*Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences*
*ISSN : 1674-2974 | CN 43-1061 / N*

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

# PRIVACY PRESERVATION FOR BIG DATA PUBLISHING: APPLYING K-ANONYMITY AND DIFFERENTIALLY PRIVATE SYNTHETIC DATA GENERATION WITH DP-CTGAN

**ANANNA HOQUE SHATHI**
Department of Computer Science and Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh.

**Dr. BOSHIR AHMED**
Department of Computer Science and Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh.

***Abstract***

*One effective privacy protection method utilized in many tech domains, including big data, is anonymization, which protects extremely sensitive information from outside parties. Extracting enough information from anonymized data while preserving privacy is still difficult, even with major developments that promote secondary use of data. Existing systems often convert large data, compromising their structure and utility. Excessive modification can hinder the performance of mechanisms and their output in real-life circumstances. To solve these problems in our work, we suggest and put into practice a hybrid anonymization method that combines k-anonymity and Differential Privacy Conditional Tabular Generative Adversarial Network (DP-CTGAN) to produce extremely superior quality data that provides insights comparable to actual data while maintaining privacy. We implemented the Mondrian and DP-CTGAN algorithms on the UCI-Adult dataset to hide extremely private information related to the income of a person from unauthorized viewers. The raw data are processed to hide unique individual information from the intermediate data frame. The Mondrian algorithm generates a range of unique information, keeping the rest of the information the same, which is considered to be a fruitful information set without showing one's private information. Our proposed approach produces more reliable anonymized data compared to the present literature.*

***Keywords:*** *Privacy Preservation; K-Anonymity; Differential Privacy; Big Data; Mondrian Algorithm; DP-CTGAN.*

## 1. INTRODUCTION

Big data pertains to the increasing volume of data that is difficult to manage, process, and analyze with traditional database technologies [1]. A firm is inundated with a tremendous amount of structured and unstructured data daily. Wu et al. (2016) proposed a far more thorough classification by distinguishing 32 unique V's, providing a multifaceted prism for assessing Big Data attributes [2]. This thorough taxonomy emphasizes how complicated things are becoming and how big data analytics needs increasingly sophisticated methods. A prevalent characteristic of big data is its diversity, meaning it can include various formats such as text, audio, images, and video, among others. Around 402.74 million terabytes, or 402.74 quintillion bytes, of data are produced per day worldwide as of 2024. Compared to the 2.5 quintillion bytes generated every day in 2012, this is a significant increase. An estimated 147 zettabytes of data will be used globally in 2024, and by 2025, that amount is expected to increase to 181 zettabytes. The extensive use of digital technologies, such as cloud computing, real-time data processing, and the Internet of Things (IoT), is driving this exponential

*ISSN：1674-2974 | CN 43-1061 / N*

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

expansion. Interestingly, more than half of all internet data traffic is now made up of video material. The extensive use of digital technologies, such as cloud computing, real-time data processing, and the Internet of Things (IoT), is driving this exponential expansion. Interestingly, more than half of all internet data traffic is now made up of video material. In recent years, several mechanisms have been created to protect the privacy of big data. The amount of data is enormous, it is generated rapidly, and the data/information landscape is worldwide (2016) [3]. Consideration should be given to lightweight incremental algorithms that can deliver robustness, high accuracy, and minimal pre-processing delay. The potential of swarm intelligence algorithms has been brought to light by recent de- velopments in feature selection techniques. For example, Jain and Purohit (2017) presented a modified particle swarm optimization technique and showed how effective it is at choosing pertinent features, which enhances classification performance and lowers computing cost [4].

