Data Mining with Privacy: A Review of Protection Techniques and Emerging Trends

C. Kosala Jayaweera, R.M. Nayanajith Rathnayaka, and M. S. Shafana

Abstract Privacy-preserving data mining (PPDM) is critical for balancing the utility and confidentiality of the data. Existing surveys analyze techniques in isolation and neglect real world domain constraints. To address this gap, this survey delivers a unified comparative analysis of traditional, statistical, cryptographic and distributed PPDM paradigms by evaluating their practical traden offs and domain specific viability. This Systematic Literature Review (SLR) follows a PICO-guided protocol analyzing literature from 2012 to 2025 across six major digital repositories including IEEE Xplore, ACM and ScienceDirect. We systematically contrast foundational methods by highlighting high dimensional limitations in traditional anonymization (k-anonymity, l-diversity), noise calibration challenges in Differential Privacy (DP) and computational bottlenecks in cryptographic schemes (FHE, SMPC) against decentralized paradigms like Federated Learning (FL). Furthermore, this survey bridges classical techniques with emerging frontiers, evaluating DP deep neural networks, GAN generated synthetic data and blockchain-based auditability. By contextualizing these techniques within conflicting domain demands such as strict compliance of healthcare and utility-driven priorities of e-commerce, we identify critical research gaps in multi objective optimization and standardization, providing a clear roadmap for scalable, domain tailored PPDM deployment.

***Index Terms*— Cryptographic Techniques, Data Mining, Hybrid Frameworks, Privacy-preserving Data Mining**

I. Introduction

IN the era of big data, the true value of information incomes from the ability to analyse it and uncover meaningful insights. The ability to analyse, interpret, and utilize data effectively has become a cornerstone of informed decision-making across a spectrum of domains, including healthcare [1], [2], finance, marketing [3], and social media analytics [4]. As data-driven approaches become more widespread, there is a growing need to tackle the challenges that come with handling large and often sensitive datasets [5]. Although the abundance of data offers great opportunities for predictive analytics and strategic decision-making, it also brings serious concerns about privacy and data security.

Data mining, defined as the computational process of discovering patterns, trends, and associations within large datasets, plays a critical role in knowledge discovery and predictive modelling. Unlike traditional statistical approaches that begin with a hypothesis and collect data to validate it, data mining is inherently data-driven such as formulating hypotheses from pre-existing data [5]. Closely aligned with machine learning, data mining focuses on uncovering latent knowledge, with machine learning typically oriented toward task performance. While supervised learning dominates machine learning, data mining traditionally emphasizes unsupervised techniques, where measuring the quality and interpretability of results remains an ongoing challenge [6], [7].

The utility of data mining is substantial, enabling organizations to forecast trends, identify anomalies, and optimize performance. However, as mining techniques delve deeper into datasets containing personal and often sensitive information, the risk of privacy breaches intensifies [8]. The inadvertent disclosure of private data such as health records, purchasing behaviours, or geolocation histories, can lead to identity theft, profiling, discrimination, and a significant erosion of public trust. Consequently, preserving privacy without compromising analytical utility has emerged as a critical and complex objective in the field of data science [9].

The increasing focus on data privacy by society and regulators—reflected in new legislation such as the General Data Protection Regulation (GDPR) [10] and the California Consumer Privacy Act (CCPA) [8]—has catalyzed the rapid development of Privacy-Preserving Data Mining (PPDM) techniques. PPDM aims to safeguard individual privacy while retaining the usefulness of mined data [11]. Figure 1 presents a taxonomy of PPDM techniques organized into three pillars such as traditional and statistical privacy techniques (syntactic models and randomization), cryptographic and distributed methods (homomorphic encryption, secure multi-party computation and federated learning) and emerging hybrid frameworks (including

Charuni Kosala Jayaweera is a postgraduate researcher pursuing an Masters by Research in the Department of Computer Science and Engineering at the University of Moratuwa, Sri Lanka. (Email: charunikosala9723@gmail.com)

R.M. Nayanajith Rathnayaka is a Lecturer (Probationary) at the Department of ICT, University of Ruhuna, Sri Lanka. (Email: mailtonayanajith@gmail.com)

