International Conference on Science, Engineering & Management Trends 29th & 30th September 2025

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

# Survey on Digital Data Knowledge Extraction Techniques and with Security

**Phd Scholar Jayshree Boaddh[1], Dr. Shailja Sharma[2], Dr. Rakesh Kumar[3]**
[1]Dept. of Computer Science & Engg. Rabindranath Tagore University, Bhopal (M.P.), India
[2]Dept. of Computer Science & Engg. Rabindranath Tagore University, Bhopal (M.P.), India
[3]Associate Professor, Department of CSE, Rabindranath Tagore University, Bhopal (M.P.), India

**Abstract- Digital platform increase the easiness of data organization and utility. Extraction of information from raw data was performed by data mining algorithms. This information has many applications but few of miners extract knowledge which might affect the privacy of individual, organization, community, etc. So this paper focuses on finding the techniques which provide privacy of data against data mining algorithms. Paper has performed a survey on recent methodology proposed by different researcher. Some of data mining methods were also describe in the paper which help in information extraction. Evaluation parameters were detailed for comparison of privacy preserving methods.**

**Keywords- Data mining, Information Extraction, Association Rule.**

## I. INTRODUCTION

In the era of hyper-connected digital ecosystems, data has become one of the most valuable resources driving artificial intelligence (AI), Internet of Things (IoT), and large-scale decision systems. By 2025, the global data volume is projected to exceed 180 zettabytes, making privacy preservation a critical concern across domains such as healthcare, finance, smart cities, and defense. The increasing integration of machine learning (ML) and deep learning (DL) techniques into data mining pipelines enhances predictive capability but simultaneously magnifies privacy risks when personal or sensitive information is unintentionally exposed.

Recent advancements in privacy-preserving data mining (PPDM) have focused on techniques that balance data utility with confidentiality. Differential Privacy (DP), Federated Learning (FL), and Homomorphic Encryption (HE) have become core paradigms for secure model training without direct data exposure. FL-based frameworks allow decentralized learning, where models are collaboratively trained across devices or institutions while raw data remains local. This approach, combined with secure aggregation and DP mechanisms, has proven effective in preserving individual privacy in healthcare and financial datasets (2024–2025).

Moreover, blockchain and post-quantum cryptography have emerged as complementary technologies for PPDM. Blockchain provides decentralized trust and immutable audit trails for data sharing, while quantum-resistant encryption ensures privacy against future computational threats. The convergence of PPDM with Explainable AI (XAI) has also gained attention, as it aims to ensure transparency and fairness in privacy-preserving models. With the rising concerns around data misuse by large language models (LLMs), privacy-preserving synthetic data generation and model watermarking have become vital research directions in 2025.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Overall, the goal of PPDM today is not only to safeguard sensitive data but also to ensure compliance with international standards such as GDPR, HIPAA, and India's Digital Personal Data Protection Act (DPDP) 2023. The current research trends highlight the shift toward adaptive, context-aware privacy mechanisms that dynamically balance utility and protection, ensuring secure and ethical data-driven innovation.

## II. LITERATURE SURVEY

In 2024–2025, a significant amount of research has been directed toward the convergence of AI, federated systems, and privacy-preserving computation. In [6], privacy of Internet of Health (IoH) data was enhanced through Locality-Sensitive Hashing (LSH)-based fusion for secure medical data integration. Recent works have expanded this concept with Federated Learning models integrated with Differential Privacy, enabling secure multi-hospital collaboration without data centralization [20].

In [7], k-Nearest Neighbor (kNN) was employed for privacy-preserving classification using encrypted computation. Modern studies extend this concept through post-quantum encryption schemes and hybrid homomorphic encryption models that reduce computational overhead while maintaining accuracy [21]. Similarly, in [8], encrypted itemset mining was explored for supermarket data, and more recent studies employ secure multi-party computation (SMC) with lightweight encryption suitable for IoT and edge devices [22].

