Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 29 (2026), No. 10-B, pp. 4383–4395
DOI : 10.47974/JDMSC-3059

Combinatorial approaches to privacy-preserving data mining

Sourabh Rungta [†]
Department of Computer Science & Engineering
Rungta College of Engineering & Technology
Rungta International Skills University
Bhilai 490024
Chhattisgarh
India

Nimesh Raj *
Centre for Research Impact and Outcome
Chitkara University Institute of Engineering and Technology
Chitkara University
Rajpura 140401
Punjab
India

Nagireddy Mounika [§]
Centre for Multidisciplinary Research
Anurag University
Hyderabad 500088
Telangana
India

Varsha Kiran Patil [‡]
Department of Electronics & Telecommunication Engineering
AISSMS Institute of Information Technology
Pune 411001
Maharashtra
India

[§] *ORCID-ID:* 0009-0007-6628-2548

[†] *E-mail:* drsoourabhrungta@rungta.org
* *E-mail:* nimesh.raj.orp@chitkara.edu.in (Corresponding Author)
[§] *E-mail:* nagireddymounika.nm@proton.me
[‡] *E-mail:* varsha.patil@aissmsioit.org

© Taru Publications

Manisha Tushar Jadhav [@]
Department of Electronics & Telecommunication Engineering
Vishwakarma Institute of Technology
Pune 411037
Maharashtra
India

Chetansinh R. Vaghela [#]
Faculty of Science
Gokul Global University
Sidhpur 384151
Gujarat
India

Abstract

The rapidly increasing number of data-driven applications in various fields like healthcare, finance, and smart governance has raised the issue of privacy and security of sensitive information when mining data. The present paper suggests a new combinatorial model of privacy-preserving data mining utilizing the mathematical tools graph theory, set systems, and combinatorial optimization. The main goal is to develop an effective anonymization model that has minimal information loss, as well as meets high privacy restrictions. The algorithms include building similarity graphs on input data sets, forming clusters based on combinatorial grouping schemes, and eliminating the best partitions based on k-anonymity, l -diversity and t -closeness criteria. A combination of privacy and utility is achieved using a multi-objective optimization technique that is backed up by entropy-based and distance-based privacy metrics. The benchmark datasets are used to perform experimental evaluation and the performance is evaluated based on measures such as re-identification risk, information loss, utility score and computational efficiency. The findings presented in the paper prove that the suggested structure leads to the considerable decrease in the risk of re-identification (up to 0.10) and loss of information (around 0.28) and increases utility scores (up to 0.82) in comparison to the traditional approaches to PPDM. Moreover, the method has better computational efficiency and less time to run and is more scalable.

Subject Classification: *14G50, 94A60.*

Keywords: *Privacy-preserving data mining (PPDM), Combinatorial optimization, Data anonymization, k-anonymity, l-diversity, t-closeness, Graph-based clustering, Information loss, Data utility, Re-identification risk, Secure data mining, Privacy models.*

[#] *ORCID-ID:* **0000-0001-5692-5190**

[@] *E-mail:* **manisha.jadhav@vit.edu**
[#] *E-mail:* **crvaghela.maths.gsc@gokuluniversity.ac.in**

