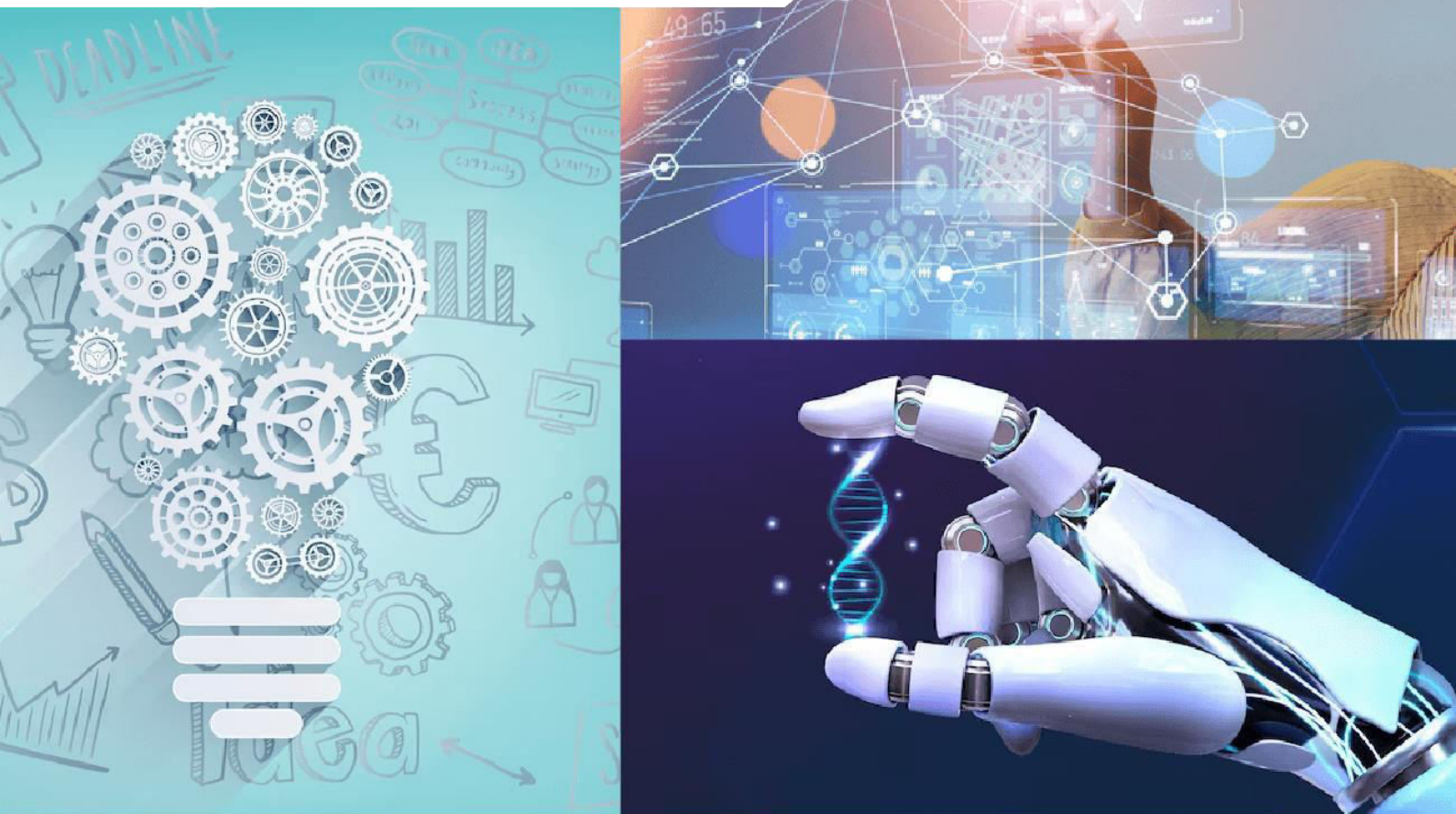




International Journal of Advanced Research in Education and Technology (IJARETY)