M.S. Shafana is a Senior Lecturer at the Department of ICT, South Eastern University of Sri Lanka, Sri Lanka. (Email: zainashareef@gmail.com)

ISSN: 2961-5992

differentially private deep nets, synthetic data generation and blockchain auditability). PPDM aims to defend individual privacy while retaining the usefulness of mined data. It encompasses a range of strategies including data anonymization, perturbation, differential privacy, cryptographic approaches and secure multi party computation. These techniques attempt to balance the inherent tradeoff between data utility and confidentiality by enabling analytics that respect ethical and legal boundaries [11].This paper presents a comprehensive review of privacy-preserving data mining techniques, examining both foundational methods and state-of-the-art innovations. Paper begins by outlining core concepts in data mining and delineating the nature of associated privacy risks. Key PPDM techniques are then evaluated, with an assessment of their respective strengths and limitations. Finally, recent advancements are explored, and critical research gaps are identified, aiming to guide future work toward privacy-aware data mining practices that uphold ethical standards and comply with regulatory mandates.

II. METHODOLOGY

This systematic literature review (SLR) adhered to a structured protocol comprising planning, conducting, and reporting phases. During the planning phase, three research questions were formulated to guide the review: RQ1: How have privacy-preserving data mining techniques evolved from early perturbation and anonymization methods? RQ2: Which cryptographic, statistical, or hybrid approaches dominate contemporary research? RQ3: What are the remaining challenges for developing regulation-compliant, next-generation solutions? A PICO (Population, Intervention, Comparison, Outcome) framework informed the development of search strings, combining terms such as "data mining" AND ("privacy-preserving" OR "confidential" OR "secure analytics") AND ("anonymization" OR "differential privacy" OR "homomorphic encryption" OR "federated learning"). Searches were conducted across six digital libraries such IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, arXiv, and Google Scholar covering peer-reviewed and preprint publications from 2012 to 2025.

The conducting phase employed a five-stage selection process to ensure rigor. First, titles, abstracts, and keywords were screened to exclude irrelevant studies. Second, duplicate records across databases were removed. Third, full-text articles were reviewed based on inclusion criteria, selecting primary studies that either (a) proposed, evaluated, or compared privacy-preserving techniques in data mining, or (b) examined utility-privacy trade-offs using real-world datasets. Exclusion criteria eliminated non-English studies, abstract-only submissions, tutorials, and editorials. Fourth, backward and forward snowballing on included studies identified additional relevant works from high-impact venues. Fifth, tracing contributions from key authors in cryptography and differential privacy ensured comprehensive coverage. This process resulted in 55 studies for final synthesis.

Data extraction captured key attributes from each study, including bibliographic details, threat model, privacy-preserving technique (e.g., statistical perturbation, secure multi-party computation, trusted execution environments, synthetic data generation), evaluation metrics, application domain, limitations, and proposed future work. A relational matrix was constructed to map techniques to publication year, sector, and privacy guarantee type, facilitating the creation of visual timelines and heatmaps to illustrate research trends and maturity. The synthesis produced three outputs: (1) a detailed SLR document compiling all included studies, (2) a taxonomy chart classifying technique families and their evolution, and (3) an evidence matrix highlighting open challenges, particularly in the composability of federated learning with homomorphic encryption and the auditability requirements of emerging regulations.