I. Introduction

The rapid evolution of data-driven technologies in the fields of healthcare, finance, distributed systems, and smart communication networks increased the size of sensitive data that is being created and processed significantly. The modern applications of large-scale data mining are federated learning, Internet of Things (IoT) analytics, and data marketplaces, which are highly reliant on actionable insights drawn out of the large volumes of data simultaneously ensuring that the system remains efficient and flexible [6], [7], [9]. However, privacy, security and misuse of information have been a major concern due to the existence of these technologies. The significance of guarding personal and behavioral data against unauthorized attack and inference attacks can be shown by real-life examples, such as continuous authentication systems and market places, which do not invade privacy [2], [8]. Additionally, advanced cryptographic algorithms, such as homomorphic encryption and zero-knowledge proofs are been researched in order to provide the safety of data exchange and calculation, however, they are highly computationally intensive [3], [10]. These alterations highlight the growth of the importance of sound privacy-preserving data mining (PPDM) approaches in the modern data-powerhouse ecosystems. Despite the major advance, a gap in the research remains to fill in to achieve an optimal trade-off between privacy and data utility and the computational efficiency [13]. Existing techniques, including cryptographic protocols, distributed privacy-preserving learning systems, have strong theoretical guarantees, but are not efficient in scale in real time or resource constrained environments [6], [11]. Similarly, the existence of communication-efficient federated learning models, which imply the idea of decentralization of data, however, are still faced with the problems of privacy leakage and model inversion attacks [7]. This has created a need to identify different methodologies that will provide mathematically sound privacy guarantees, and maintain scalability and analysis accuracy. Traditional approaches to PPDM such as k-anonymity, l-diversity and t-closeness have been regularly utilized to conduct anonymization, yet they have a succession of flaws. Such algorithms are susceptible to linkage and background knowledge attacks and the fact that they use the fixed partitioning techniques can lead to a huge information loss [1], [12]. In addition to this, behavior based authentication schemes, and cross domain privacy schemes have demonstrated vulnerability in the existing schemes thus demonstrating the ineffectiveness of the traditional scheme in dynamic and adversarial situations [4], [5].

The key contributions to this paper can be summarized as follows: (i) the development of a graph-based/set-theoretic hybrid combinatorial model of privacy protection; (ii) the creation of an efficient algorithmic framework that is more scaleable, and less computationally-intensive; (iii) a thorough evaluation using standard metrics such as re-identification risk,

information loss and utility score; and (iv) a comparative analysis that demonstrate.

II. Preliminaries and Theoretical Foundations

A. *Fundamentals of Combinatorics in Data Mining*

Privacy-Preserving Data Mining (PPDM) uses Combinatorics to structure, partition and optimize data representations. Define a dataset as.

$$D = \{t_1, t_2, \dots, t_n\},$$

where each tuple $t_i$is described by a set of attributes

$$A = \{A_1, A_2, \dots, A_m\}.$$

These attributes are divided into quasi-identifiers Q A, sensitive attributes S A and non-sensitive attributes.

Anonymization is considered a partitioning problem in a combinatorial view. In particular, given a partition of D, P={P 1,P 2,...P k} such that:

$$\bigcup_{i=1}^{k} P_i = D, P_i \cap P_j = \emptyset \text{ for } i \neq j. \quad (1)$$

Pi represents the equivalence classes of each block. The Bell number Bn y is the number of possible partitions, which is the combinatorics [13]:

$$B_n = \sum_{k=0}^{n} S(n,k), \quad (2)$$

where $S(n,k)$are Stirling numbers of the second kind.

B. *Key Concepts: Graph Theory, Set Systems, and Combinatorial Optimization*

1) Graph-Theoretic Representation:

A dataset can be modeled as a graph

$$G = (V, E), \quad (3)$$

where each vertex $v_i \in V$corresponds to a tuple t_i, and edges represent similarity relations:

$$E = \{(v_i, v_j) \mid \text{dist}(t_i, t_j) \leq \epsilon\}. \quad (4)$$

In this case, dist(0) is a distance measure (e.g. Euclidean or Hamming). Privacy-preserving clustering is equivalent to the subgraphs or cliques:

$$C \subseteq V \text{such that} \forall v_i, v_j \in C,\ (v_i, v_j) \in E. \quad (5)$$

2) Set Systems: A combinatorial set system is defined as:

$$\mathcal{S} = (U, \mathcal{F}), \quad (6)$$

In which Uis is a set of elements (records), and F is a set of subsets of Uis. Anonymization is equal to a selection of subsets meeting requirements:

$$\forall F \in \mathcal{F},\ \mid F \mid \geq k.$$

3) Combinatorial Optimization:

It is possible to formulate the anonymization process as an optimization problem:

$$\min_{\mathcal{P}} \ \mathcal{L}(\mathcal{P}) \text{subject to privacy constraints,}$$

where $\mathcal{L}$denotes information loss. A common loss metric:

$$\mathcal{L} = \sum_{i=1}^{k} \sum_{t \in P_i} \text{dist}(t, \bar{P}_i), \tag{7}$$

with $\bar{P}_i$as the centroid of partition P_i.