Similar to this, Teng et al. (2017) suggested an adaptive feature selection technique that made use of V-shaped binary PSO. This method successfully assessed feature subsets as coherent units, improving the model's overall accuracy and global search capabilities [5]. These investigations highlight the adaptability and resilience of PSO-based methods in handling the challenges associated with feature selection in high-dimensional datasets. In their thorough analysis of big data analytics in smart grids, Zhang, Huang, and Bompard (2018) showed how the enormous volumes of data gathered from smart meters and sensors may be used for fault detection, predictive maintenance, and real-time monitoring [6]. In order to properly handle and evaluate this data and guarantee the dependability and effectiveness of smart energy systems, their study highlights the need for sophisticated data analytics. In a similar vein, Madadi et al. (2018) investigated how big data analysis might be used to operate smart power systems [7]. They talked about how integrating big data approaches can improve system operation, protection, and control, which will increase power networks' sustainability and efficiency. Furthermore, the use of big data analytics in smart grid technologies was explained by Misra and Bera (2018) [8]. They emphasized how crucial it is to handle and examine sizable, intricate datasets to maximize smart grid efficiency and dependability, especially in real-time situations. A range of privacy models, such as k-anonymity, l-diversity, and t-closeness, are utilized to protect against potential threats to the privacy of disseminated data. In another research [9], the authors provide a threat model as in figure 1 that describes the potential privacy hazards associated with disclosing or sharing sensitive datasets. In this context, a threat model is a structured depiction of the many sorts of adversaries, the information they may hold, and the potential assaults they can launch to breach the privacy of persons within a dataset. By incorporating this threat model into our research, we can better appreciate the limitations of each privacy technique and argue the necessity for a hybrid approach that uses both k-anonymity and differential privacy.

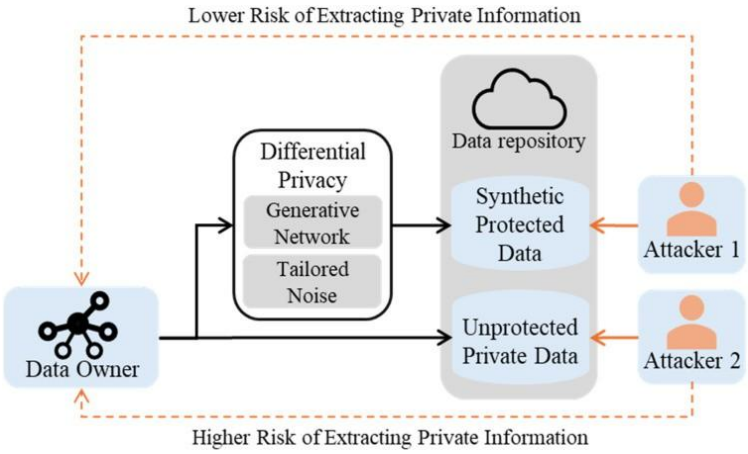


**Figure 1: Threat Model for Privacy Attack**

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

**Privacy-Preserving Data Publishing (PPDP):** This is a collection of strategies and tactics designed to publish data while safeguarding the privacy of the people whose information is included in the dataset. The PPDP attempts to strike a compromise between privacy safeguards and data value. Common PPDP techniques include:

- K-Anonymization: Each record is guaranteed to be identical to at least k-1 others.
- L-diversity: Within k-anonymous groups, l-diversity introduces variation in sensitive traits.
- T-closeness: Guarantees that each group's sensitive attribute distribution closely resembles the distribution as a whole.
- Perturbation: Including noise in data or query results to safeguard personal information used in differential privacy. Publishing manufactured data that closely resembles real data but does not directly relate to actual people is known as synthetic data generation.
- Data generalization/suppression is the process of lowering the level of detail in data or completely re- moving certain sections for example, displaying age as ”50–60” rather than ”54”.

The exponential expansion of digital data has raised concerns about the privacy of sensitive information. Or- ganizations and researchers must strike a balance between the demand for data utility and the requirement to protect individual privacy. Researchers have created a range of privacy-preserving strategies in response to these issues. A few of the present methodologies are observed for the benchmark of our proposed work.