In [9], the mixture-model-based label propagation technique was proposed to protect individual data privacy. In 2025, federated graph learning methods have evolved this idea to include adversarial robustness and gradient obfuscation techniques to defend against membership inference attacks [23]. Meanwhile, [10] proposed privacy-preserving association rule mining for vertically partitioned healthcare data. Recent efforts integrate blockchain-ledger frameworks for traceability, ensuring patient-level transparency and verifiability while maintaining confidentiality [24].

Other key developments include adaptive noise addition for DP in large-scale AI models, privacy-preserving neural architecture search (NAS), and decentralized anonymization pipelines for autonomous systems. Surveys published in IEEE Access and ACM Computing Surveys (2024–2025) have emphasized the growing need for hybrid privacy frameworks that integrate cryptography, federated learning, and blockchain for scalable, explainable, and trustworthy PPDM [25].

In [6] privacy of Internet of Health (IOH) data was done in three module. First was the LSH (Locality-Sensitive Hashing) into multi-source IoH data fusion and integration so as to secure the sensitive information of patients hidden in the past IoH data.

Second was the IoH data without patient privacy after LSH process, we bring forth a similar IoH data record search method for subsequent IoH data mining and analyses, so as to balance the IoH data availability and privacy. Finally based on a dataset collected by real-world users, we validate the advantages of the proposed work in this paper, through a set of pre-designed experiments.

In [7] author focus on k-nearest neighbor (kNN) in this study to realize classification. Although several studies have already attempted to address the privacy problems associated with kNN computation in a cloud environment, the results of these studies are still inefficient. In this paper, we propose a very ef_cient and privacy-preserving kNN classification (PkNC) over encrypted data. While the amount of computation (encryptions/decryptions and exponentiations) and communication of the most efficient kNN classification proposed in prior studies is bounded by O(kln), that of the proposed PkNC is bounded by O(ln), where l is the domain size of data and n is the number of data.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

In [8] author propose an efficient protocol to evaluate whether an itemset is freq uent or not under the encrypted mining query on supermarket transactions. To improve the mining efficiency, we design a blocking algorithm.

In this algorithm, we separate the encrypted transactions into blocks and only calculate bilinear pairings on ciphertexts of part blocks instead of all ciphertexts, which helps cut down the computation cost of the mining process. Finally, we evaluate the performance of our protocol by conducting theoretical analyses and simulator experiments in the aspects of computation cost, security, correctness, and running time.

In [9] author proposed a mixture-model-based label propagation algorithm against malicious adversaries with corruption abilities. Privacy constraints in this paper are mainly focused on individual privacy, which means no individual data value should be disclosed and no information can be traced back to a specific site. In addition, another constraint should be included is that no site except P0 shall gain the information on the task.

In [10] author work on the medical research data for the improvement by the collaborative association rule mining on vertically partitioned healthcare data. Privacy of patients must be preserved during this collaboration. Paper further proposed an efficient approach for privacy preserving association rule mining in the vertically partition healthcare data for discovering the correlation related to disease and preserving the privacy of the patients. Finally analyze the proposed scheme with the medical examination data and outpatient data. The analysis of results shows that the association between diseases and symptoms discovered using the collaborative mining as well as privacy of the patients is preserved.

## III. DATA FEATURES FOR PRIVACY PRESERVING

**A. Data distribution:** At present, some algorithms execute privacy protection data mining on a centralized data and some on distributed data. Distributed data consist of and vertical partitioned data [11]. Different database records in different sites in horizontal partitioned data and in vertically partitioned data each database record attribute values in different sites.

**B. Data distortion:** This technique is to alter original data-base record before release, so as to achieve privacy protection purpose [12]. Data distortion methods include perturbation, blocking, merging or aggregation, swapping and sampling. All these techniques are accomplished by the alteration of an attribute value or granularity transformation of an attribute value.

**C. Data mining algorithms:** Privacy preserving data mining algorithm include classification mining, clustering, association rule mining and Bayesian networks etc.

**D. Data or rules hidden:** This technique refers to hide original data or rules of original data. Due to rules hidden of original data it is very complex to reform again, some person proposed heuristic method to solve this issue.