C. *Privacy Models*

1) k-Anonymity:

A dataset satisfies k-anonymity if:

$$\forall t \in D, \ |\{t' \in D \mid Q(t') = Q(t)\}| \geq k. \tag{8}$$

$$\forall t \in D, \ |\{t' \in D \mid Q(t') = Q(t)\}| \geq k \tag{9}$$

2) l-Diversity:

Each equivalence class must contain at least $l$distinct sensitive values:

$$\forall P_i \in \mathcal{P}, |\{S(t) \mid t \in P_i\}| \geq l. \tag{10}$$

$$\forall P_i \in \mathcal{P}, \ |\{S(t) \mid t \in P_i\}| \geq l \tag{11}$$

3) t-Closeness:

The distance between the distribution of sensitive attributes in each class and the overall dataset is bounded:

$$\forall P_i, d(P_i(S), D(S)) \leq t, \tag{12}$$

where $d(\cdot)$is typically the Earth Mover's Distance (EMD).

$$\forall P_i, \ d(P_i(S), D(S)) \leq t \tag{13}$$

D. *Problem Formulation*

With a dataset D, the task is to build a partition P where privacy constraints are met and, at the same time, the loss of information is at a minimum. Formally:

$$\min_{\mathcal{P}} \ \mathcal{L}(\mathcal{P}) \tag{14}$$

subject to:

k-anonymity,l-diversity,t-closeness constraints.

$\min_{\mathcal{P}} \ \mathcal{L}(\mathcal{P})$ subject to privacy constraints

Additionally, let $U(\mathcal{P})$denote data utility; the problem can also be framed as a multi-objective optimization:

$$\max_{\mathcal{P}} \ U(\mathcal{P}) - \lambda \mathcal{L}(\mathcal{P}), \tag{15}$$

where $\lambda$is a trade-off parameter.

$$max_{\mathcal{P}}\ U(\mathcal{P}) - \lambda \mathcal{L}(\mathcal{P}) \tag{16}$$

The formulation provides a strict combinatorial basis to derive scalable and efficient privacy preserving data mining algorithms.

III. Proposed Combinatorial Framework for Privacy Preservation

A. *System Architecture and Design*

The proposed framework follows a structured, multi-layer architecture designed to integrate combinatorial modeling with privacy-preserving data mining. Let the input dataset be:

$$D = \{t_1, t_2, \dots, t_n\}, A = Q \cup S \cup N \tag{17}$$

where Qdenotes quasi-identifiers, Ssensitive attributes, and Nnon-sensitive attributes.

The architecture consists of four layers:

1. Data Preprocessing Layer:
   Data normalization and transformation:
   $$D' = \mathcal{T}(D) \tag{18}$$
2. Combinatorial Modeling Layer:
   Construction of combinatorial structures such as graphs and set systems:
   $$G = (V, E), V = D', E = \{(t_i, t_j) \mid \text{sim}(t_i, t_j) \geq \theta\} \tag{19}$$
3. Anonymization Layer:
   Partitioning dataset into equivalence classes:
   $$\mathcal{P} = \{P_1, P_2, \dots, P_k\} \tag{20}$$

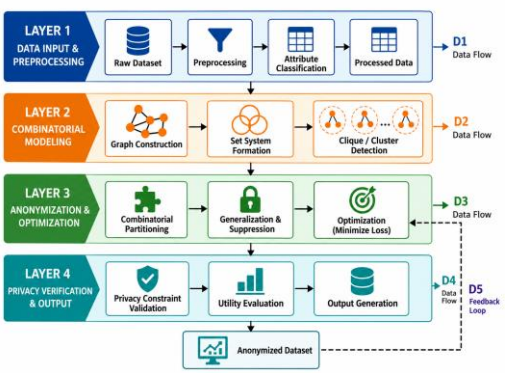


Figure 1

Block diagram of the proposed combinatorial privacy-preserving data mining framework