Raj, Anushree and D'Souza (2019) focus on anonymization methods to protect privacy for data stored in the cloud using a k-anonymity algorithm with a MapReduce framework [10]. This paper discusses anonymization techniques for privacy protection of data published on the cloud, focusing on the top-down specification algorithm within k-anonymity. It explores the adaptation of the MapReduce framework to process large amounts of big data for anonymization. The implementation involves a generalized method using map and reduce phases in two different phases of top-down specification. Techniques for anonymizing data are essential for safeguarding sensitive information, particularly given the growing volume of data collected by businesses and government organizations. Traditional privacy-preserving data mining algorithms are in- sufficient for big data analytics, which are computed using MapReduce in cloud environments. The proposed algorithm parallelizes k-anonymity using MapReduce, implementing top-down specialization to anonymize data and support privacy preservation.

Kwatra and Torra (2022) suggest a privacy-preserving structure that utilizes k-anonymity by the Mondrian algorithm alongside decision trees within a federated learning environment for data that is horizontally divided. [11]. In federated learning, data heterogeneity results in non-IID (non-independent and identically distributed) data. A new method is introduced to create non-IID data partitions by addressing an optimization problem. Each device develops a decision tree classifier and shares the root node of its tree with an aggregator. The aggregator consolidates the trees by selecting the split attribute that occurs with the highest frequency and subsequently develops the branches in accordance with the split values associated with that attribute. Until every node that needs to be merged is a leaf node, this cyclical process keeps going. Upon the completion of the merging process, the aggregator disseminates the consolidated decision tree to the respective devices. The objective is to develop a cohesive machine learning model that integrates data from multiple devices while simultaneously maintaining k-anonymity for the individuals involved.

Vikas Thammanna Gowda, Mason Lee, and Vanessa Campagna propose Enhanced Stratified Sampling (ESS), integrating k-anonymity, l-diversity, and t-closeness to preserve privacy in big data

*ISSN：1674-2974 | CN 43-1061 / N*

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

publishing [12]. ESS generates data subsets, evaluating them based on privacy and information loss and selecting the optimal set for publication. This method enhances privacy while maintaining high data utility for large-scale data analysis without compromising privacy. The paper highlights several privacy risks associated with the publishing of big data, including

1. Attribute Disclosure: Individuals may be identified through their sensitive attributes, leading to unautho- rized access to personal information. This occurs when enough quasi-identifiers are present to re-identify individuals within a dataset.
2. Homogeneity Attacks: In scenarios where a sensitive attribute has a limited number of distinct values, attackers may infer the sensitive information of individuals belonging to small Equivalence Classes (ECs). This means that if all records in an EC have the same sensitive attribute value, it increases the risk of attribute disclosure.
3. Privacy Breaches: The mishandling of sensitive information can lead to privacy violations such as identity theft, discrimination against individuals, and financial losses related to exposed personal information.
4. Balancing Privacy and Data Usability: Ensuring privacy often involves generalizing or suppressing data, which in turn can lead to increased information loss and reduced data utility, making it challenging to extract meaningful insights from the published data.

These risks underline the importance of developing effective methods for protecting privacy within the realm of big data publishing to safeguard sensitive information while still allowing for valuable data analysis. Abdul Majeed and Seong Oun Hwang present a new method for data publishing that combines differential privacy with k-anonymity in [13]. This paper tackles the issue of extracting sufficient knowledge from anonymized data while maintaining privacy by introducing a combined Differential Privacy (DP) and k-anonymity approach. The method separates the dataset into partitions that either violate privacy or do not, applying a relaxed privacy budget $\epsilon$ to numerical attributes in the non-privacy-violating partition while keeping most categorical attributes intact. Experiments conducted on three real-world datasets demonstrate that this approach retains 60.81% of the original data in its anonymized form, decreases privacy risks by 20.05%, and improves utility by 54.01% and 15.33% when evaluated using Information Loss (IL) and accuracy metrics, respectively. The proposed hybrid scheme in the paper effectively balances data utility and privacy through several key mechanisms:

