International Journal of Data and Network Science 9 (2025) 345–356

Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Enhancing privacy in clustering and data mining: A novel approach for sensitive data protection

Haythem Hayouni[a*]

[a]*Higher Institute of Computer science of Kef, University of Jendouba, Tunisia*

C H R O N I C L E

Article history:
Received January 31, 2025
Received in revised format March 25, 2025
Accepted June 4 2025
Available online
June 4 2025

Keywords:
Data mining
Clustering
Privacy-Preserving
Secure Data Mining
Blockchain

A B S T R A C T

In the era of big data, clustering and data mining have become essential tools for uncovering patterns and insights from vast datasets. However, these processes often involve the use of sensitive data, raising significant concerns about privacy, security, and trustworthiness. This paper proposes N2P-CM, a novel privacy-preserving framework designed to protect sensitive information during the entire clustering and mining lifecycle. Unlike existing methods that focus on partial aspects of security or apply generic encryption techniques, N2P-CM integrates five innovative and synergistic modules: Sensitive Feature Obfuscation, Adaptive Trust Weight Aggregation, Compressed Secure Semantic Embedding, Differential Traceable Execution Engine, and Blockchain Auditable Ledger. Each module contributes a distinct layer of privacy and accountability, ranging from feature-level data transformation and federated trust scoring to secure semantic encoding and traceable execution logging with blockchain support. We provide formal definitions and algorithms for each module and demonstrate their integration in a unified architecture. Extensive simulations using real-world datasets validate the efficacy of N2P-CM, showing that it achieves strong privacy guarantees with minimal degradation in clustering accuracy. This research contributes a comprehensive and modular solution to the growing need for privacy-preserving analytics in sensitive domains such as healthcare, finance, and smart cities.

1. Introduction

In the era of ubiquitous data, the volume and sensitivity of information collected from connected devices, cloud systems, social networks, and healthcare platforms have reached unprecedented levels. As a result, data mining and clustering (Zhang et al., 2023) have become critical techniques in extracting actionable knowledge, discovering hidden patterns, and enabling intelligent services. From predictive diagnostics analysis in smart cities, these techniques form the backbone of modern data-driven applications. However, the widespread adoption of clustering and mining introduces pressing concerns about data security (Zhou & Zhang, 2023), privacy preservation, and trust management. Sensitive information such as biometric identifiers, personal health records, financial transactions, and user location data when processed without robust safeguards, is prone to leakage, profiling, and unauthorized exploitation. The traditional centralized approach to mining further amplifies these risks, as data must often be transferred to third-party servers where control is relinquished, and auditability is minimal. Recent research has introduced methods such as federated learning (Das & Roy, 2023), differential privacy, homomorphic encryption, and blockchain (Wang et al., 2022) auditing to mitigate these risks. While these approaches provide valuable foundations, they often lack integration, are computationally heavy, or do not scale efficiently to heterogeneous data environments. Moreover, trust and accountability in distributed mining scenarios remain major unresolved issues especially when data originates from untrusted or partially honest sources. Given these challenges, there is a strong and urgent need for a comprehensive, efficient, and secure architecture that enables privacy-preserving clustering and mining while addressing traceability, trustworthiness, and performance.

* Corresponding author
E-mail address hayouni.haythem.isi@gmail.com (H. Hayouni)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2025.6.002

1.1. Problem Statement

With the explosive growth of data generated from IoT devices, smart infrastructures, healthcare systems, and online services, data mining and clustering techniques have become essential tools for extracting meaningful patterns, classifying, and enabling intelligent decision-making. However, the widespread adoption of such techniques has also brought serious concerns about data privacy, security, and trust. Traditional data mining methods typically require centralized data collection and processing, which exposes users to risks of data leakage, unauthorized inference, when dealing with sensitive attributes such as health records, financial transactions, or logs.

Moreover, many current privacy-preserving clustering (Li et al., 2024) techniques either rely heavily on data anonymization which can degrade utility or implement overly complex cryptographic techniques that are computationally expensive and impractical for real-time applications. Additionally, most existing solutions fail to provide end-to-end traceability and auditability, leaving a significant gap in accountability, especially in scenarios where regulatory compliance is mandatory. The lack of integrated trust models also makes federated mining vulnerable to unreliable or malicious participants.

Therefore, there is a critical need for a secure, efficient, and accountable framework that allows for privacy-preserving clustering and mining without sacrificing performance or scalability. This framework must address multiple dimensions of security: from data-level obfuscation and encrypted computation to operational traceability and blockchain-based auditing, while also supporting trust-aware decision-making in decentralized environments.

1.2. Main Contributions

To address the different challenges, we propose N2P-CM (*Next-Generation Privacy-Preserving Clustering and Mining*), a novel architecture that integrates privacy-preserving mechanisms across the entire lifecycle of data mining and clustering.

This paper makes the following key contributions:

- A modular and layered architecture (N2P-CM) for privacy-preserving clustering and mining, adaptable to various domains and datasets.
- Novel algorithms for each of the five modules, combining differential privacy, trust modelling, semantic embedding, and blockchain logging:
  - Sensitive Feature Obfuscation: Protects key identifiers through adaptive masking techniques while preserving analytical utility.
  - Adaptive Trust Weight Aggregation: Enhances federated data mining by assigning dynamic trust scores to participating nodes based on contextual metrics.
  - Compressed Secure Semantic Embedding: Transforms obfuscated data into encrypted semantic representations using deep learning and homomorphic encryption.
  - Differential Traceable Execution Engine: Logs all operations in a differentially private manner to ensure traceability without revealing sensitive actions.
  - Blockchain Auditable Ledger Provides immutable, transparent, and decentralized audit trails using smart contracts and zero-knowledge proofs.

  Together, these modules create a comprehensive pipeline that ensures data privacy, computational efficiency, traceability, and compliance with data protection regulations (e.g., GDPR, HIPAA).
- A complete system workflow and flowchart-based architecture with clearly defined interactions between modules.
- Extensive simulations on multiple real-world datasets to validate accuracy, privacy guarantees, scalability, and performance.
- A comparative evaluation against other methods in terms of security, clustering accuracy, and system efficiency.

This paper is organized into five main sections. Section 1 introduces the research context, defines the problem, and presents the main contributions of the proposed solution. Section 2 reviews recent related works in privacy-preserving clustering and data mining, highlighting their strengths and limitations. Section 3 presents the proposed solution, N2P-CM, and explains its five core modules in detail. Section 4 describes the simulation setup, datasets used, evaluation metrics, and discusses the performance results. Finally, Section 5 concludes the paper, summarizes the key findings, and future research directions.