III. LITERATURE REVIEW

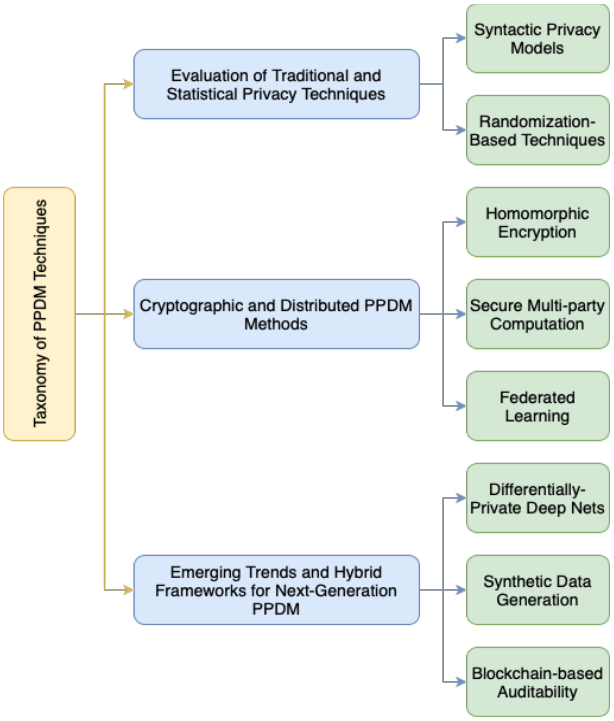

Fig. 1. Taxonomy of PPDM Techniques

ISSN: 2961-5992

A. Evaluation of Traditional and Statistical Privacy Techniques

Traditional privacy-preserving methods such as k-anonymity, l-diversity, and t-closeness remain relevant due to their conceptual simplicity and low computational demands. These methods perform well on structured, low-dimensional datasets where quasi-identifiers are clearly defined. For instance, k-anonymity often preserves a reasonable balance between privacy and utility in such cases. However, as data dimensionality increases, these techniques face challenges due to data sparsity and heightened re-identification risks. More recent analyses have confirmed that syntactic models like t-closeness tend to suffer from excessive information loss in high-dimensional or dynamic environments [12], [13]. Comparative evaluations highlight a persistent trade-off: stronger privacy protections typically come at the cost of reduced data granularity and analytical value [14].

Randomization-based techniques such as data perturbation and noise injection have become central to privacy-preserving data mining, particularly under the differential privacy (DP) framework. The Laplace and Gaussian mechanisms offer formal privacy guarantees that are provably resistant to inference attacks. However, the strength of privacy governed by the ε (epsilon) parameter has a direct impact on utility. Recent studies demonstrate that ε values in the range of 0.1–1.0 offer strong privacy but lead to significant degradation in tasks like classification and clustering, whereas values above 2.0 favour utility but weaken privacy [15], [16]. Additionally, the cumulative nature of the privacy budget under repeated queries introduces further challenges, making privacy budget management critical in real-world applications [17], [18].

In summary, cryptographic and distributed paradigms such as FHE, SMPC, and FL provide a robust answer to the data leakage vulnerabilities of traditional methods by enabling computations strictly on encrypted or localized data. While they offer mathematically provable security and are highly suitable for sensitive sectors like healthcare and finance, they shift the core challenge from a privacy utility trade off to a computational-communication bottleneck. As a result, current research is heavily focused on optimizing these protocols to reduce latency and bandwidth overheads.

B. Cryptographic and Distributed PPDM Methods

Lattice-based fully homomorphic encryption (FHE) schemes built on the Learning with Errors (LWE) problem enable complex computation on ciphertexts while resisting quantum attacks and improving efficiency [19]. A recent multi-key FHE variant eliminates the need for a common reference string by using distributed key generation, boosting decentralisation for scenarios such as privacy-preserving crowd-sensing [20]. Applied to smart-grid analytics, lattice-based FHE outperforms Paillier and Boneh–Goh–Nissim schemes in speed and ciphertext size, thanks to optimised bootstrapping [21]. These advances support polynomial-time operations suitable for tasks like secure genomic processing and smart-meter aggregation.

Secure multi-party computation (SMPC) lets multiple parties compute jointly over private inputs such as vital for fraud detection and healthcare collaboration without revealing raw data [22]. Hybrid protocols that mix arithmetic and Boolean circuits improve performance on diverse, high-dimensional data. Function secret sharing (FSS) cuts communication overhead by about 50% for non-linear functions such as neural-network activations [23]. These protocols achieve provable security under both semi-honest and malicious models while mitigating cost and latency. Deployments in finance and healthcare networks confirm SMPC's practicality for secure, large-scale analytics [24].

Federated learning (FL) trains models across devices without moving raw data, addressing privacy in healthcare, industrial diagnostics, and cloud analytics [25], [26]. A class-imbalanced FL framework for wind-turbine fault diagnosis combines biometric authentication with secure aggregation, sustaining high accuracy despite skewed data [27]. Frameworks such as SF-XGBoost integrate DP and secure aggregation, achieving 99.31% accuracy and 0.9994 ROC-AUC while preserving privacy [26], [28]. Adaptive-budget methods like FedProx adjust privacy loss to data sensitivity, improving stability over FedAvg and delivering up to 30% higher accuracy in heterogeneous medical datasets [29], [30].