1. Partitioning Data: The scheme classifies the dataset into two partitions: non-privacy-violating and privacy- violating. In the non-privacy-violating partition, most of the data values are kept in their initial state, which helps to maintain data utility. In contrast, the privacy-violating partition undergoes minimal necessary anonymization, applying stringent privacy protections only where needed.
2. Relaxed Privacy Budget: A lenient privacy budget ($\epsilon$) is utilized for numerical attributes within the partition that does not violate privacy, which allows for better preservation of the data's original statistical properties. This approach effectively enhances the utility of the data while still adhering to privacy requirements. The privacy-violating partition, however, utilizes a more conservative privacy budget to ensure stronger privacy guarantees.
3. Minimization of Information Loss: The scheme is designed to minimize IL by carefully anonymizing only the necessary data portions. It aims to retain high accuracy and utility while safeguarding sensitive information against potential breaches. This is achieved by leveraging the diversity of sensitive attributes (SAs) and employing k-anonymity techniques that produce compact and diverse clusters.

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

Overall, the hybrid approach they proposed allows for effective anonymization without severely compromising the usefulness of the data, making it particularly suitable for data-hungry applications, such as AI.

**Table 1: Performance of Privacy-Preserving Techniques in Big Data**

| Authors (Year) | Methodology | Key Contributions |
|---|---|---|
| Raj, Anushree et al. (2019), | k-Anonymity with MapReduce | Demonstrates scalable k-anonymity using MapReduce to preserve privacy in cloud-based big data settings |
| Kwatra & Torra (2022) | Mondrian k-Anonymity and Decision Trees in Federated Learning | Builds a privacy-preserving FL framework using local k-anonymized training and collaborative decision tree fusion |
| Gowda et al. (2024) | ESS integrating k-Anonymity, l-Diversity, t-Closeness | Addresses attribute disclosure, homogeneity attacks, and utility trade-offs in published big data |
| Majeed & Hwang (2024) | Hybrid of Differential Privacy and k- Anonymity | Proposes partitioned anonymization with relaxed ε for numerical attributes to optimize both privacy and utility |

The other sections are divided to explain the experimental results of our proposed hybrid approach for privacy preservation using the Mondrian algorithm and DP-CTGAN. Section 2 discusses the methodology for the proposed hybrid strategy, which includes a description of the research design and dataset. Section 3 illustrates the results and discussion of the hybrid privacy for big data publishing, including a comparison of data types and DP levels, while Section 4 depicts the paper's conclusion.

## 2. METHODOLOGY

The emergence of data-driven applications has heightened concerns about personal privacy. This paper presents a two-phase privacy-preserving data publishing pipeline that combines k-anonymity and differential privacy. The suggested method protects privacy by first performing syntactic anonymization with the Mondrian algorithm, followed by semantic anonymization using a differentially private synthetic data generator.

### 2.1. Research Design

This work uses an experimental methodology to examine the effectiveness of two privacy-preserving approaches, k-anonymity and differential privacy, as applied to the Adult Income dataset. The study assesses confidentiality risk, data utility, and computing efficiency to obtain the following objectives:

- To assess the effectiveness of k-anonymity (Mondrian) and differential privacy (DP-CTGAN) on sensitive datasets.
- To assess the usefulness and privacy trade-offs of classical anonymization vs synthetic data production.
- To create a hybrid pipeline that incorporates both technologies for increased privacy.

The research design procedures are implemented in a methodical manner to ensure a structured and coherent approach throughout the project. This includes explicitly identifying the study problem, setting specified ob- jectives, and selecting acceptable data gathering and analysis procedures. The design also describes how to implement privacy-preserving approaches such as k-anonymity using the Mondrian algorithm and differential privacy via synthetic data generation. Each step is meticulously planned to ensure that the results are reliable and valid, that the data remains intact, and that the

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

research objectives are met. Following these procedures provides a logical foundation for the research, which improves the study's overall rigor and effectiveness.

**2.1.1. Data Preprocessing**

Before the experiments were carried out, the data sets were carefully preprocessed to ensure consistency and quality of the data. Encoding categorical data, standardizing formats, managing missing values, and eliminating any irregularities or contradictions were all part of this. To prepare the data for an accurate application of the privacy-preserving approach and dependable experimental results, such preparation measures were necessary.