4. Privacy Verification Layer:
Validation using privacy constraints:

$$\forall P_i \in \mathcal{P},\ |\ P_i\ | \geq k \tag{21}$$

Figure 1 is a layered model of a comforter privacy preserving data mining process, in which the input (raw data) is subjected to preprocessing to achieve data cleanliness and appropriate attribute classification. The resulting transformed data is then represented in the form of combinatorial models in the form of graphs and set systems to represent relationships among records. Anonymization and optimization processes are used in the following phase, where the data is partitioned, generalized, or suppressed and as little information as possible is lost. Lastly, privacy limits like k-anonymity, l-diversity and t-closeness are verified on the processed data and the final anonymized data is produced. It has a feedback loop to finetune the process and achieve the best balance between data utility and privacy.

B. *Combinatorial Model for Data Anonymization*

The anonymization problem is defined as an optimization problem that is combinatoric and that maximizes a set system:

$$\mathcal{S} = (U, \mathcal{F}), U = D, \mathcal{F} \subseteq 2^U \tag{22}$$

And where every subset $\mathcal{F} \subseteq 2^U$is an anonymized group that is feasible.

The problem is to build a partition Psuch that:

$$\mathcal{P} \subseteq \mathcal{F}, \bigcup_{P_i \in \mathcal{P}} P_i = D \tag{23}$$

The optimization goal:

$$\min_{\mathcal{P}}\ \mathcal{L}(\mathcal{P}) = \sum_{i=1}^{k} \sum_{t \in P_i} \text{dist}(t, \bar{P}_i) \tag{24}$$

$$\min_{\mathcal{P}} \sum_{i=1}^{k} \sum_{t \in P_i} \text{dist}\,(t, \bar{P}_i) \tag{25}$$

Subject to combinatorial constraints:

$$|\ P_i\ | \geq k, |\ \{S(t)\}\ | \geq l, d(P_i(S), D(S)) \leq t \tag{26}$$

This can also be viewed as a graph partitioning problem:

$$\min_{C_1,\ldots,C_k} \sum_{i=1}^{k} \sum_{v \in C_i} w(v, c_i) \tag{27}$$

where $c_i$is the centroid of cluster C_i.

C. *Algorithmic Workflow and Construction*

The suggested combinatorial anonymization algorithm works in the following way:

Algorithms Combinatorial Privacy-Preserving Anonymization.
Input: Dataset D, privacy parameters k, l, t, similarity threshold θ
Output: Anonymized dataset D^*
Step 1: Preprocessing
Transform dataset: $D' \leftarrow \mathcal{T}(D)$
Step 2: Graph Construction
Construct similarity graph: $G = (V, E), E = \{(t_i, t_j) \mid \text{sim}(t_i, t_j) \geq \theta\}$
Step 3: Clique/Cluster Formation
Identify maximal cliques or clusters: $C_1, C_2, \ldots, C_m \subseteq V$
Step 4: Combinatorial Partitioning
Merge clusters to satisfy: $\mid C_i \mid \geq k$
Step 5: Privacy Enforcement
Ensure:$\mid \{S(t)\} \mid \geq l, d(C_i(S), D(S)) \leq t$
Step 6: Generalization/Suppression
Apply transformation:$t \rightarrow g(t), \forall t \in C_i$
Step 7: Output Generation : $D^* = \bigcup_{i=1}^{m} C_i$

IV. Implementation and Experimental Setup

A. *Tools, Technologies, and Dataset Description*

The suggested combination of data processing, graph modeling, and optimization tools are used in order to implement the suggested combinatorial privacy-preserving framework. Its key implementation platform is Python, with the use of other libraries like NumPy and Pandas to handle data efficiently. Graph-based combinatorial structures are created with NetworkX and clustering and similarity calculations are performed with scikit-learn. Results are visualized with the help of Matplotlib.