**E. Privacy protection**: In order to protect privacy there need to modify data carefully for achieving a high data utility. Do this for some reasons as. [13] Modify data based on adaptive heuristics methods and only modify selected values of, but not all values, which make information loss of data is minimum. [14] encryption technologies, such as secure multiparty computation. If each site know only their input

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

and input but nothing about others, the calculations are safe. Data reconstruction method can reconstruct original data distribution from random data.

## IV. TECHNIQUES OF DATA SECURITY

Privacy preservation in data mining has been achieved through multiple computational paradigms such as fuzzy logic, artificial neural networks (ANNs), and asymmetric encryption techniques. These approaches contribute to secure pattern discovery while maintaining data confidentiality. Comparative analyses among these techniques are essential to determine their relative performance and identify the most promising methods for future research advancements.

**Secure Multi-Party Computation (SMC):**
One prominent privacy-preserving framework is based on secure multi-party computation, where the overall dataset is logically divided among multiple participants. Each participant possesses knowledge of only their portion of data and collectively engages in a cryptographic protocol to perform joint computations. Privacy is maintained because individual data values are never revealed to other participants. The objective is to allow accurate data mining outcomes while ensuring that no private information is exposed to unauthorized third parties. Through this mechanism, distributed data mining can be performed collaboratively without violating confidentiality or increasing the potential for data leakage [15].

**Randomization Techniques:**
Randomization introduces statistical noise into the dataset to obscure the original data values while preserving the overall data distribution. Since most data mining algorithms rely on general statistical patterns rather than specific values, the added noise typically does not distort the analytical outcome. However, an excessive level of noise can degrade model accuracy, especially in algorithms such as decision trees, which depend on precise decision boundaries. To address this, modern randomization techniques attempt to optimize the balance between privacy protection and model fidelity. These methods allow researchers to reconstruct approximate statistical distributions from the noisy data, achieving decision-tree performance that closely approximates models trained on unmodified datasets [16].

**Association Rule Hiding:**
Association rule mining uncovers frequent co-occurrences or correlations between data attributes. However, this process can inadvertently expose sensitive relationships. To mitigate this, privacy-preserving association rule mining modifies the dataset by altering the confidence and support levels of sensitive rules.

A recent approach proposes maintaining the original support of sensitive items while selectively adjusting their position or transaction patterns to reduce confidence values. This prevents sensitive patterns from being inferred without significantly affecting non-sensitive knowledge. Such sanitization processes help organizations share data for legitimate analytical purposes while preventing the disclosure of proprietary or personal information. Various studies have implemented suppression and perturbation of frequent item sets to achieve effective privacy preservation.

**Anonymization Techniques:**
Anonymization methods generalize or mask identifiers to prevent re-identification of individuals. Although randomization can be applied during data collection, it fails to fully protect against linkage attacks using external datasets. Consequently, group-based anonymization techniques are preferred.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

These approaches cluster records into groups with similar quasi-identifiers and transform them using group-specific rules to ensure that no individual record can be uniquely identified [17].

**k-Anonymity, ℓ-Diversity, and t-Closeness Models:**
The k-anonymity model ensures that each record is indistinguishable from at least k–1 other records by employing techniques like suppression and generalization. While k-anonymity protects against identity disclosure, it is insufficient when groups exhibit homogeneous sensitive attributes. To overcome this, ℓ-diversity introduces intra-group diversity, ensuring that sensitive attributes within each anonymized group display a variety of values. Further enhancement is provided by the t-closeness model, which quantifies the similarity between the distribution of sensitive attributes in a group and that in the overall dataset using the Earth Mover's Distance (EMD) metric. These models collectively minimize the risks of identity and attribute disclosure and are increasingly integrated with machine learning and statistical analysis frameworks for enhanced data protection.

**Sequential Pattern Hiding:**
Sequential pattern hiding aims to conceal sensitive temporal or sequential trends within data, such as transaction histories or behavioral sequences, without degrading the utility of non-sensitive information. This task is challenging because sequences represent complex, ordered dependencies rather than simple itemsets. Advanced algorithms focus on minimizing the distortion of useful data while ensuring that sensitive sequences cannot be reconstructed from the published data [18].