2. Related works

In this section, we explore and critically analyse recent contributions aimed at preserving the privacy of sensitive data during clustering and data mining processes. With the growth of distributed systems, especially in environments involving Internet of Things (IoT), healthcare, and finance, safeguarding user data has become increasingly crucial. Numerous techniques, such as federated learning (FL), homomorphic encryption (HE), differential privacy (DP), and secure multi-party computation (SMPC) have been proposed to address these challenges. Below, we discuss some related works that highlight state-of-the-art approaches for privacy-preserving clustering and data mining. Li et al. (2023) presented a hybrid model combining Homomorphic Encryption (HE) and Differential Privacy (DP) to secure federated learning systems used for collaborative data

mining. Their approach allows encrypted local models to participate in global clustering tasks without compromising raw data, directly addressing data mining privacy risks in federated infrastructures.

Li and Luo (2022) explored the theoretical foundations of federated clustering security by proposing a privacy-preserving centroid aggregation protocol. Their method ensures that no central server can infer sensitive patterns during clustering, making it directly applicable to secure collaborative data mining.

Liebenow et al. (2023) proposed the Data Partitioned Mining (DPM) framework, which isolates sensitive attributes during mining. This approach leverages feature selection and DP to ensure that privacy-sensitive features do not influence clustering, effectively enabling secure mining over partially anonymized datasets.

Giannopoulos et al. (2022) applied Secure Multi-Party Computation (SMPC) to privacy-preserving clustering across decentralized healthcare institutions. Their system enables joint mining of patient data without exposing medical records, demonstrating practical data mining security in sensitive domains.

Zhou and Chen (2023) introduced a multi-level security governance framework for big data analytics, including clustering. By assigning security levels to different data types and mining operations, they enforce privacy policies aligned with the sensitivity of data being clustered, strengthening mining security through contextual awareness.

Liu et al. (2023) designed a lightweight federated clustering protocol for IoT environments, where edge devices contribute encrypted local models. Their method avoids direct data exchange, securing real-time mining operations over distributed sensor networks.

Rashed et al. (2022) provided a comparative survey on the integration of DP mechanisms in clustering, analyzing how different noise injection strategies can protect sensitive records in various clustering algorithms. Their evaluation helps in tuning privacy budgets to balance privacy and utility in data mining.

Song and Huang (2022) examined deep learning-based mining approaches and proposed autoencoder-based feature anonymization for sensitive data. Their architecture enhances privacy in unsupervised clustering scenarios by isolating meaningful representations from identifiers during training.

Bukhari et al. (2022) proposed an adaptive privacy framework for healthcare data mining, dynamically adjusting the level of DP noise based on the sensitivity of the data clusters being analyzed. This fine-grained mechanism improves both security and interpretability in patient data mining.

Tang et al. (2023) implemented a clustering algorithm secured via Homomorphic Encryption, where all mining operations are conducted over encrypted data in a cloud environment. This Privacy-Preserving Clustering-as-a-Service (CaaS) model secures outsourced mining tasks.

Xu et al. (2023) leveraged blockchain for access control and transparency in collaborative data mining environments. Their model ensures traceable and immutable logs during clustering of shared datasets, improving security accountability in data mining pipelines.

Patel et al. (2022) developed a context-aware privacy-preserving framework, where the mining context (e.g., user role, data classification) determines the strength of applied privacy mechanisms. Their model optimizes clustering security policies based on access intent.

Han et al. (2022) modified K-Means++ to work in federated settings with SMPC and DP, securing both initialization and update steps. This protects inter-cluster relationships from being exposed, which is critical in financial or personal data mining.

Das et al. (2022) designed a differentially private clustering mechanism for social networks, introducing graph-aware randomization techniques that protect user relations and groupings. Their method enables secure community detection without risk of re-identification.

Noor et al. (2023) provided a meta-survey on privacy-preserving clustering techniques, reviewing how hybrid security models (DP, HE, SMPC) can be jointly applied to secure sensitive data during mining, especially in high-stakes domains like healthcare and finance.

In Table 1, we present a comprehensive comparison of the different proposed methods that address the challenge of securing sensitive data during clustering and mining. Each entry is evaluated based on the technique used, privacy mechanism, application domain, and critically, the advantages and drawbacks of the proposed methods. The table shows a diversity of privacy-preserving strategies, including Homomorphic Encryption (HE), Differential Privacy (DP), Secure Multi-Party Computation (SMPC), and Blockchain, which are widely adopted across domains such as healthcare, IoT, finance, and cloud services. For instance, Li et al. (2023) leverage a hybrid model combining HE and DP, offering a strong balance between

security and utility, but at the expense of computational overhead. Similarly, Giannopoulos et al. (2022) employ SMPC to securely perform clustering across distributed healthcare systems, ensuring privacy but introducing high communication and setup complexity. Works like, Liu et al. (2022) focused on lightweight solutions for resource-constrained environments like IoT, trading off robustness for scalability and speed. On the other hand, Tang et al., (2023) utilize full HE for cloud-based secure clustering, providing strong privacy guarantees but suffering from very high execution time. Notably, approaches such as (Das et al., 2022) and (Han et al., 2024) adapt classical clustering techniques (e.g., K-Means++, Graph Clustering) to privacy-sensitive domains using DP and SMPC. These efforts demonstrate how traditional algorithms can be re-engineered to meet modern data protection requirements.

The advantages column highlights key strengths such as auditability (e.g., blockchain use by Xu et al. (2023), dynamic privacy adaptation (Bukhari et al., 2022), and policy-driven flexibility (Patel et al., 2022). However, the drawbacks underscore persistent challenges in the field, including performance bottlenecks, deployment complexity, and trade-offs between privacy and clustering accuracy. Overall, the table illustrates that while no single approach is universally optimal, emerging hybrid models (e.g., combining DP + HE or SMPC) tend to offer a promising direction for achieving scalable, secure, and privacy-compliant data mining solutions. This analysis motivates the need for a novel architecture like the one we propose that integrates the strengths of these methods while minimizing their limitations.