Volume 13, Issue 1, January - February 2026

Impact Factor: 8.152



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



Enhanced Privacy Preservation in Mixed Data Sharing using Correlation – Aware Differential Privacy and Data Balancing

A. Lakshmi Pathi Rao¹, Maddi Gopichand², Katroth Suman³, Suruguri Kamal Reddy⁴

Assistant Professor, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India¹

Student, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India ^{2,3,4}

ABSTRACT: In modern digital environments, sensitive data such as healthcare and financial records are frequently shared for analysis, raising serious privacy concerns. Traditional anonymization and differential privacy techniques often struggle to maintain a balance between data privacy and utility, especially in mixed and imbalanced datasets. This paper presents a web-based system that implements correlation-aware differential privacy combined with data balancing techniques to ensure secure and efficient data sharing. The system identifies data quality issues and enhances datasets through minimal synthetic data generation. It groups correlated attributes and applies optimized privacy mechanisms, reducing unnecessary noise while preserving data truthfulness. Numerical data is protected using the Laplace mechanism, while categorical data is handled through the exponential mechanism. Additionally, AES encryption and SHA-256 hashing ensure secure storage and access control. The proposed approach achieves an improved trade-off between privacy, accuracy, and usability, making it suitable for real-world applications.

KEY WORDS

- **Differential Privacy** – A mathematical technique used to protect sensitive information by adding controlled noise to datasets while preserving overall data utility.
- **Data Balancing** – A preprocessing technique that improves dataset quality by handling imbalanced or incomplete data through controlled synthetic data generation.
- **Correlation-Aware Processing** – A method that identifies relationships between attributes and applies privacy mechanisms selectively to preserve meaningful data patterns.
- **Web-Based Application** – A browser-accessible system that enables users to upload, process, and securely share datasets through an online platform.
- **AES Encryption** – A symmetric encryption algorithm used to secure stored data and ensure confidentiality during data transmission and storage.
- **SHA-256 Hashing** – A cryptographic hashing technique used for secure password storage and data integrity verification.
- **Privacy Preservation** – Techniques and methods used to protect sensitive user information from unauthorized access or disclosure.
- **Mixed Data Handling** – The ability of the system to process datasets containing both numerical and categorical attributes efficiently.

I. INTRODUCTION

With the rapid growth of digital technologies such as IoT, cloud computing, and data-driven systems, large volumes of sensitive data are generated and shared across platforms. However, this data often contains personal and confidential information, leading to privacy risks such as identity disclosure and data misuse.

Traditional privacy-preserving techniques like k-anonymity and basic differential privacy fail to maintain a balance between privacy protection and data utility. They either introduce excessive distortion or fail to preserve meaningful relationships within the data.

To address these challenges, this work proposes a **correlation-aware differential privacy model combined with data balancing techniques**. Instead of applying privacy mechanisms uniformly, the system intelligently identifies sensitive data portions and processes them selectively. This ensures better preservation of data structure while maintaining strong privacy guarantees.

II. LITERATURE REVIEW

In recent years, privacy-preserving data sharing has gained significant attention due to the increasing use of sensitive data in domains such as healthcare, finance, and smart systems. Several research works have been proposed to address challenges related to privacy, scalability, and data utility.

R. Wang et al. (2024), in their paper titled “**A Divide-and-Conquer Approach to Privacy-Preserving High-Dimensional Big Data Release**,” proposed a method that partitions high-dimensional datasets into smaller subsets. This approach improves scalability and reduces computational complexity while preserving privacy. However, it introduces additional system complexity and requires careful partitioning strategies.

M. You et al. (2024), in “**Hierarchical Adaptive Evolution Framework for Privacy-Preserving Data Publishing**,” introduced a multi-layered framework that dynamically adjusts privacy mechanisms based on data sensitivity. This adaptive approach improves flexibility and achieves a better balance between privacy and utility, but it increases implementation complexity.