Ultimately, emerging hybrid frameworks represent a strategic convergence of privacy mechanisms and modern machine learning capabilities. By leveraging techniques like concentrated DP in deep neural networks, generative synthetic data, and immutable blockchain ledgers, these approaches address both the utility degradation of traditional models and the trust limitations of decentralized systems. However, standardizing evaluation metrics that define robust validation frameworks and overcoming integration complexities remain critical hurdles before these hybrid systems can achieve widespread, seamless adoption.

C. Emerging Trends and Hybrid Frameworks for Next-Generation PPDM

Differentially private deep learning concentrated differential privacy (CDP) with dynamic budget allocation reduces privacy loss while retaining accuracy for crowdsourced data [31], [32]. Per-layer and per-device gradient clipping accelerates and streamlines private training, surpassing flat clipping and even improving results when fine-tuning models as large as GPT-3 [33], [34]. For Transformers, Re-Attention and Phantom Clipping tackle attention distraction and

ISSN: 2961-5992

inefficient clipping in a model-agnostic way [35], [36], further refining the privacy–utility balance.

Synthetic data generation (SDG), GAN-based synthetic tabular data offers a compelling alternative to anonymisation and even FL, supplying cost-effective, shareable datasets with strong privacy guarantees, though the field still lacks universal evaluation metrics [37]. Recent studies show generative models can yield data indistinguishable from real-world samples, supporting cross-organisational collaboration [38], [39]. Ongoing work evaluates privacy, utility, and fairness, but robust validation frameworks are still needed [30].

Blockchain-based auditability, Blockchain enhances transparency and non-repudiation in PPDM. A hierarchical identity-based scheme logs data usage via auditable smart contracts, preventing unauthorised processing while preserving owner privacy [40]. In IoT, blockchain-backed attribute-based access control secures private data with high throughput [34]. BlockCryptoAudit pairs blockchain with Paillier encryption and smart-contract-based role controls, realising 98% risk mitigation and 99% audit quality at low overhead [41]. Although promising, seamless encryption integration and scalability remain open challenges.

A comparative analysis of these three paradigms reveals an evolutionary trajectory governed by the ongoing tradeoffs among data utility, privacy guarantees and computational cost. Traditional and statistical techniques offer low computational complexity, yet they struggle to preserve utility in high dimensional datasets due to severe data perturbation. Cryptographic and distributed methods such as FHE, secure multi-party computation and federated learning effectively eliminate centralized data leakage risks. Even though introduce substantial computational latency and communication overhead that limit their application in resource constrained environments such as IoT. To bridge these limitations, emerging hybrid frameworks synthesize strengths across paradigms by integrating differential privacy into federated learning, employing GAN-based synthetic data and leveraging blockchain for transparent auditability to maximize high dimensional utility and trust. Ultimately, because lack of universal solution in a single paradigm, technique selection must be strictly context-aware such that traditional models suit simple low-dimensional queries; cryptographic schemes fit compute-heavy environments requiring strict regulatory compliance; and emerging hybrid frameworks excel in complex, collaborative machine learning workflows.

IV. Discussion And Synthesis

A significant research trend involves optimizing the privacy-utility trade-off through multi-objective techniques. Studies leveraging Pareto-based optimization, genetic algorithms, and reinforcement learning have demonstrated promising results in identifying optimal configurations for privacy mechanisms. For instance, some models dynamically adjust perturbation levels based on sensitivity analysis or predicted utility degradation [42].

These adaptive methods outperform static configurations in tasks such as recommendation systems, where personalization must coexist with user privacy. The integration of context-aware privacy budgets and user-level control mechanisms also enhances practical utility without compromising confidentiality. However, a lack of standardization and benchmark datasets makes it difficult to generalize these optimization strategies across domains.