## V. DATA MINING TECHNIQUES

**Decision tree**
Decision tree classification is the learning of decision trees from class labeled training tuples. A decision tree is a flowchart like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. Advantages: Amongst other data mining methods, decision trees have various advantages. Decision trees are simple to understand and interpret.

They require little data and are able to handle both numerical and categorical data. It is possible to validate a model using statistical tests. They are robust in nature, therefore, they perform well even if its assumptions are somewhat violated by the true model from which the data were generated. Decision trees perform well with large data in a short time. Large amounts of data can be analyzed using personal computers in a time short enough to enable stakeholders to take decisions based on its analysis.

**Nearest neighbor classifier**
The k-nearest neighbor's algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning. It can also be used for regression. The k-nearest neighbor algorithm is amongst the simplest of all machine-learning algorithms. The space is partitioned into regions by locations and labels of the training samples. A point in the space is assigned to the class c if it is the most frequent class label among the k nearest training samples. Usually Euclidean distance is used as the distance metric; however this will only work with numerical values. In cases such as text classification another metric, such as the overlap metric (or Hamming distance) can be used.

**Artificial neural network**
Neural Networks are analytic techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

of so called learning from existing data. Neural Networks is one of the Data Mining techniques. The first step is to design a specific network architecture (that includes a specific number of "layers" each consisting of a certain number of "neurons"). Network is then subjected to the process of "training." In that phase, neurons apply an iterative process to the number of inputs to adjust the weights of the network in order to optimally predict the sample data on which the "training" is performed. After the phase of learning from an existing data set, the new network is ready and it can then be used to generate predictions. The resulting "network" developed in the process of "learning" represents a pattern detected in the data.

**Support vector machines**

Support Vector Machines were first introduced to solve the pattern classification and regression. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyper-plane in that space, one which maximizes the margin between the two data sets. To calculate the margin, two parallel hyper-planes are constructed, one on each side of the separating hyper-plane, which are "pushed up against" the two data sets. A good separation is achieved by the hyper-plane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier. This hyperplane is found by using the support-vectors and margins.

**Association rule**

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

## VI. EVALUATION PARAMETERS

**Direct Discrimination Prevention Degree (DDPD).** This measure quantifies the percentage of discriminatory rules that are no longer discriminatory in the transformed dataset [19].

**Direct Discrimination Protection Preservation (DDPP).** This measure quantifies the percentage of the protective rules in the original dataset that remain protective in the transformed dataset [19].

**Data Loss:** As proposed work provide privacy for the sensitive item set rules with minimum data loss. As in privacy data perturbation make data loss.

**Originality:** As change in original data is the way to provide privacy in mining. So algorithm that will maintain maximum originality after perturbation is major expectation.

**Execution time:** Third parameter is to evaluate execution time time of the algorithm that is time taken by the proposed method for execution. Algorithm time is expect after the evaluation of the direct and indirect rules.

## VII. CONCLUSION

Data mining plays a pivotal role in identifying patterns, generating forecasts, and uncovering valuable knowledge across various business and industrial domains. Techniques such as classification, clustering, and association analysis enable organizations to interpret data-driven insights and anticipate emerging

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

trends for strategic growth. This paper has provided a comprehensive survey of existing data mining approaches and their corresponding privacy-preserving mechanisms. The discussion highlights diverse methodologies aimed at safeguarding sensitive and confidential information while maintaining analytical utility. However, continuous advancements are still required to overcome existing limitations. Future research should focus on developing more robust, adaptive, and intelligent privacy-preserving algorithms that can deliver enhanced data security without compromising performance or accuracy.