X. Han et al. (2024), in “**HyObscure: Hybrid Obscuring for Privacy-Preserving Data Publishing**,” proposed a hybrid approach that combines multiple privacy techniques such as generalization, suppression, and perturbation. This method enhances privacy protection and reduces risks of inference attacks. However, it requires careful parameter tuning to avoid excessive data distortion.

K. Zhang et al. (2024), in “**Bounded and Unbiased Composite Differential Privacy**,” focused on improving differential privacy by controlling noise distribution. Their approach reduces excessive distortion and improves accuracy, but it still faces limitations in preserving correlations among attributes.

Q. Razi et al. (2024), in “**Privacy Utility Tradeoff Between PETs: Differential Privacy and Synthetic Data**,” analyzed the trade-off between privacy and utility. The study highlights that differential privacy provides strong guarantees, while synthetic data offers better utility. However, both approaches require careful balancing to avoid privacy leakage or loss of accuracy.

Overall, existing methods either compromise data utility or fail to preserve meaningful relationships between attributes. These limitations highlight the need for a correlation-aware differential privacy approach combined with data balancing, as proposed in this work.

No.	Paper Title	Author Name	Key Points	Remark
1	A Divide-and-Conquer Approach to Privacy-Preserving High-Dimensional Big Data Release	Wang et al., 2024	Splits high-dimensional datasets into smaller subsets for efficient processing and scalability	Improves scalability but increases complexity
2	Hierarchical Adaptive Evolution Framework for Privacy-Preserving Data Publishing	You et al., 2024	Uses hierarchical structure and dynamically adjusts privacy levels based on data sensitivity	Provides flexibility but requires complex implementation
3	HyObscure: Hybrid Obscuring for Privacy-Preserving Data Publishing	Han et al., 2024	Combines generalization, suppression, and perturbation techniques for enhanced privacy	Strong protection but requires parameter tuning
4	Bounded and Unbiased Composite Differential Privacy	Zhang et al., 2024	Introduces bounded noise to improve accuracy and reduce distortion	Improves utility but still limited in preserving relationships

5	Privacy Utility Tradeoff Between PETs: Differential Privacy and Synthetic Data	Razi et al., 2024	Compares privacy strength and utility between DP and synthetic data methods	Highlights trade-off but lacks unified solution.
---	--	-------------------	---	--

III. METHODOLOGY OF PROPOSED SURVEY

The proposed system is designed as a **privacy-preserving web-based framework** that integrates data balancing, correlation-aware differential privacy, and security mechanisms to ensure safe and efficient data sharing.

System Overview

The system allows users to upload datasets containing sensitive information. Before sharing the data, it undergoes preprocessing, balancing, correlation analysis, and privacy protection. Finally, the processed data is securely stored and shared with authorized users.

Working Process

- User Authentication:** The user registers and logs into the system using secure credentials. Passwords are protected using SHA-256 hashing to ensure secure authentication.
- Data Upload:** The user uploads a dataset containing mixed data (numerical and categorical attributes). The system stores metadata and prepares the dataset for processing.
- Data Preprocessing and Balancing:** The system analyzes the dataset to identify issues such as missing values, redundancy, and imbalance. It generates minimal synthetic records to balance the dataset and improve quality.
- Correlation Analysis:** Attributes within the dataset are analyzed to identify relationships and dependencies. Correlated attributes are grouped together to ensure that privacy mechanisms do not destroy meaningful patterns.
- Application of Differential Privacy:** The system applies privacy protection selectively:
 - Numerical data is protected using the Laplace mechanism
 - Categorical data is protected using the exponential mechanism
- Privacy Budget Management:** A privacy parameter (epsilon) is assigned to control the level of privacy. The system ensures that privacy loss is monitored and does not exceed allowed limits.
- Data Encryption and Storage:** The processed dataset is encrypted using AES encryption before being stored in the database. This ensures confidentiality and protection from unauthorized access.
- Access Control and Data Sharing:** Only authorized users can request and access the processed data. The system verifies permissions before allowing data download.