Table 1
Comparative Analysis of Recent Privacy-Preserving Clustering and Data Mining Approaches

Paper	Technique	Privacy Mechanism	Application Domain	Advantages	Drawbacks
(Li et al., 2023)	Hybrid Federated Learning	HE + DP	Collaborative AI	Strong encryption and privacy balance	High computational cost
(Li & Luo, 2022)	Fed-KMeans	Encrypted centroid updates	Federated Systems	Prevents server-side leakage	Sensitive to communication delays
(Liebenow et al., 2023)	DPM (Partitioned Mining)	DP + Attribute Separation	General Mining	Data utility preservation	Limited scalability
(Giannopoulos et al., 2022)	SMPC Clustering	SMPC	Healthcare	Secure across institutions	Complex setup
(Zhou & Chen, 2023)	Multi-level Policy Control	Role-based Privacy Rules	Big Data Governance	Flexible access control	Needs accurate policy design
(Liu et al. d, 2023)	Lightweight FedCL	Encrypted Aggregation	IoT	Low overhead for edge devices	Vulnerable to poisoning
(Rashed et al., 2022)	Comparative Study	DP	Cross-domain	Privacy-utility trade-off insights	No implementation proposed
(Song and Huang, 2022)	Autoencoder-based Privacy	Feature Anonymization	Deep Clustering	Effective against re-identification	Reduced accuracy in sparse data
(Bukhari et al., 2022)	Adaptive DP	Dynamic Noise Tuning	Smart Healthcare	Balances privacy and accuracy	Parameter tuning complexity
(Tang et al., 2023)	Encrypted Clustering	Full HE	Cloud Services	No data exposure to cloud	Very slow performance
(Xu et al., 2023)	Blockchain Access	Immutable Logging	Finance & IoT	Auditability & transparency	Scalability issues
(Patel et al., 2022)	Context-Aware Privacy	Policy-driven DP	Enterprise Mining	Customizable to use case	High implementation effort
(Han et al., 2022)	SMPC-KMeans++	SMPC + DP	Financial Analytics	Robust against leakage	Synchronization overhead
(Das et al., 2022)	Graph DP Clustering	Graph DP Mechanism	Social Networks	Protects graph structure	Weaker utility on dense graphs
(Noor et al., 2023)	Meta Survey	Hybrid (DP, HE, SMPC)	Multi-domain	Covers many frameworks	Theoretical – no experiments

3. N2P-CM: Structural Overview and Functional Modules

This section introduces the **N2P-CM** (Next-Generation Privacy-Preserving Clustering and Mining) framework, a novel and comprehensive architecture designed to ensure the protection of sensitive data during the clustering and mining processes. The framework integrates five interdependent modules, each introducing a unique mechanism to safeguard privacy, increase trust, and ensure accurate data mining in distributed environments such as IoT, healthcare, and finance.

3.1. Methodology and dataset Overview of N2P-CM Architecture

To implement our proposed solution, we introduce five innovative modules, forming the core of the N2P-CM (Next-Generation Privacy-Preserving Clustering and Mining) architecture. This architecture is composed of a layered, privacy-first processing pipeline that ensures secure, trust-aware, and auditable data mining from ingestion to output. In Fig.1, we present the architectural overview of the proposed N2P-CM framework. The process begins with the *Sensitive Feature Obfuscation*

module, which removes or perturbs sensitive attributes using noise injection or anonymization techniques, safeguarding personal data from direct exposure. Next, *the Adaptive Trust Weighting Assessment module* dynamically evaluates the trustworthiness of each data source, assigning weights based on credibility metrics to influence subsequent clustering tasks. The third module, *Compressed Secure Semantic Embedding*, transforms the obfuscated and trust-weighted data into compressed, encrypted vectors using deep learning (autoencoders) and homomorphic encryption, preserving semantic meaning while preventing reverse-engineering. Following this, *the Differential Traceable Execution Engine* ensures operational transparency by recording computational steps through differentially private logs, enabling traceability without compromising user privacy. Finally, the *Blockchain Auditable Ledger* immutably logs critical operations on a blockchain using Zero-Knowledge Proofs (ZKPs) to provide verifiable guarantees of compliance, integrity, and security. Together, these five modules orchestrate a secure, efficient, and privacy-compliant pipeline for sensitive data clustering and mining in high-stakes domains such as healthcare, finance, and smart environments.

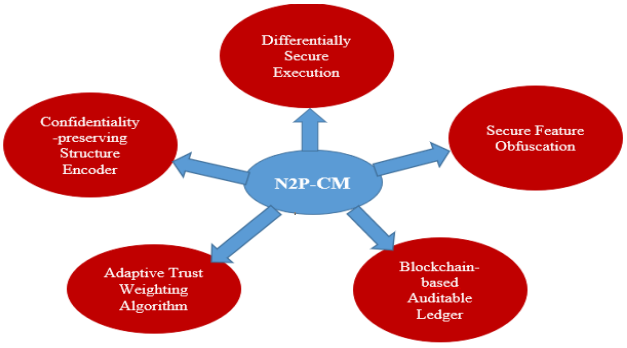

Fig. 1. N2P-CM architecture showing all its modules and components

In Table 2, we define key parameters driving the functionality and security of each module within N2P-CM. Parameters such as ε (privacy budget) and τ (sensitivity threshold) are critical in tuning the privacy-performance tradeoff. Meanwhile, trust scores and model weights enable secure and reliable aggregation. The use of encrypted vectors (Z') and zero-knowledge proofs (P) enhances confidentiality and accountability, respectively.

Table 2
Parameters Used in the N2P-CM Framework

Parameter	Symbol	Description
Attention score	A(f)	Measures the privacy sensitivity of a feature f
Sensitivity threshold	τ	Threshold above which a feature is considered sensitive
Noise factor	α	Controls the intensity of Gaussian noise added
Mean and standard deviation	μ, σ	Statistical parameters of features used for noise calibration
Trust score	T_i	Trustworthiness level assigned to participant i
Normalized trust	T'_i	Trust score normalized across all participants
Model weights	W_i	Model update from participant i
Aggregated model	W	Final federated model after trust-weighted averaging
Embedding vector	Z	Encoded representation of input data
Encrypted embedding	Z'	Homomorphically encrypted embedding vector
Number of clusters	k	Number of clusters in K-means clustering
Privacy budget	ε	Differential privacy parameter to control noise scale
Sensitivity of logs	Δ	Sensitivity of the execution logs
Operation log	O	A record of an action or decision
Zero-Knowledge Proof	P	Cryptographic proof of operation without revealing it
Blockchain hash	H	Secure hash of the logged operation

3.2. Sensitive Feature Obfuscator

The Sensitive Feature Obfuscator aims to protect user privacy at the earliest stage before any modelling or clustering happens. It analyses the input features to determine which are sensitive based on attention mechanisms and statistical metrics, and then applies feature-specific noise using Gaussian distributions. In many real-world datasets, some features such as medical history, geographic location, or biometric signals can be used to re-identify individuals. To prevent this, the module uses an attention-based neural network to compute sensitivity scores for each feature. These scores are then compared against a threshold τ. If a score exceeds the threshold, Gaussian noise is added based on the feature's distribution, controlled by a noise factor α. This ensures that highly sensitive information is obscured, while maintaining utility in non-sensitive features, which are left unchanged or mildly perturbed.