Domain-specific analyses reveal varying adoption and performance of privacy-preserving methods. In the healthcare sector, strict compliance requirements (e.g., HIPAA, GDPR) drive the use of strong privacy guarantees such as anonymization combined with DP or cryptography [43], [44]. In contrast, e-commerce and social media applications prioritize data utility for personalization and trend analysis, often adopting lighter privacy mechanisms. In financial applications, encrypted data mining and SMC are increasingly used to detect fraud while preserving customer confidentiality. Meanwhile, smart grid and IoT systems benefit from FL and edge-based anonymization due to their distributed nature and latency constraints [45]. These variations emphasize the importance of contextualizing privacy-preserving techniques according to domain-specific risk, regulation, and performance expectations.

Cryptographic solutions such as SMC and HE offer robust security guarantees [46], [47]. In theory, these methods ensure data confidentiality throughout computation, making them suitable for highly sensitive domains such as healthcare and finance.

ISSN: 2961-5992

TABLE I

EVIDENCE MATRIX – OPEN CHALLENGES AT THE INTERSECTION OF FEDERATED LEARNING (FL), HOMOMORPHIC ENCRYPTION (HE) AND NEW AUDITABILITY MANDATES

Focus Area	Specific Open Challenge	Impact on Practice	Research / Engineering Gap
FL & HE composability	High computational latency caused by bootstrapping and ciphertext arithmetic on large-scale models	Slows round-time; impractical for edge devices and time-sensitive inference	Hardware-assisted FHE kernels and mixed-precision selective encryption pipelines
	Communication overhead – encrypted updates are an order of magnitude larger than plaintext	Mobile bandwidth & battery drain; drop-out rates rise	Ciphertext compression, quantised FHE, and sparsity-aware aggregation
	Multi-key key-management complexity in cross-organisation FL	On-boarding delays, revocation pain-points	Standardised DKG APIs and post-compromise key-rotation for MK-FHE
	Model-utility drop from encrypted fixed-point maths	Hard to justify in high-stakes domains (clinical, finance)	Adaptive precision schemes & hybrid FHE-DP noise budgeting
Regulatory auditability	Mandatory event-logging for high-risk AI (EU AI Act Art 19) conflicts with ciphertext opacity	Providers must expose behaviour traces yet keep data encrypted	Cryptographically signed meta-logs & zero-knowledge attestation of FL rounds
	Traceability v. privacy tension – encrypted gradients are "black boxes" for auditors	Regulators may deem FL-only logs insufficient	Explainable-FL toolchains that derive audit features from encrypted updates
	Blockchain-based audit trails face GDPR/PDPA data-minimisation limits	Immutability v. "right to erasure" & cross-border transfer rules	Layer-2 off-chain reaction or chameleon-hash ledgers for PET scenarios
	Multi-jurisdiction retention & oversight (EU AI Act 2026; U.S. state AI laws)	Firms face fragmented timelines and overlapping audits	Harmonised "once-for-many" audit schemas and provenance-aware policy engines

FL=Federated Learning, HE=Homomorphic Encryption, FHE=Fully Homomorphic Encryption, ML=Machine Learning, DKG=Distributed Key Generation, API=Application Programming Interface, MK-FHE=Multi-Key Fully Homomorphic Encryption, DP=Differential Privacy, EU AI Act=European Union Artificial Intelligence Act, ZK=Zero-Knowledge, PET=Privacy-Enhancing Technologies, GDPR=General Data Protection Regulation, PDPA=Personal Data Protection Act, U.S.=United States.

HE, while enabling operations on encrypted data, introduces substantial computational overhead. Recent benchmarks show that even partially homomorphic schemes (e.g., Paillier encryption) exhibit latency that scales poorly with dataset size. FHE, although conceptually ideal, remains impractical for large-scale mining tasks due to exorbitant computational costs and memory consumption [48]. Secure multiparty computation, though more efficient than FHE in certain cases, struggles with communication overhead and synchronization issues in distributed systems. These limitations restrict the scalability and applicability of cryptographic solutions in real-time or high-throughput environments [49].