- Remove direct identifiers.
- Clean and normalize data.
- Identify quasi-identifiers and sensitive attributes.

**2.1.2. Apply the Mondrian Algorithm to Implement k-Anonymity**

The Mondrian technique, which effectively achieves k-anonymity while maintaining data utility, was used to accomplish anonymization. In order to guarantee that each group comprises at least k indistinguishable items, this algorithm iteratively divides the dataset into smaller groups based on quasi-identifiers. Because of its multidimensional partitioning technique, which balances privacy protection with information loss, it works well with structured datasets.

- Implement Mondrian multidimensional partitioning.
- Choose appropriate values for k, which are 10 and 15.
- Produce an anonymized data set.

The following figure from [14] shows how partitioning occurs in the Mondrian algorithm.

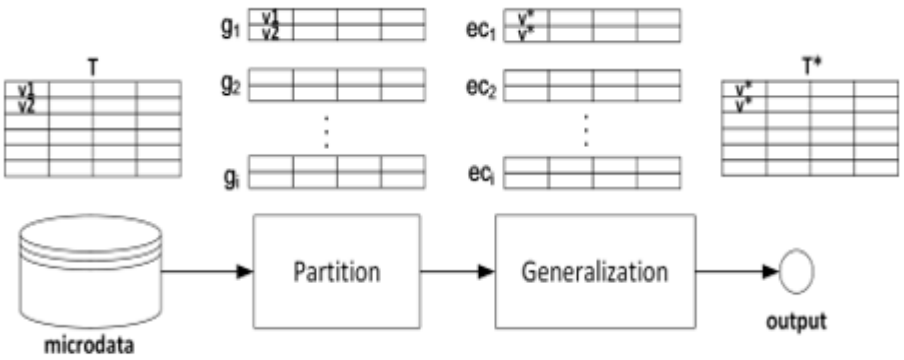


**Figure 2: Illustration of Mondrian Multidimensional Partitioning**

**2.1.3. Apply Differential Privacy using DP-CTGAN**

Deep generative models utilize neural networks to understand the features of a dataset and can produce synthetic data that closely mimics actual data, as shown in figure 3 [15]. To lower the chance of re-identification, the dataset was initially anonymized using the Mondrian algorithm, which satisfied k-anonymity.

The DP-CTGAN, which further improves privacy by introducing differential privacy mechanisms during the data generation process, relied on this anonymized data as a more privacy-conscious basis for training. The result was synthetic data that protects individual privacy while maintaining statistical properties.

*ISSN : 1674-2974 | CN 43-1061 / N*

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

Key points to provide differential privacy are listed below:

- Used the Mondrian-generalized dataset as input to DP-CTGAN.
- Train a synthetic data generator with $\epsilon$ - differential privacy (tune $\epsilon$ values like 0.5, 1, 2).
- Generate a synthetic dataset.