Modules of the System

User Module

- User registration and login
- Dataset upload
- View processed data

Admin/Analyst Module

- Verify uploaded datasets
- Approve user requests
- Monitor system activity

Key Components

1. Data Balancing Module

This module improves dataset quality by identifying imbalance and generating minimal synthetic data. It ensures fair representation of all data categories.

2. Correlation-Aware Differential Privacy Module

This is the core module of the system. It analyzes attribute relationships and applies privacy mechanisms selectively to maintain data structure and accuracy.

3. AES Encryption Module

Ensures secure storage of data by encrypting datasets before saving them in the database.

4. SHA-256 Module

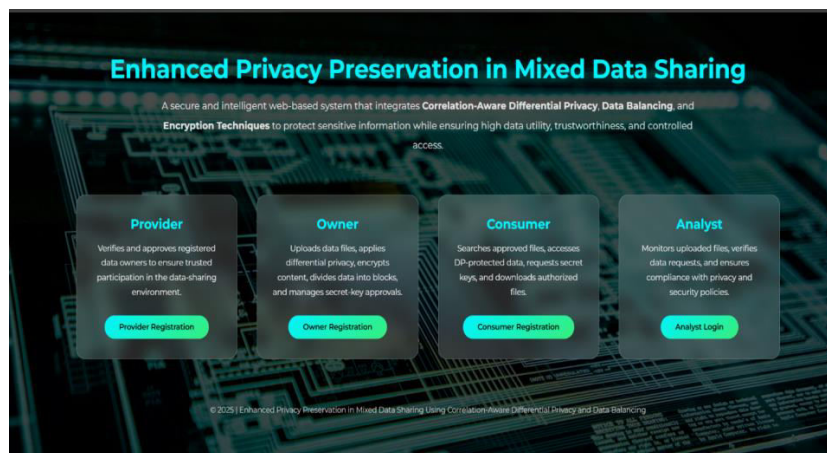
Provides secure hashing for passwords and ensures data integrity.

System Benefits:

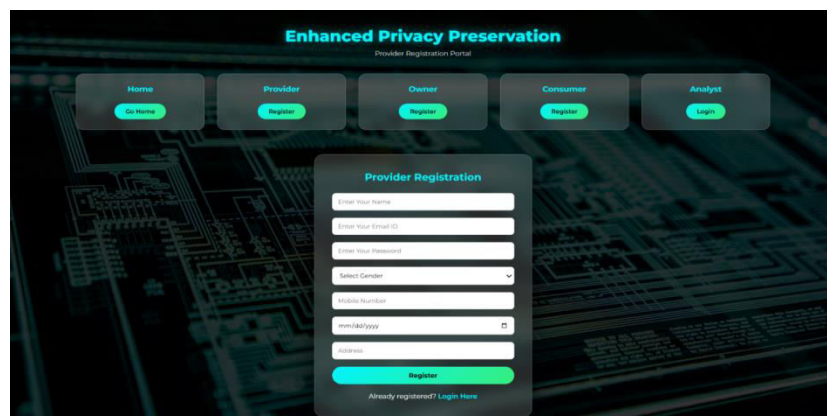
- Preserves data relationships and patterns
- Reduces excessive noise in data
- Improves accuracy and usability
- Provides strong privacy guarantees
- Supports real-world mixed datasets