Emerging approaches like Federated Learning (FL) are gaining traction as viable privacy-preserving strategies, particularly in mobile and edge computing environments [50]. FL enables collaborative model training without centralizing raw data, thereby minimizing the risk of direct data leakage. Studies indicate that FL performs well in heterogeneous environments, maintaining acceptable levels of model accuracy across diverse datasets [27], [51]. However, FL is not inherently private. Model updates can leak sensitive information through gradient inversion attacks or reconstruction methods [52], [53]. To counter this, recent works integrate differential privacy or secure aggregation protocols within federated frameworks [52], [54]. Experimental results show that while these integrations enhance privacy, they often reduce convergence speed and final model accuracy. Furthermore, FL introduces challenges such as device heterogeneity, unreliable communication, and incentive mechanisms for client participation [55]. Table 1 synthesizes the primary architectural and regulatory open challenges to bridge the gap between theoretical PPDM models and practical development. Furthermore, this discusses the real-world impacts and highlights the critical research gaps necessary to advance the field. The synthesis reveals that combining FL and HE offers robust privacy. it creates severe computational and bandwidth bottlenecks such as FHE bootstrapping latency, massive ciphertext updates and multi key management friction that require solutions like hardware assisted FHE kernels, quantized FHE and

ISSN: 2961-5992

standardized distributed key generation APIs. Furthermore, the table emphasizes a critical conflict between regulatory transparency mandates and privacy-preserving opacity. Encrypted model gradients act as black boxes to auditors while immutable blockchain audit trails clash with statutory right to erasure provisions. Resolving these bottlenecks demands next generation technical interventions including zero knowledge attestation frameworks, chameleon-hash ledgers and harmonized cross jurisdictional audit schemas. Ultimately, Table 1 supports the paper's overarching conclusion that scalable, regulation compliant PPDM cannot rely on cryptographic primitives in isolation, but instead requires a co-designed paradigm that aligns low level hardware optimization with high level legal and policy frameworks. Despite these challenges, FL represents a promising shift in privacy-aware data mining by decentralizing learning processes and reducing reliance on trusted intermediaries.

V. Conclusion

In conclusion, while PPDM has evolved from basic data concerns to sophisticated hybrid frameworks by bridging the gap between theoretical security and production deployment, it requires addressing three high impact open research problems across distinct domains. Technically, the field requires hardware accelerated FHE kernels, mixed precision encryption pipelines and adaptive DP budgeting to eliminate the severe computational and bandwidth bottlenecks inherent in FL–HE composability. Regulative, researchers must resolve the tension between ciphertext opacity and statutory transparency (EU AI Act, GDPR) by developing zero knowledge attestation schemes and mutable chameleon hash ledgers that permit verifiable auditing and data erasure without compromising underlying privacy. Socio technically, the absence of standardized evaluation metrics and domain tailored governance models hinders real world adoption; future work must establish cross-domain benchmarking frameworks and provenance aware policy engines to harmonize conflicting stakeholder demands across healthcare, IoT and finance. Advancing these targeted research vectors will transition PPDM from isolated cryptographic primitives to scalable, regulation compliant systems ready for real world deployment.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the author(s) used ChatGPT, an AI generative tool, to support language editing and refinement. After using this tool, the author(s) reviewed and edited the manuscript as needed and take full responsibility for the content of the published article.

References

[1] M. Hernandez et al., "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing*, vol. 493, pp. 28–45, 2022.

[2] Y. Zhou, X. Wu, and Y. Li, "PrivTree: A differential privacy framework for hierarchical data," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 9, pp. 1701–1715, 2020.

[3] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Security Privacy (SP)*, 2017, pp. 3–18.

[4] A. Marandi et al., "Lattice-based homomorphic encryption for privacy-preserving smart meter data analytics," *The Computer Journal*, vol. 67, no. 5, pp. 1687–1698, 2023.

[5] T. Calders and B. Custers, "What is data mining and how does it work?," in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, B. Custers, Ed. Berlin, Heidelberg: Springer, 2013, pp. 27–42.

[6] J. O. Arowoogun et al., "A comprehensive review of data analytics in healthcare management: Leveraging big data for decision-making," *World J. Adv. Res. Rev.*, vol. 21, no. 2, pp. 1810–1821, 2024.