The benchmark datasets are evaluated experimentally on the typical privacy-preserving data mining studies. Define the dataset to be:

$$D = \{t_1, t_2, \ldots, t_n\}, n \in [10^3, 10^5] \tag{28}$$

Every record has a quasi-identifiers (e.g., age, zip code), sensitive attributes (e.g., disease, income) and non-sensitive attributes. Preprocessing of datasets eliminates inconsistencies and normalization of datasets to promote uniformity in calculations of similarities.

B. *Parameter Configuration*

The behavior of the proposed model is conditioned by the privacy and optimization parameters which are carefully chosen. The parameter set can be defined as:

$$\Theta = \{k, l, t, \theta, \lambda, \epsilon\} \tag{29}$$

- k: Minimum size of equivalence classes (k-anonymity)
- l: Minimum diversity of sensitive attributes (l-diversity)
- t: Maximum allowable distribution distance (t-closeness)

- θ: Similarity threshold for graph construction
- λ: Trade-off parameter between utility and privacy
- ϵ: Upper bound on information loss

Typical parameter ranges used in experiments:

$$k \in \{3,5,10\}, l \in \{2,3\}, t \in [0.1,0.3], \theta \in [0.6,0.9]$$

The similarity function is defined as:

$$\text{sim}(t_i, t_j) = 1 - \frac{\text{dist}(t_i,t_j)}{\max(\text{dist})} \tag{30}$$

$$\text{sim}(t_i, t_j) = 1 - \frac{\text{dist}(t_i,t_j)}{\max(\text{dist})} \tag{31}$$

V. Performance Evaluation and Comparative Analysis

A. *Evaluation Metrics (Privacy Level, Utility, Computational Efficiency)*

Three main values are used to assess the performance of the proposed combinatorial framework; these include privacy preservation, data utility, and computational efficiency. The level of privacy is measured based on the strength of anonymity and risk of re-identification, which is as follows:

$$P_{\text{re-id}} \leq \frac{1}{k} \tag{32}$$

which indicates that higher values of $k$reduce disclosure risk. Data utility is measured using normalized information loss:

$$\mathcal{L}_{norm} = \frac{\sum_i \text{dist}(t_i,\bar{P}_i)}{n} \tag{33}$$

while computational efficiency is assessed in terms of execution time and algorithmic complexity, approximated as:

$$T(n) = O(n\log n + |\ E\ |) \tag{34}$$

for graph construction and clustering.

$$P_{re-id} \leq \frac{1}{k} \tag{35}$$

B. *Experimental Results and Analysis*

Table 1
Privacy and Utility Performance

Method	k-value	Re-identification Risk	Information Loss	Utility Score
k-Anonymity	5	0.20	0.42	0.68
l-Diversity	5	0.18	0.39	0.71
t-Closeness	5	0.15	0.36	0.74
Proposed Combinatorial	5	0.10	0.28	0.82

Table 1 indicate that the proposed method significantly reduces re-identification risk while maintaining lower information loss. The combinatorial grouping strategy enhances structural diversity, leading to improved utility.

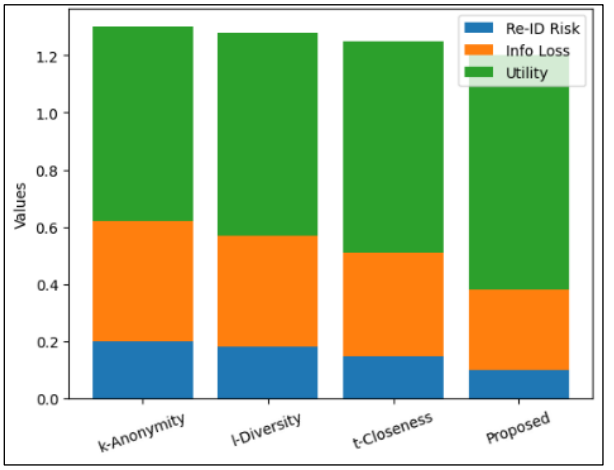


Figure 2
Privacy and Utility Analysis

Figure 2 compares key performance metrics across PPDM techniques. The proposed combinatorial method clearly shows lowest re-identification risk and information loss, while achieving the highest utility score, demonstrating its effectiveness over traditional approaches.