IV. RESULTS

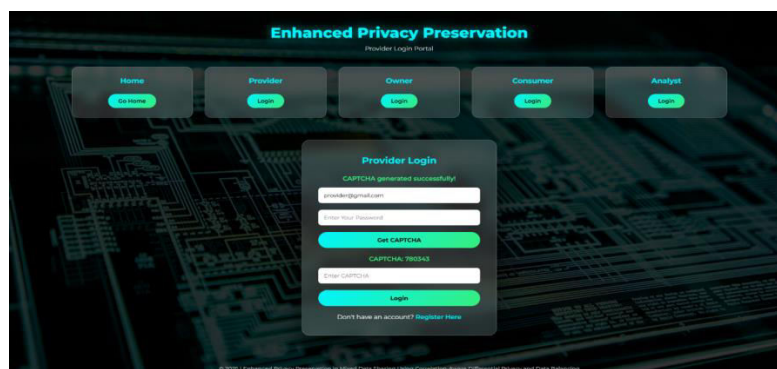
4.1 Home Page

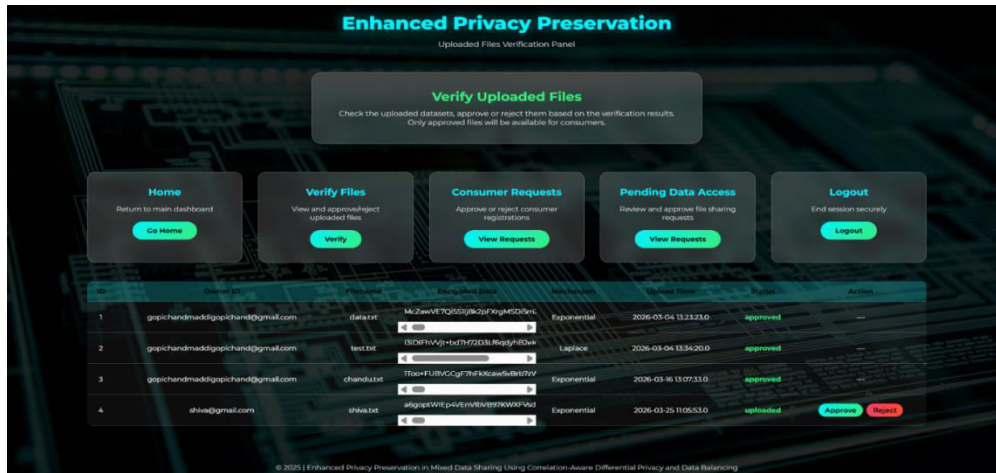


4.2 Provider Registration Page



4.3 Provider Login Page

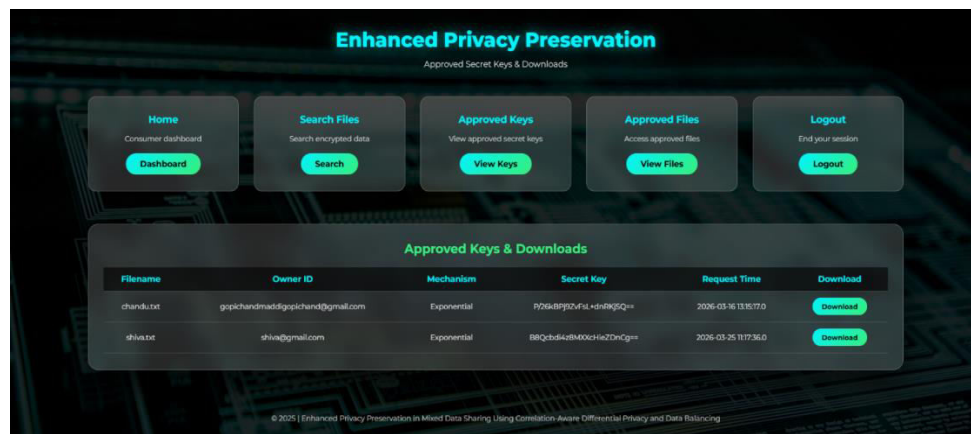




4.9 Consumer Access DP Data



4.10 Consumer Download Original Data



V. CONCLUSION AND FUTURE WORK

This paper presents a privacy-preserving data-sharing system using correlation-aware differential privacy and data balancing. The proposed approach effectively addresses limitations of traditional methods by preserving data relationships while ensuring strong privacy protection.