[7] X. Gao et al., "Enhancing topic interpretability for neural topic modeling through topic-wise contrastive learning," arXiv preprint, 2024. [Online]. Available: https://arxiv.org/abs/2412.17338

[8] C. Gilbert and M. Gilbert, "Privacy-preserving data mining and analytics in big data environments," *SSRN Electron. J.*, 2025. [Online]. Available: https://doi.org/10.2139/ssrn.5258795

[9] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: Privacy and data mining," *IEEE Access*, vol. 2, pp. 1149–1176, 2014.

[10] L. Franke et al., "An exploratory mixed-methods study on General Data Protection Regulation (GDPR) compliance in open-source software," in *Proc. Association for Computing Machinery*, 2024, pp. 325–336.

[11] N. Yuvaraj, K. Praghash, and T. Karthikeyan, "Privacy preservation of the user data and properly balancing between privacy and utility," *Int. J. Business Intell. Data Mining*, vol. 20, no. 4, pp. 394–411, 2022.

[12] Y. Zhang, S. Ji, and T. Wang, "Differentially private releasing via deep generative model," arXiv preprint arXiv:1801.01594, 2018.

[13] K. Singh, M. Thakur, and N. P. Rana, "Data privacy and anonymization: A survey," *Inf. Security J.: A Global Perspective*, vol. 28, no. 1, pp. 24–43, 2019.

[14] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 97–107, 2016.

[15] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *Proc. 28th USENIX Security Symp.*, Santa Clara, CA, USA, Aug. 2019, pp. 1895–1912.

[16] Q. Liu, X. Meng, Y. Chen, and Z. Qin, "A systematic review of differential privacy in machine learning," *Inf. Fusion*, vol. 72, pp. 160–183, 2021.

[17] L. Fan, C. Y. Lin, and S. Jha, "Differential privacy with bounded adversaries," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1874–1889, 2020.

[18] Y. Liu, J. Zhang, Y. Tang, and L. T. Yang, "Differential privacy budget allocation for federated learning: Challenges and opportunities," *IEEE Trans. Network Sci. Eng.*, vol. 9, no. 3, pp. 1723–1734, 2022.

[19] J. Yuan et al., "Approximate homomorphic encryption based privacy-preserving machine learning: A survey," *Artif. Intell. Rev.*, vol. 58, no. 3, 2025.

[20] H. Zhang et al., "Lattice-based multi-key homomorphic encryption scheme without common random strings," *Symmetry*, vol. 17, no. 5, p. 722, 2025.

[21] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statistics*, 2017, pp. 1273–1282.

[22] V. Kumar and C. Lakshmi, "An efficient secure computation for privacy preserving data mining in multi party computation (MPC) – A review," *Int. J. Adv. Res. Eng. Technol. (IJARET)*, vol. 11, no. 10, pp. 855–870, 2020.