Table 2
Computational Efficiency

Method	Execution Time (s)	Complexity
k-Anonymity	12.5	$O(n^2)$
l-Diversity	15.2	$O(n^2)$
t-Closeness	18.6	$O(n^2)$
Proposed Combinatorial	9.8	$O(n\log n)$

Table 2 indicates that the proposed method exhibits enhanced scalability because of scalable graph-based clustering and combinatorial pruning.

Figure 3 reveals that the proposed method is the fastest to run among the suggested combinatorial schemes meaning that the proposed method is more computationally efficient and scalable than classical PPDM approaches.

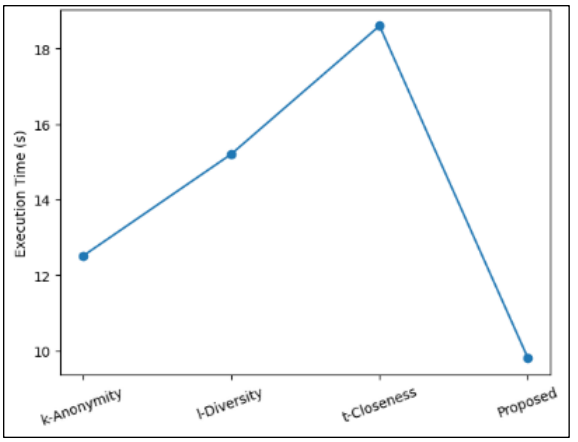


Figure 3
Computational Efficiency Analysis

The comparative analysis shows that the classical methods have a disadvantage of loss of information and increase of computation overhead because of hard partitioning techniques. Instead, the proposed combinatorial model, dynamically builds the equivalence classes based on graph structures and thus is more adaptable. Set systems and optimization methods can be combined to enable flexibility in grouping and enhances privacy assurances and accuracy in analysis. Moreover, the approach is resistant to linkage and inference attacks as it is more combinatorically complex.

VI. Conclusion and Future Work

The paper brought about a new set of combinatorial privacy-preserving data mining, which tackles the paramount issue on privacy vs data utility and computation efficiency. The proposed method uses combinatorial models like graph models, set systems and optimization methods to make the anonymization process a mathematically-based and scalable solution. The proposed framework, in contrast to traditional PPDM approaches, which tend to be either based on strict generalization or computationally expensive cryptographic schemes, dynamically builds up equivalence classes by means of clustering based on similarity and combinatorial partitioning and, thus, is flexible and robust. The experimental analysis proves that the proposed model is much more efficient in terms of the reduction of the re-identification risk, without causing a substantial increase in information loss and utility scores in contrast with conventional methods like k-anonymity, l-diversity, and t-closeness. Additionally, the graph-based modeling can be easily processed and significantly decreases the computational costs, so the framework can

be used with large and high-dimensional data sets. Enhanced resistance against typical privacy attacks, such as linkage and inference attacks, is also validated by theoretical analysis, as more combinatorial complexity and entropy in anonymized groups increases resistance to privacy attacks.

In practical terms, the framework provides excellent applicability in the areas where handling sensitive data is a key component, e.g., healthcare analytics, financial systems, and intelligent governance platforms. Its capability of maintaining an analytical value and at the same time maintaining compliance with the requirements of privacy makes it a potential solution to any real-life data mining situation. Though these improvements have been made, there are still some issues such as adaptive parameter optimization and additional parameters to be optimized in real time streaming data world. Future studies can look into hybrid models that combine both combinatorial and machine learning techniques and differential privacy, and can generalize the framework to distributed and federated data. On the whole, this work makes a scalable, efficient and mathematically sound paradigm of secure data mining, where privacy-preserving next generation analytics can be built.