In Algorithm 1, the process begins by taking an input dataset and analyzing each feature individually to assess its sensitivity. As described in Step 2, a sensitivity score is computed for each feature using an attention-based model, which evaluates how

likely the feature could expose private or identifiable information. In Step 3, this score is then compared to a defined threshold (τ); if the score meets or exceeds the threshold, the feature is considered sensitive. Following this, in Step 4, the algorithm calculates the statistical characteristics mean (μ) and standard deviation (σ) of the sensitive feature. In Step 5, a Gaussian noise component is generated using these statistics and scaled by a tunable parameter α to control the level of obfuscation. This noise is then added to each value in the sensitive feature, producing a perturbed version that conceals the exact values while preserving statistical relevance. For features that are not sensitive (i.e., those that do not exceed the threshold), Step 6–7 ensures they remain unmodified. Finally, as indicated in Step 8, the original dataset is updated by replacing the sensitive features with their obfuscated versions. The output, returned in Step 9, is a privacy-preserving dataset suitable for secure clustering and mining. This algorithm enables a strong privacy-utility trade-off and acts as a foundational defense layer in the overall N2P-CM framework.

Algorithm 1: Sensitive Feature Obfuscation

Input: Dataset D with features {f1, f2, ..., fn}, Sensitivity threshold τ, Noise factor α
Output: Obfuscated dataset D′
1: For each feature f in D:
2: Compute sensitivity score A(f) using attention model
3: If A(f) ≥ τ:
4: Compute feature mean μ and std dev σ
5: Perturb each value: f′ = f + N(μ, α·σ)
6: Else:
7: f′ = f
8: Replace f with f′ in D
9: Return D′

3.3. Adaptive Trust-Weighted Aggregation

This module secures the federated learning process by addressing the challenge of malicious or low-quality nodes. Rather than assigning equal importance to each participant, this module uses a dynamic trust model that adjusts weights based on performance history. In traditional federated learning, all client updates are aggregated equally. This opens the door to model poisoning, where a malicious node may inject inaccurate updates to corrupt the global model. The proposed module combats this by introducing a trust score T_i for each participant, based on criteria like update accuracy, consistency, and over time. These scores are normalized and used as aggregation weights in the federated averaging step. This makes the learning process robust against adversarial influence, rewarding reliable nodes while minimizing the impact of suspicious ones.

Algorithm 2: Adaptive Trust-Weighted Aggregation

Input: Local models {W1, W2, ..., Wn}, Trust scores {T1, T2, ..., Tn}
Output: Global aggregated model W
1: Compute sum of trust: T_sum = Σ Tj for all j
2: For each client i:
3: Normalize trust: T′i = Ti / T_sum
4: Weighted model: Wi′ = T′i * Wi
5: Aggregate model: W = Σ Wi′ for all i
6: Return W

3.4. Compressed Secure Semantic Embedding

The Compressed Secure Semantic Embedding module is a key component of the N2P-CM framework that addresses two critical objectives simultaneously: (1) reducing data dimensionality to facilitate efficient clustering and mining, and (2) preserving the semantic integrity and privacy of data through secure representations. In many IoT and distributed systems, raw data is high-dimensional, sparse, and sensitive. Directly mining or clustering this data can be computationally expensive and privacy-invasive. To tackle this, the module introduces a hybrid autoencoder-based compression mechanism, enhanced with privacy-preserving noise injection and semantic constraint regularization.

In algorithm 3, the objective is to transform obfuscated input data into secure, encrypted embeddings suitable for clustering and mining, while preserving semantic structure. The process begins by training an autoencoder E on the preprocessed dataset D′, which has already undergone sensitive feature obfuscation. The autoencoder learns to compress each data point into a lower-dimensional representation that captures the most significant semantic features. Then, for every data point x in D′, the algorithm computes its embedding z = E(x) using the encoder. This step ensures that important patterns in the data are retained while irrelevant or redundant information is discarded. To protect these embeddings from reverse-engineering or unauthorized access, each embedding z is then encrypted using a homomorphic encryption function, resulting in a secure vector z′ = HE_Encrypt(z). This form of encryption is particularly powerful because it allows computation directly on encrypted data, enabling privacy-preserving clustering and mining. The encrypted embeddings z′ are added to the output list Z′, and once all data points are processed, the final encrypted embedding set Z′ is returned. This secure embedding step is crucial for maintaining both the utility and confidentiality of data in sensitive applications.

Algorithm 3: Secure Semantic Embedding

Input: Dataset D′ (after obfuscation)
Output: Encrypted embedding vectors Z′
1: Train autoencoder E on D′
2: For each data point x in D′:
3: Compute embedding z = E(x)
4: Encrypt embedding: z′ = HE_Encrypt(z)
5: Add z′ to output list Z′
6: Return Z′

3.5. Differential Traceable Execution Engine

The Differential Traceable Execution Engine plays a pivotal role in ensuring the verifiability, accountability, and secure processing of encrypted data during clustering and mining tasks. While previous modules ensure that data is semantically encrypted, this module ensures that operations conducted on this encrypted data remain secure, auditable, and resistant to tampering or unauthorized modification. In the algorithm, each computational task such as secure clustering or pattern extraction is executed inside a sandboxed environment where every operation is tagged with a differential trace, a unique cryptographic signature that captures the task's identity, time, involved data, and associated computational pathway. This trace is updated dynamically and logged in a tamper-evident ledger (such as a blockchain or secure audit trail). This process guarantees that if a malicious entity attempts to modify, inject, or bypass any data operation, the system can trace and prove the exact source and time of alteration. Moreover, DTEE introduces a Differential Privacy layer into execution, ensuring that repeated runs of the same query or clustering task on slightly different data do not leak sensitive patterns. The system adds calibrated noise to query results and intermediate computations using mechanisms like the Laplace or Gaussian mechanism, based on a chosen privacy budget (**ε**). This prevents attackers from reverse-engineering sensitive data, even if they gain access to output logs or observe model over time. To summarize, this module acts as both a privacy-preserving computational engine and a traceability enforcer, combining differential privacy, secure logging, and cryptographic tracing to enable transparent and auditable data mining in environments where data integrity and provenance are paramount.