ISSN: 2961-5992

[23] E. Boyle et al., "Function secret sharing for mixed-mode and fixed-point secure computation," IACR Cryptology ePrint Archive, 2020. [Online]. Available: https://eprint.iacr.org/2020/1392
[24] G. Premi et al., "Secure multi-party computation for privacy preservation in collaborative networks," in *Proc. 2025 Int. Conf. Automation and Computation (AUTOCOM)*, 2025, pp. 637–642.
[25] D. R. Kale et al., "Federated learning for privacy-preserving data mining," in *Proc. 2024 Int. Conf. Intelligent Systems and Advanced Applications (ICISAA)*, 2024, pp. 1–6.
[26] J. Tyagi and M. Sharma, "Enhancing privacy-preserving data mining in cloud-based data warehouses: A federated learning approach for secure multi-tenant environments," in *Proc. 2025 IEEE 10th Int. Conf. Smart Cloud (SmartCloud)*, New York City, NY, USA, 2025, pp. 8–13.
[27] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Machine Learning and Knowledge Discovery in Databases*, Springer, 2020, pp. 429–450.
[28] Y. Lindell, "Secure multiparty computation," *Commun. ACM*, vol. 64, no. 1, pp. 86–96, 2020.
[29] N. Kaaniche and M. Laurent, "A blockchain-based data usage auditing architecture with enhanced privacy and availability," in *Proc. 2017 IEEE 16th Int. Symp. Network Computing and Applications (NCA)*, 2017, pp. 1–5.
[30] A. Kiran, P. Rubini, and S. S. Kumar, "Comprehensive review of privacy, utility and fairness offered by synthetic data," *IEEE Access*, p. 1, 2025.
[31] S. Lu et al., "Class-imbalance privacy-preserving federated learning for decentralized fault diagnosis with biometric authentication," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 9101–9111, 2022.
[32] P. Mohassel and Y. Zhang, "SecureML: A system for scalable privacy-preserving machine learning," in *Proc. IEEE Symp. Security and Privacy (SP)*, 2017, pp. 19–38.
[33] J. He et al., "Exploring the limits of differentially private deep learning with group-wise clipping," arXiv preprint, 2022. [Online]. Available: https://doi.org/10.48550/arxiv.2212.01539
[34] D. Han et al., "A blockchain-based auditable access control system for private data in service-centric IoT environments," *IEEE Trans. Ind. Informat.*, vol. 18, no. 5, pp. 3530–3540, 2021.
[35] Y. Ding et al., "Delving into differentially private transformer," arXiv preprint, 2024. [Online]. Available: https://doi.org/10.48550/arxiv.2405.18194
[36] L. Gao, "Enterprise internal audit data encryption based on blockchain technology," *PLoS ONE*, vol. 20, no. 1, p. e0315759, 2025.
[37] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the GAN: Information leakage from collaborative deep learning," in *Proc. 2017 ACM SIGSAC Conf. Computer and Communications Security*, 2017, pp. 603–618.
[38] A. Kotal, S. S. L. Chukkapalli, and A. Joshi, "Synthetic data for privacy preservation in distributed data analysis systems," in *Springer Optimization and Its Applications*, 2024, pp. 393–407.
[39] Y. Zhou et al., "Comprehensive review of differential privacy in deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 6, pp. 3456–3470, 2025.
[40] H. Zhang, Y. Xiao, H. Chen, and J. Wu, "A survey on privacy-preserving data publishing using generalization and suppression," *IEEE Access*, vol. 8, pp. 36217–36236, 2020.
[41] X. Gao, "Enterprise internal audit data encryption based on blockchain technology," *PLoS ONE*, vol. 20, no. 1, p. e0315759, 2025.
[42] L. Yu et al., "Differentially private model publishing for deep learning," in *Proc. 40th IEEE Symp. Security and Privacy (IEEE S&P 2019)*, 2019.
[43] J. Jia, Y. Zhang, N. Z. Gong, and P. Mittal, "Privacy risks of general-purpose language models," arXiv preprint arXiv:2110.08205, 2021.
[44] U. O. Nnaji et al., "A review of strategic decision-making in marketing through big data and analytics," *Magna Scientia Adv. Res. Rev.*, vol. 11, no. 1, pp. 084–091, 2024.
[45] I. Damgård, M. Geisler, and M. Krøigaard, "Homomorphic encryption and secure comparison for multiparty computation," *Int. J. Appl. Cryptography*, vol. 2, no. 1, pp. 22–31, 2012.
[46] M. Gaboardi, H. Lim, R. Rogers, and S. Vadhan, "Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing," in *Proc. ICML 2016*, 2016.
[47] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theoretical Computer Sci.*, vol. 9, no. 3–4, pp. 211–407, 2014.
[48] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol. (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.
[49] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.
[50] K. Bonawitz et al., "Practical secure aggregation for federated learning on user-held data," in *NIPS Workshop on Private Multi-Party Machine Learning*, 2017.
[51] O. Farràs and D. Rebollo-Monedero, "Optimal trade-off between information leakage and utility in data anonymization," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 12, pp. 2743–2755, 2016.
[52] O. Oyedokun, S. E. Ewim, and O. P. Oyeyemi, "Leveraging advanced financial analytics for predictive risk management and strategic decision-making in global markets," *Global J. Res. Multidisciplinary Studies*, vol. 2, no. 2, pp. 016–026, 2024.
[53] D. Nugroho and P. Angela, "The impact of social media analytics on SME strategic decision making," *IAIC Trans. Sustainable

ISSN: 2961-5992