The system improves data quality before applying privacy mechanisms, resulting in better analytical accuracy and usability. Integration of encryption and hashing enhances overall system security.

REFERENCES

- [1] R.Wang, J. Liang, S. Wang, and C. Chang, “A divide-and-conquer approach to privacy-preserving high-dimensional big data release,” J. Inf. Secur. Appl., vol. 83, Apr. 2024, Art. no. 103756.
- [2] M. You, Y.-F. Ge, K. Wang, H. Wang, J. Cao, and G. Kambourakis, “Hierarchical adaptive evolution framework for privacy-preserving data publishing,” World Wide Web, vol. 27, no. 4, p. 49, Jul. 2024.
- [3] A. Majeed and S. O. Hwang, “Rectification of syntactic and semantic privacy mechanisms,” IEEE Secur. Privacy, vol. 21, no. 5, pp. 2–16, Sep./Oct. 2023.
- [4] L. Sweeney, “K-anonymity: A model for protecting privacy,” Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 557–570, Oct. 2002.
- [5] S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati, “k anonymity: From theory to applications,” Trans. Data Privacy, vol. 16, no. 1, pp. 25–49, 2023.
- [6] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian, “L diversity: Privacy beyond k-anonymity,” ACM Trans. Knowl. Discovery From Data, vol. 1, no. 1, p. 3, Jan. 2006.
- [7] N. Li, T. Li, and S. Venkatasubramanian, “T-closeness: Privacy beyond k-anonymity and l-diversity,” in Proc. IEEE 23rd Int. Conf. Data Eng., Apr. 2007, pp. 106–115.
- [8] Y. Yan, E. A. Herman, A. Mahmood, T. Feng, and P. Xie, “A weighted K member clustering algorithm for K-anonymization,” Computing, vol. 103, no. 10, pp. 2251–2273, Oct. 2021.
- [9] L. Caruccio, D. Desiato, G. Polese, G. Tortora, and N. Zannone, “A decision-support framework for data anonymization with application to machine learning processes,” Inf. Sci., vol. 613, pp. 1–32, Oct. 2022.
- [10] X. Han, Y. Yang, J. Wu, and H. Xiong, “HyObscure: Hybrid obscuring for privacy-preserving data publishing,” IEEE Trans. Knowl. Data Eng., vol. 36, no. 8, pp. 3893–3905, Aug. 2024.
- [11] C. Dwork, “Differential privacy,” in Proc. Int. Colloq. Automata, Lang., Program. Cham, Switzerland: Springer, Jan. 2006, pp. 1–12.
- [12] R. Bild, K. A. Kuhn, and F. Prasser, “SafePub: A truthful data anonymization algorithm with strong privacy guarantees,” Proc. Privacy Enhancing Technol., vol. 2018, no. 1, pp. 67–87, Jan. 2018.
- [13] D. Pujol and A. Machanavajjhala, “Equity and privacy: More than just a tradeoff,” IEEE Secur. Privacy, vol. 19, no. 6, pp. 93–97, Nov. 2021.
- [14] M. Rafiei, F. Wang, M. Pourbafrani, and W. M. vanderAalst, “TraVaG: Differentially private trace variant generation using GANs,” in Proc. Int. Conf. Res. Challenges Inf. Sci. Cham, Switzerland: Springer, Jan. 2023, pp. 415–431.
- [15] B.-C. Tai, Y.-T. Tsou, S.-C. Li, Y. Huang, P.-Y. Tsai, and Y.-C. Tsai, “User-driven synthetic dataset generation with quantifiable differential privacy,” IEEE Trans. Services Comput., vol. 16, no. 5, pp. 3812–3826, Sep./Oct. 2023.
- [15] M. Strobel and R. Shokri, “Data privacy and trustworthy machine learning,” IEEE Secur. Privacy, vol. 20, no. 5, pp. 44–49, Sep. 2022.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 8.152