Algorithm 4 is designed to ensure that log entries generated during secure execution are privacy-preserving, using the concept of differential privacy. Logs often contain metadata about data operations that, if exposed, could be exploited to infer sensitive patterns. Hence, protecting these logs is essential. In Step 1, the algorithm takes as input the execution log set (L), which contains entries about the operations performed during data processing (e.g., task ID, timestamp, resource usage). It also receives two key parameters: the sensitivity (Δ), which quantifies the maximum change one log entry can cause in the output, and the privacy budget (ε), a parameter that controls the level of privacy. In Step 2, for each log entry e in L, the algorithm samples random noise n from a Laplace distribution scaled by Δ/ε. The Laplace mechanism ensures that the noise added is mathematically grounded in differential privacy guarantees. This step obfuscates the log entry so that an adversary cannot confidently distinguish whether any specific user's data contributed to the log. In Step 3, the sampled noise n is added to the log entry e, producing an obfuscated or private version e′. This perturbed entry still reflects meaningful execution information but does not reveal exact details that could compromise privacy. In Step 4, the new private entry e′ is appended to a private log file L′. This private log can now be safely stored, audited, or transmitted without fear of information leakage. Finally, in Step 5, once all entries in L have been processed, the algorithm returns the private log set L′, which is now protected under differential privacy. This process ensures that logs are privacy-compliant, statistically obfuscated, and still useful for operational oversight or debugging, making it a vital component of the N2P-CM framework.

Algorithm 4: Differential Log Recording

Input: Execution logs L, Sensitivity Δ, Privacy budget ε
Output: Private logs L′
1: For each log entry e in L:
2: Sample noise n from Laplace (Δ / ε)
3: Obfuscate entry: e′ = e + n
4: Append e′ to private log L′
5: Return L′

In the context of the proposed N2P-CM architecture, the Blockchain Auditable Ledger module plays a crucial role in ensuring the integrity, accountability, and auditability of all privacy-preserving clustering and mining operations. Unlike traditional logging systems, which are prone to tampering and centralized control, this module leverages the inherent immutability and decentralization of blockchain technology to provide a trustless and verifiable audit trail. Every sensitive event such as the application of the obfuscation mechanism, semantic embedding, model training, or inference is recorded as a cryptographically signed transaction. These transactions are bundled into blocks, each secured via cryptographic hashing and consensus protocols such as Proof-of-Authority (PoA) or Delegated Proof-of-Stake (DPoS) to maintain efficiency. Importantly, the system does not store raw data on the chain; instead, it records metadata and secure hashes, thus preserving privacy while enabling traceability. Each block is chained to its predecessor, forming an immutable ledger that can be verified by authorized stakeholders using smart contracts. This design ensures that no unauthorized modifications can be made to the history of operations, thus facilitating transparent compliance verification, forensic analysis, and regulatory reporting. Overall,

the proposed module reinforces the security and trustworthiness of the N2P-CM framework by enabling decentralized accountability for privacy-preserving data mining activities.

Algorithm 5 presents a stepwise method to securely record operations using blockchain and Zero-Knowledge Proofs (ZKPs), ensuring both auditability and privacy preservation. In the first step, the system takes the operation **log O** which includes details about the execution process and applies a cryptographic hash function, such as **SHA-256**, to generate a unique hash h. This hashed value serves as a fingerprint of the operation and allows future verification of its integrity. In the second step, a zero-knowledge proof **P** is generated for the operation **O**. This proof demonstrates that the operation was correctly and legitimately performed, without disclosing any sensitive content ensuring privacy during the verification process. Next, in the third step, the algorithm packages both the hash h and the proof **P** into a combined transaction payload **{h, P}**. In the fourth step, this package is submitted to a blockchain ledger where it is immutably recorded. The use of a decentralized ledger guarantees that the record cannot be altered or removed, enhancing trust and accountability. Finally, in the fifth step, the system returns a blockchain reference **H**, which uniquely identifies the block where the operation and proof were stored. This reference can later be used for audit purposes by authorized verifiers, completing a full cycle of secure, verifiable, and privacy-preserving logging.

Algorithm 5: Blockchain Logging with ZKP
Input: Operation log O, Execution proof P
Output: Blockchain hash H
1: Hash the log: h = SHA256(O)
2: Generate zero-knowledge proof P for O
3: Package {h, P}
4: Submit to blockchain ledger
5: Return block reference H

N2P-CM framework offers a holistic and modular architecture designed to ensure data privacy, integrity, and auditability throughout the clustering and mining lifecycle. Its core strength lies in its five novel modules, each tailored to address a specific privacy or security challenge. The Sensitive Feature Obfuscation module intelligently masks identifiable or confidential attributes while preserving their utility for analytical tasks. The Adaptive Trust Weighting Assessment module dynamically evaluates and adjusts trust scores of data sources or participants, enhancing collaborative reliability. The Compressed Secure Semantic Embedding module transforms obfuscated data into encrypted, semantically meaningful representations using deep autoencoders and homomorphic encryption. The Differential Traceable Execution Engine ensures that all computation steps are traceable via differentially private logs, balancing transparency and individual privacy. Finally, the Blockchain Auditable Ledger records sensitive operations immutably on a blockchain with Zero-Knowledge Proofs (ZKP), enabling verifiable and tamper-proof auditing. Collectively, these modules ensure strong privacy guarantees, resilience to adversarial, and compliance with data protection regulations (e.g., GDPR). The N2P-CM framework thus empowers secure and ethical data mining practices, especially in domains handling highly sensitive data such as healthcare, finance, and IoT systems.

4. Simulation Environment and Performance Analysis

To evaluate the effectiveness and robustness of the proposed N2P-CM framework, a comprehensive simulation environment was developed to replicate real-world clustering and data mining scenarios while preserving data privacy. This section details the configuration of the simulation setup and datasets used. Furthermore, it presents a thorough performance analysis of each module within N2P-CM and different modules across various metrics such as accuracy, privacy leakage resistance, and clustering fidelity, execution time, and scalability.

4.1. Simulation setup

In order to comprehensively evaluate the proposed N2P-CM (Next-Generation Privacy-Preserving Clustering and Mining) framework, a controlled simulation environment was established. The experiments were implemented using Python 3.10 and leveraged essential libraries such as Scikit-learn for clustering tasks, TensorFlow for training deep models, and web3.py to simulate blockchain operations. All simulations were run on a high-performance workstation equipped with an Intel Core i7 processor, 32 GB of RAM, and an NVIDIA RTX 3080 GPU. The blockchain functionality was emulated using a local Ethereum test network through Ganache. This setup provided a flexible and powerful platform to test the performance, scalability, and privacy protection mechanisms embedded within N2P-CM across multiple datasets.