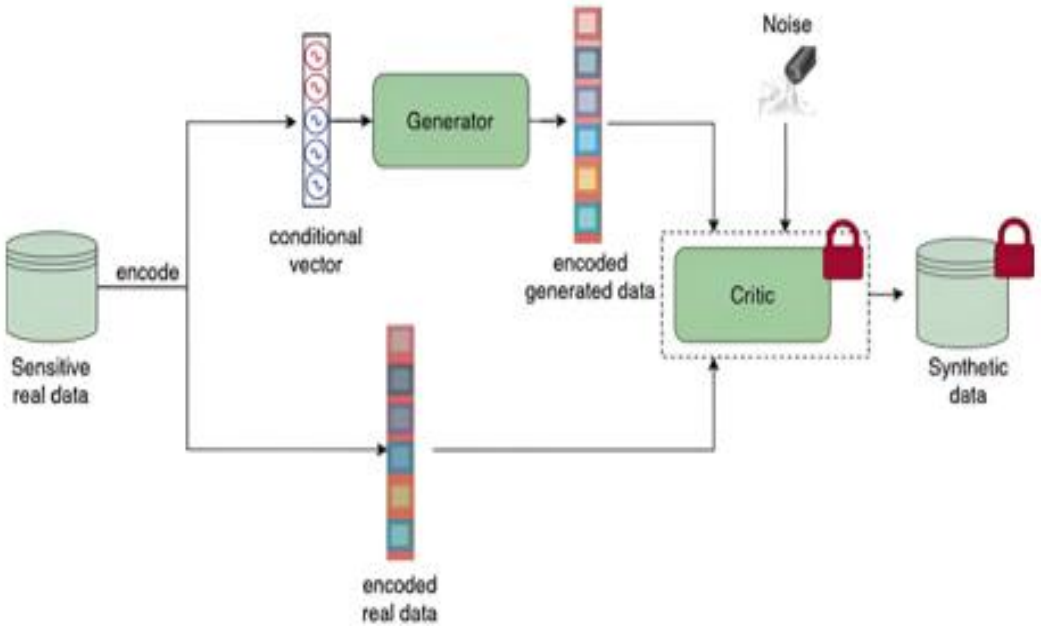


**Figure 3: DP-CTGAN Architecture**

The above figure illustrates the fundamental architecture of DP-CTGAN. Confidential training data is input into a conditional generator, which produces samples that adequately represent all potential discrete values.

Con- currently, a random perturbation is introduced to the critic to ensure privacy safeguards. During preprocessing, mode-specific normalization is applied to continuous columns, allowing the data representation to capture complex distributions. The conditional generator addresses the issue of imbalanced categorical columns, facilitating more effective and uniform data generation.

**2.1.4. Hybrid Approach**

A two-stage privacy-preserving pipeline was used to examine the effectiveness of differential privacy and k-anonymity together. The Mondrian algorithm, a multidimensional partitioning method, was initially employed to anonymize the original dataset, ensuring that each record is unidentifiable from at least k - 1 other records concerning quasi-identifiers.

The k-anonymized dataset created by this procedure reduces the possibility of re-identification via linkage attacks. In the second step, the k-anonymized data was used to train the DP-CTGAN model, which is an adaptation of the Conditional Tabular Generative Adversarial Network (CTGAN) that incorporates differential privacy guarantees.

This approach satisfies formal differential privacy restrictions by synthesizing fresh, statistically comparable records and introducing calibrated noise to mask the contribution of individual data points.

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

To obtain synthetic data, mainly we followed the following steps.

- Train DP-CTGAN on the k-anonymized dataset.
- Measure whether combining methods improves or harms utility/privacy.

This dual-layered solution seeks to find a balance between keeping data utility for downstream analytics and improving privacy protection beyond what either method can do on its own. By combining k-anonymity's structural anonymization with differential privacy's noise-based assurances, the technique aims to provide a more resilient solution to privacy concerns in data publishing and sharing. The following diagram shows the hybrid approach for enhanced big data privacy preservation.