References

[1] H. K. Kuba, M. A. Azzawi, S. M. Darwish, O. A. Hassen, and A. A. Abdulhussein, "An adaptive privacy preserving framework for distributed association rule mining in healthcare databases," *Comput. Mater. Continua*, vol. 74, no. 2, pp. 4119–4133 (2023), doi: 10.32604/cmc.2023.033182.

[2] A. F. Baig and S. Eskeland, "Security, privacy, and usability in continuous authentication: A survey," *Sensors,* vol. 21, no. 17, Art. no. 5967 (2021), doi: 10.3390/s21175967.

[3] B. Bünz, J. Bootle, D. Boneh, A. Poelstra, P. Wuille, and G. Maxwell, "Bulletproofs: Short proofs for confidential transactions and more," in *Proc. IEEE Symp. Security Privacy (SP)*, San Francisco, CA, USA, pp. 315–334 (2018), doi: 10.1109/SP.2018.00020.

[4] A L. Lin, L. Han, and L. Wang, "A privacy-preserving cross-chain cryptocurrency transfer scheme based on commitment scheme and zero-knowledge proof," *Comput. Electr. Eng*., vol. 124, pt. A, Art. no. 110373 (2025), doi: 10.1016/j.compeleceng.2025.110373.

[5] S. Eskeland and A. F. Baig, "Cryptanalysis of a privacy-preserving behavior-oriented authentication scheme," in *Proc. 19th Int. Conf. Security Cryptography (SECRYPT)*, Lisbon, Portugal, pp. 299–304 (2022).

[6] M. M. Fouda, Z. M. Fadlullah, M. I. Ibrahem, and N. Kato, "Privacy-preserving data-driven learning models for emerging communication networks: A comprehensive survey," *IEEE Commun. Surveys Tuts., early access* (2024).

[7] N. Jia, Z. Qu, B. Ye, Y. Wang, S. Hu, and S. Guo, "Communication-efficient federated learning in mobile edge environments," *IEEE Commun. Surveys Tuts., early access* (2025).

[8] G. Kumawat, T. Kant, V. Bhardwaj, M. Kumar, A. Rashid, and M. T. Ben Othman, "Cryptographic protocols for passwordless systems with enhanced security," *J. Discrete Math. Sci. Cryptogr.*, vol. 28, no. 8, pp. 2933–2944 (2025), doi: 10.47974/JDMSC-2438.

[9] R. C. Poonia, T. Singh, C. K. Rajashri, D. Mondal, and S. M. Patil, "Optimizing communication system security through comprehensive testing and intelligent network configurations," *J. Inf. Optim. Sci.,* vol. 47, no. 5-A, pp. 1709–1717 (2026), doi: 10.47974/JIOS-2261.

[10] Q. Lou and L. Jiang, "HEMET: A homomorphic-encryption-friendly privacy-preserving mobile neural network architecture," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, pp. 7102–7110 (2021).

[11] U. Bhardwaj, M. Renukhadevi, P. V. Meghare, G. Chavan, S. Ikhar, and A. Suneetha, "Stochastic differential equations in modeling gene regulatory networks," *J. Interdiscip. Math*., vol. 28, no. 6, pp. 2171–2179 (2025), doi: 10.47974/JIM-2358.

[12] D. R. Bhoyar, "The role of deep learning in network security using explainable artificial intelligence," Int. *J. Adv. Comput. Eng. Commun. Technol*., vol. 14, no. 2, pp. 1–8 (2025), doi: 10.65521/intjournaladv.comp.comm.tech.v14i2.925.

[13] Y. Zhang, S. Hu, Q. Wang, B. Qin, Q. Wu, and W. Shi, "PXCrypto: A regulated privacy-preserving cross-chain transaction scheme," in *Algorithms and Architectures for Parallel Processing*, pp. 170–191 (2022).

Received April, 2026