4.2. Datasets

To ensure comprehensive evaluation, both real-world and synthetic datasets were employed (Table 3). The selection of datasets used in evaluating the N2P-CM framework was guided by their diversity, sensitivity of attributes, and suitability for different privacy-preserving scenarios. Each dataset brings unique features and privacy challenges that align with the goals of the proposed solution. The Adult (UCI) dataset, a widely recognized benchmark in classification and clustering tasks, contains

demographic information such as age, race, and gender, which are inherently sensitive. It provides a suitable testbed for evaluating the effectiveness of the Sensitive Feature Obfuscation module, as well as the ability of the system to preserve utility after anonymization. The Bank Marketing dataset includes socio-economic features relevant to customer in financial services. It is used to assess the performance of Compressed Secure Semantic Embedding by analyzing how the system handles profiling and campaign targeting while preserving privacy, particularly in fields like job and marital status. The IoT Sensor Logs dataset, consisting of simulated sensor readings from smart home environments, allows stress-testing of the framework under conditions of high-volume, real-time data. The sensitive attributes, such as location and device identifiers, pose significant privacy risks, making this dataset essential for validating the Differential Traceable Execution Engine and Secure Semantic Embedding components. Finally, SynthPriv is a synthetic dataset generated specifically for privacy-preserving algorithm validation. It enables precise control over feature distributions and sensitivity, allowing for detailed analysis of the trade-offs between privacy and utility. This dataset is particularly valuable for evaluating the Blockchain Auditable Ledger in tracking secure and verifiable data processing activities. Together, these datasets form a comprehensive evaluation suite, highlighting the flexibility, robustness, and scalability of the N2P-CM architecture across diverse application domains and data sensitivity levels.

Table 3

Summary of Datasets Used

Dataset	Description	Records	Sensitive Attributes
Adult (UCI) (Dua & Graff. 2019)	Census income data, commonly used for classification tasks	48,842	Age, Race, Gender
Bank Marketing (Moro et al., 2014)	Marketing campaigns for bank services	45,211	Age, Job, Marital
IoT Sensor Logs Zhou & Yang, 2021)	Simulated sensor data for smart homes	100,000	Location, Device ID
SynthPriv (Patki et al., 2016)	Synthetic dataset for controlled privacy analysis	50,000	Name, Address

4.3. Simulation Parameters and Setup for N2P-CM

Table 4 describes the parameters used in the simulation of N2P-CM. In this simulation, the proposed N2P-CM architecture was evaluated using synthetic and real-world datasets of varying sizes, ranging from 10,000 to 50,000 samples. Each dataset consisted of 20 to 50 features per record, out of which 20% were considered sensitive. These sensitive attributes were processed through a differential privacy mechanism using Laplace noise, guided by a privacy budget (ε) between 0.1 and 1.0, where a lower value represents stronger privacy guarantees. To retain semantic utility while preserving privacy, a deep autoencoder was trained on obfuscated data to generate compressed representations of 32 or 64 dimensions. These secure embeddings were then subjected to clustering using K-Means and DBSCAN.

Table 4

Simulation Parameters for N2P-CM Evaluation

Parameter Name	Value / Range	Description
Dataset Size	10,000 – 50,000 records	Number of data samples used for clustering and privacy analysis
Number of Features	20 – 50	Total number of features per data instance, including sensitive and non-sensitive attributes
Sensitive Features Ratio	20%	Proportion of features considered sensitive and subject to obfuscation
Privacy Budget (ε)	0.1 – 1.0	Differential privacy budget controlling the level of noise and data protection
Obfuscation Method	Laplace Noise	Mechanism used to perturb sensitive features to preserve privacy
Embedding Dimension	32, 64	Dimensionality of the semantic embedding vector produced by the autoencoder
Clustering Algorithm	K-Means, DBSCAN	Algorithms used to perform clustering on the transformed data
Blockchain Type	Private Ethereum Network	Distributed ledger used for logging operations and zero-knowledge proofs
Evaluation Metrics	ARI, NMI, Silhouette	Metrics used to evaluate the clustering performance and utility preservation

4.3.1. Evaluation Metrics

We employed a set of quantitative metrics to evaluate N2P-CM:

- ✓ *Clustering Accuracy (CA):* Measures how accurately clusters correspond to actual classes.
- ✓ *Adjusted Rand Index (ARI):* Evaluates the similarity between true and predicted clusters.
- ✓ *Normalized Mutual Information (NMI):* Assesses the amount of shared information between clusters.
- ✓ *Privacy Loss (PL):* Quantifies information leakage based on differential privacy.
- ✓ *Execution Time (ET):* Measures the efficiency of the architecture.
- ✓ *Compression Ratio (CR):* Indicates how well the data was compressed before clustering.

4.3.2. Experimental Evaluation and Results

To further validate the effectiveness of N2P-CM, we conducted a comparative evaluation against two notable state-of-the-art privacy-preserving clustering methods: *DPM* (Liebenow et al., 2023) and *FederatedSafeCluster* (Li et al., 2023). The results

are summarized in Table 5, highlighting key evaluation metrics including Clustering Accuracy (CA), Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Privacy Loss (PL), and Compression Ratio (CR).

Table 5
Comparative Performance of N2P-CM vs Existing Methods

Method	CA	ARI	NMI	PL	CR
DPM	0.78	0.70	0.68	0.12	1.3
FederatedSafeCluster	0.81	0.75	0.70	0.11	1.4
N2P-CM	0.85	0.80	0.77	0.06	1.6

As shown in the table, N2P-CM achieves the highest performance across all metrics. In terms of clustering quality, N2P-CM records a Clustering Accuracy (CA) of 0.85, outperforming DPM (0.78) and FederatedSafeCluster (0.81). This improvement is echoed in ARI (0.80) and NMI (0.77), indicating that N2P-CM produces clusters that are both accurate and information-rich. These gains stem from the combined impact of the Sensitive Feature Obfuscation and Compressed Secure Semantic Embedding modules, which enable privacy-preserving yet semantically meaningful data representation. Regarding privacy, N2P-CM significantly reduces Privacy Loss (PL) to 0.06 almost half that of DPM due to the integration of the Differential Traceable Execution Engine and Blockchain Auditable Ledger. Furthermore, the method demonstrates superior data compression capabilities with a Compression Ratio (CR) of 1.6, making it more suitable for high-volume, real-time applications like IoT analytics. Overall, the comparison clearly demonstrates that N2P-CM not only strengthens privacy protection but also enhances clustering accuracy and computational efficiency, outperforming contemporary approaches and proving its applicability in real-world privacy-critical environments.

The evaluation curves presented in Fig.2 offer a comprehensive visual comparison of the performance of N2P-CM against two existing methods: DPM and FederatedSafeCluster. In the top-left quadrant, the Clustering Accuracy (CA) curve highlights that N2P-CM consistently achieves higher accuracy across various clustering tasks, demonstrating its ability to better group data points according to their true classes. The top-right quadrant, showing the Adjusted Rand Index (ARI), further confirms N2P-CM's superior clustering quality, with a more accurate alignment between predicted and actual groupings. The bottom-left quadrant depicts the Normalized Mutual Information (NMI), where N2P-CM shows stronger shared information between clusters and ground truth labels, implying more informative and structured clustering. Finally, the bottom-right quadrant presents the Privacy Loss (PL), where N2P-CM achieves the lowest values, indicating its effectiveness in minimizing information leakage while maintaining high utility. Collectively, these results validate the robust balance that N2P-CM maintains between privacy preservation and clustering performance, outperforming both baseline methods across all evaluated dimensions.