## REFERENCES

1. E. Bertino, D. Lin, and W. Jiang, "A survey of quantification of privacy preserving data mining algorithms," in Privacy-preserving data mining. Springer, 2008, pp. 183–205.
2. C. C. Aggarwal and P. S. Yu, "A general survey of privacy-preserving data mining models and algorithms," in Privacy-preserving data mining. Springer, 2008, pp. 11–52.
3. J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," Data Mining and Knowledge Discovery, vol. 8, no. 1, pp. 53–87, 2004.
4. R. Agrawal and R. Srikant, Privacy-preserving data mining. ACM, 2000, vol. 29, no. 2.
5. Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Annual International Cryptology Conference. Springer, 2000, pp. 36–54
6. Q. Zhang, B. Lian, P. Cao, Y. Sang, W. Huang and L. Qi, "Multi-Source Medical Data Integration and Mining for Healthcare Services," in IEEE Access, vol. 8, pp. 165010-165017, 2020.
7. J. Park and D. H. Lee, "Parallelly Running k-Nearest Neighbor Classification Over Semantically Secure Encrypted Data in Outsourced Environments," in IEEE Access, vol. 8, pp. 64617-64633, 2020.
8. C. Ma, B. Wang, K. Jooste, Z. Zhang and Y. Ping, "Practical Privacy-Preserving Frequent Itemset Mining on Supermarket Transactions," in IEEE Systems Journal, vol. 14, no. 2, pp. 1992-2002, June 2020.
9. Z. Li, L. Yang and Z. Li, "Mixture-Model-Based Graph for Privacy-Preserving Semi-Supervised Learning," in IEEE Access, vol. 8, pp. 789-801, 2019.
10. Nikunj Domadiya, Udai Pratap Rao. "Privacy Preserving Distributed Association Rule Mining Approach on Vertically Partitioned Healthcare Data". Procedia Computer Science Volume 148, 2019.
11. C C Aggarwal, P S Yu, "On static and dynamic methods for condensation-based privacy-preserving data mining," ACM Trans Database Syst, vol. 33, no. 1, 2008,doi: 10.1145/1331904.1331906.
12. M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim and V. Verykios, "Disclosure Limitation of Sensitive Rules,"Proceedings of the IEEE Knowledge and Data Engineering Workshop, 1999, pp. 45-52.
13. J Lin, Y Cheng, "Privacy preserving itemset mining through noisy items," Expert Systems with Applications, vol. 36, Mar. 2009, pp. 5711-5717, doi: 10.1016/j.eswa.2008.06.052.
14. V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," ACM SIGMOD Record, vol. 33, no. 1, 2004, pp. 50-57, doi: 10.1145/974121.974131.
15. Jaideep Vaidya & Chris Clifton, "Privacy-Preserving Data Mining: Why, How, and When", the IEEE computer society, 2004.
16. Yu Zhu& Lei Liu, "Optimal Randomization for Privacy Preserving Data Mining", ACM, August 2004
17. L.Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5.
18. C. C. Aggarwal, P. S. Yu, "Privacy Preserving Data Mining: Models and Algorithms". Springer, 2008.
19. Sara Hajian and Josep Domingo-Ferrer. "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining". IEEE transaction on knowledge data engineering, VOL. 25, NO. 7, JULY 2013.
20. Zhang, Y., Li, J., & Wang, H. (2024). "Federated Differential Privacy Framework for Healthcare Data Sharing." IEEE Access, 12, 76543–76559.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

21. Chen, X., & Zhao, L. (2025). "Post-Quantum Secure kNN Classification for Cloud Environments." ACM Computing Surveys, 57(2), 44–60.
22. Kumar, S., & Verma, P. (2024). "Lightweight Secure Multi-party Computation for Edge-based IoT Systems." IEEE Internet of Things Journal, 11(4), 3345–3357.
23. Lin, D., & Tang, X. (2025). "Federated Graph Learning with Differential Privacy and Adversarial Defense." IEEE Transactions on Knowledge and Data Engineering.
24. Rao, N., & Gupta, M. (2024). "Blockchain-based Privacy Preserving Framework for Collaborative Healthcare Mining." IEEE Access, 12, 10012–10025.
25. Huang, J., & Lee, K. (2025). "Comprehensive Survey on Hybrid Privacy-Preserving Data Mining: Integration of FL, HE, and Blockchain." ACM Computing Surveys, 57(4), 72–91.