). Correlated differential privacy of multiparty data release in machine learning. *Journal of Computer Science and Technology, 37*, 231–251. https://doi.org/10.1007/s11390-021-1754-5
12. Li, X., Chen, Y., Wang, C., & Shen, C. (2021). When deep learning meets differential privacy: Privacy, security, and more. *IEEE Network, 35*(6), 148–155. https://doi.org/10.1109/MNET.001.2100256
13. Mackenzie, I. S., Rogers, A., Poulter, N. R., Williams, B., Brown, M. J., Webb, D. J., Ford, I., Rorie, D. A., Guthrie, G., Grieve, J. W. K., Pigazzani, F., Rothwell, P. M., Young, R., McConnachie, A., Struthers, A. D., Lang, C. C., & MacDonald, T. M. (2022). Cardiovascular outcomes in adults with hypertension with evening versus morning dosing of usual antihypertensives in the UK (TIME study): A prospective, randomised, open-label, blinded-endpoint clinical trial. *The Lancet, 400*(10361), 1417–1425. https://doi.org/10.1016/S0140-6736(22)01786-X
14. Lin, C. F. (2002). Attribute-consequence-value linkages: A new technique for understanding customers' product knowledge. *Journal of Targeting, Measurement and Analysis for Marketing, 10*, 339–352. https://doi.org/10.1057/palgrave.jt.574005
15. Kargupta, H., Datta, S., Wang, Q., & Sivakumar, K. (2003, November). On the privacy preserving properties of random data perturbation techniques. In *Proceedings of the Third IEEE International Conference on Data Mining* (pp. 99–106). IEEE. https://doi.org/10.1109/ICDM.2003.1250908
16. Li, X. B., & Sarkar, S. (2006). A tree-based data perturbation approach for privacy preserving data mining. *IEEE Transactions on Knowledge and Data Engineering, 18*(9), 1278–1283. https://doi.org/10.1109/TKDE.2006.136
17. He, J., Cai, L., & Guan, X. (2018). Preserving data privacy with added noises: Optimal estimation and privacy analysis. *IEEE Transactions on Information Theory, 64*(8), 5677–5690. https://doi.org/10.1109/TIT.2018.2842221
18. Shen, J., Yang, H., Vijayakumar, P., & Kumar, N. (2022). A privacy-preserving and untraceable group data sharing scheme in cloud computing. *IEEE Transactions on Dependable and Secure Computing, 19*(4), 2198–2210. https://doi.org/10.1109/TDSC.2021.3050517
19. Wang, M., Guo, Y., Zhang, C., Wang, C., Huang, H., & Jia, X. (2023). MedShare: A privacy-preserving medical data sharing system by using blockchain. *IEEE Transactions on Services Computing, 16*(1), 438–451. https://doi.org/10.1109/TSC.2021.3114719
20. Tan, L., Yu, K., Shi, N., Yang, C., Wei, W., & Lu, H. (2022). Towards secure and privacy-preserving data sharing for COVID-19 medical records: A blockchain-empowered approach. *IEEE Transactions on Network Science and Engineering, 9*(1), 271–281. https://doi.org/10.1109/TNSE.2021.3101842
21. Mahanan, W., Chaovalitwongse, W. A., & Natwichai, J. (2021). Data privacy preservation algorithm with k-anonymity. *World Wide Web, 24*, 1551–1561. https://doi.org/10.1007/s11280-021-00922-2
22. Pan, H., Jha, D., Biswas, K., & Bagci, U. (2025). Frequency-based federated domain generalization for polyp segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICASSP49660.2025.10889662.

***Corresponding Author:** Ankit Kumawat