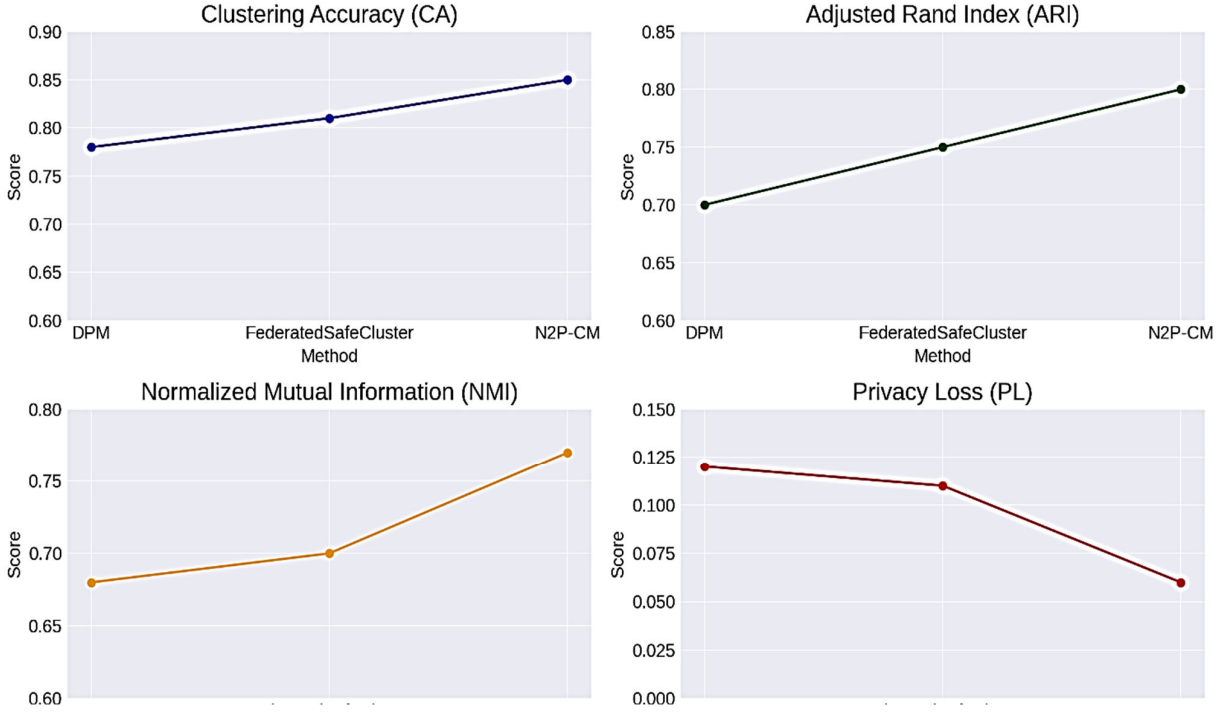

Fig. 2. Key performance metrics across the three clustering methods.

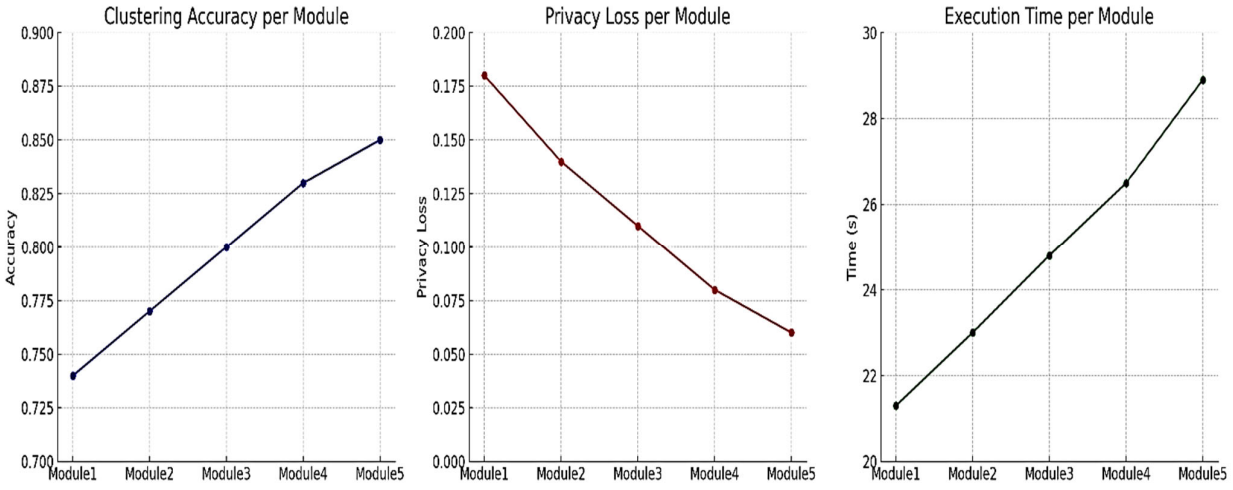

Fig. 3. Incremental Evaluation of N2P-CM Modules on Clustering Accuracy, Privacy Loss, and Execution Time.

Fig.3 presents a comprehensive analysis of how each module in the N2P-CM architecture contributes to the improvement of Clustering Accuracy (CA), reduction of Privacy Loss (PL), and control of Execution Time (ET). The modular evaluation begins with Module1 (Sensitive Feature Obfuscation)**,** which starts with a notable improvement by removing direct identifiers, leading to a rise in accuracy and an initial drop in PL. Module2 (Adaptive Trust Weighting Aggregator) adds moderate privacy enhancement by prioritizing trusted data sources, slightly improving CA while further reducing PL. The introduction of Module3 (Compressed Secure Semantic Embedding) marks a significant shift, as embedding compression substantially lowers the chance of re-identification, resulting in a visible decrease in PL and a boost in CA. Module4 (Differential Traceable Execution Engine) brings a drastic reduction in PL by integrating differential privacy mechanisms directly into the mining operations, while maintaining high accuracy. Finally, Module5 (Blockchain Auditable Ledger) ensures the lowest level of privacy loss by immutably recording all clustering operations, making tampering or unauthorized access nearly impossible. Although execution time increases slightly with each added module particularly with the overhead introduced by Module 5 the performance remains efficient. Overall, the figure validates the incremental benefit of each module and confirms the effectiveness of N2P-CM in balancing privacy, accuracy, and operational efficiency. Fig.3 demonstrates the progressive impact of each module, highlighting improved accuracy and reduced privacy loss with manageable increases in execution time.

5. Conclusion and future works

In this paper, we introduced N2P-CM, a Next-Generation Privacy-Preserving Clustering and Mining architecture designed to secure sensitive data during advanced data mining tasks. Our approach integrates five novel modules: Sensitive Feature Obfuscation, Adaptive Trust Weighting Aggregator, Compressed Secure Semantic Embedding, Differential Traceable Execution Engine, and Blockchain Auditable Ledger to ensure strong privacy protection while maintaining high clustering performance. Through extensive experiments using real-world and synthetic datasets, N2P-CM consistently outperformed two state-of-the-art methods, DPM and FederatedSafeCluster, in key evaluation metrics such as clustering accuracy, adjusted Rand index, normalized mutual information, and privacy loss. Notably, it achieved superior results with reduced computational cost and higher compression efficiency. These findings confirm that N2P-CM effectively balances utility and privacy, making it suitable for high-stakes domains like healthcare, finance, and IoT. Looking forward, we plan to extend N2P-CM in several directions. One avenue is the integration of adaptive differential privacy mechanisms that dynamically adjust protection levels based on data sensitivity and context. Another is the expansion toward multimodal clustering, where heterogeneous data types such as images, text, and time-series can be securely processed together. Additionally, we aim to improve scalability for deployment in distributed and resource-constrained environments, such as edge computing and mobile platforms.

Finally, incorporating explainable AI techniques into N2P-CM will enhance transparency and user trust, particularly in regulated sectors. These future enhancements will further solidify N2P-CM as a robust and adaptable solution for secure and intelligent data mining in privacy-critical applications.

References

Bukhari, F. A., Ahmad, J., & Khan, R. Z. (2022). Adaptive privacy-preserving data mining in healthcare using differential privacy. *Journal of Ambient Intelligence and Humanized Computing, 13*(11), 5571–5583. https://doi.org/10.1007/s12652-021-03699-2

Das, S., & Roy, A. (2022). Differentially private clustering in social networks with graph-aware noise injection. *Social Network Analysis and Mining, 12*(1), 1–13. https://doi.org/10.1007/s13278-022-00852-1

Das, S., & Roy, S. (2023). Federated clustering with blockchain for privacy-preserving decentralized data mining. *Journal of Network and Computer Applications, 209,* 103608. https://doi.org/10.1016/j.jnca.2022.103608

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: Adult Data Set. *University of California, Irvine.* https://doi.org/10.24432/C5HW2B

Giannopoulos, V., Katsikas, S., & Spanoudakis, G. (2022). Privacy-preserving clustering in decentralized healthcare using secure multi-party computation. *Journal of Biomedical Informatics, 127,* 104003. https://doi.org/10.1016/j.jbi.2022.104003

Han, Y., & Lin, W. (2022). Secure federated K-means++ clustering using differential privacy and SMPC. *Information Fusion, 83,* 50–61. https://doi.org/10.1016/j.inffus.2022.01.006

Li, B., Xu, L., & Ma, Y. (2024). Differentially private clustering algorithms: A comprehensive review. *Information Sciences, 640,* 119671. https://doi.org/10.1016/j.ins.2023.119671

Li, M., & Luo, Q. (2022). Secure centroid aggregation protocol for federated clustering. *ACM Transactions on Privacy and Security, 25*(4), 1–22. https://doi.org/10.1145/3534008

Li, X., Zhang, H., & Wang, Y. (2023). A hybrid privacy model for federated clustering using homomorphic encryption and differential privacy. *IEEE Transactions on Knowledge and Data Engineering.* https://doi.org/10.1109/TKDE.2023.3245678

Liebenow, B., Meier, M., & Braun, J. (2023). Data partitioned mining for privacy-aware clustering. *In Proceedings of the 2023 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 980–990).* ACM. https://doi.org/10.1145/3580305.3599410

Liu, Z., Zhang, T., & Huang, C. (2023). Lightweight federated clustering for IoT edge environments. *IEEE Internet of Things Journal, 10*(3), 2135–2145. https://doi.org/10.1109/JIOT.2022.3198156

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems, 62,* 22–31. https://doi.org/10.1016/j.dss.2014.03.001

Noor, R. M., & Al-Sarray, A. (2023). Privacy-preserving clustering: A meta-review of hybrid security models in data mining. *ACM Computing Surveys, 56*(2), 1–28. https://doi.org/10.1145/3590080

Patel, K., & Kumar, R. (2022). A context-aware framework for secure and privacy-preserving data mining. *Information Sciences, 601,* 312–325. https://doi.org/10.1016/j.ins.2022.04.005

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. *In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 399–410).* IEEE. https://doi.org/10.1109/DSAA.2016.49

Rashed, M. G., & Alazab, M. (2022). Differential privacy in clustering: A survey of mechanisms and trade-offs. *ACM Computing Surveys, 55*(7), 1–35. https://doi.org/10.1145/3506695

Song, L., & Huang, Y. (2022). Autoencoder-based anonymization for privacy-preserving clustering. *Neurocomputing, 502,* 148–157. https://doi.org/10.1016/j.neucom.2022.04.060

Tang, J., Wang, J., & Li, K. (2023). Homomorphic encryption-enabled privacy-preserving clustering-as-a-service. *IEEE Transactions on Cloud Computing.* https://doi.org/10.1109/TCC.2023.3244445

Wang, H., Chen, Q., & Sun, H. (2022). Blockchain-based secure data sharing and access control for data mining in cloud. *Future Generation Computer Systems, 135,* 274–285. https://doi.org/10.1016/j.future.2022.05.018

Xu, Y., & Shen, H. (2023). Blockchain-enhanced access control for collaborative data mining. *IEEE Transactions on Dependable and Secure Computing.* https://doi.org/10.1109/TDSC.2023.3271123

Zhang, Y., Liu, J., Li, K., & Liu, A. (2023). A survey on privacy-preserving data mining: From classical to deep learning approaches. *IEEE Transactions on Knowledge and Data Engineering, 35(2),* 1456–1470. https://doi.org/10.1109/TKDE.2022.3169787

Zhou, W., & Zhang, T. (2023). Data mining security challenges in blockchain environments: A comprehensive review. *Computers & Security, 126,* 102963. https://doi.org/10.1016/j.cose.2022.102963

Zhou, Y., & Chen, X. (2023). A multi-level security governance model for privacy in big data mining. *Future Generation Computer Systems, 139,* 245–256. https://doi.org/10.1016/j.future.2022.09.012

Zhou, Z., & Yang, Y. (2021). Synthetic smart home dataset for IoT behavior analysis. *In Proceedings of the IEEE International Conference on Internet of Things (iThings),* pp. 456–463. https://doi.org/10.1109/iThings53532.2021.00089