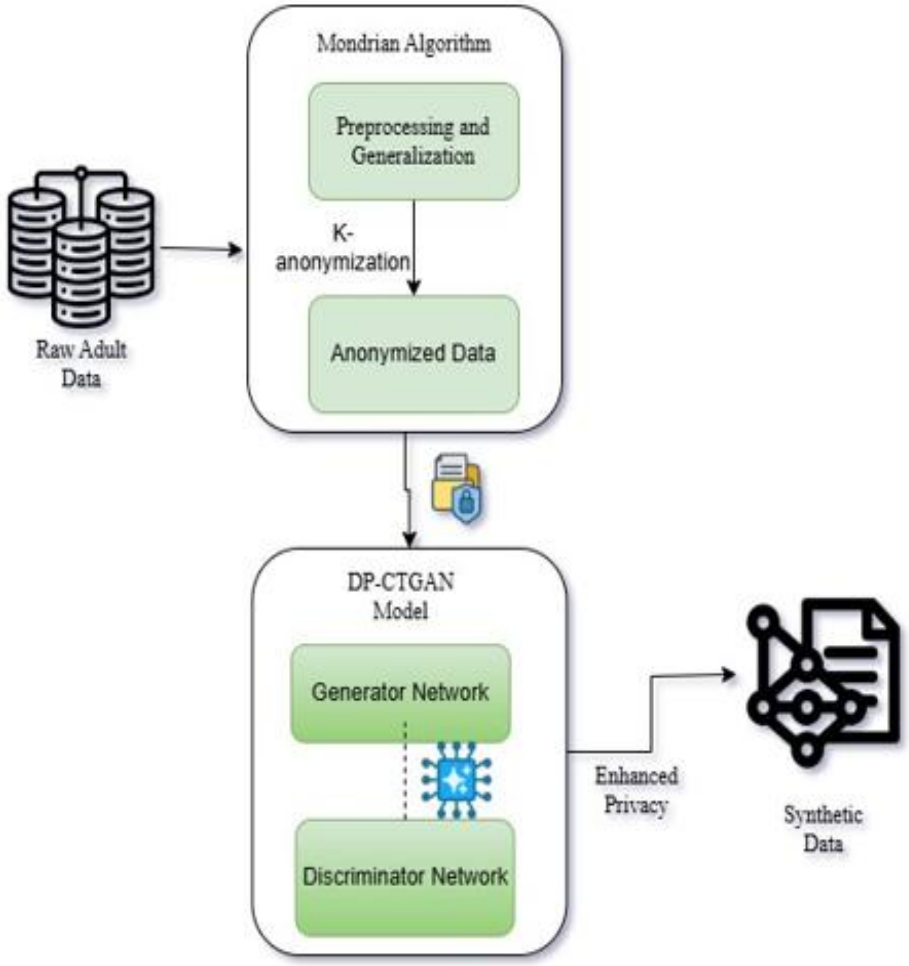


**Figure 4: Hybrid Model for Privacy Preservation**

### 2.2. Dataset Description

The UCI Adult Income dataset, often known as the" Census Income" dataset, is a popular benchmark in machine learning research, especially for classification tasks and privacy-preserving data processing. It is based on the US Census Bureau's 1994 and 1995 Current Population Surveys and is hosted via the UCI Machine Learning Repository. The dataset consists of 32,561 occurrences and 15 attributes, with a combination of categorical and continuous variables. Its major goal is to forecast whether a person earns more than $50,000 per year, i.e., has an income of > $50,000, based on demographic and job characteristics. The income class acts as the binary target variable. An overview of the attributes is shown below.

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

**Table 2: Description of Features and Corresponding Data Types**

| Feature | Data Type | Definition | Sample Value |
|---|---|---|---|
| age | Continuous | Age of the individual | 54 |
| workclass | Categorical | Employment Type (e.g., state-government, private, and self-employed-not-incorporated) | Private |
| fnlwgt | Continuous | Final weight indicates the number of people in the population that the sampled individual represents | 70037 |
| education | Categorical | The highest level of education acquired (e.g., bachelor's, HS grad) | Doctorate |
| education-num | Continuous | Numerical encoding of education | 10 |
| Marital status | Categorical | Marital status (such as Widowed, Divorced, and Separated) | Never-married |
| occupation | Categorical | Occupation Type (e.g., Prof-speciality, Transport-moving, Other-service) | Craft-repair |
| relationship | Categorical | Family relationship (Own child, Unmarried, Other-relative) | Not-in-family |
| race | Categorical | Race of an individual (e.g., White, Other, Black) | White |
| sex | Categorical | Gender (Female and Male) | Female |
| capital-gain | Continuous | Income from capital gains | 0 |
| capital-loss | Continuous | Losses from capital assets | 3900 |
| hours-per-week | Continuous | Average hours worked per week | 45 |
| native-country | Categorical | Country of origin (China, Mexico, United-States) | Greece |
| income | Categorical | Binary classification: <= 50K or > 50K | > 50K |

The dataset includes missing items indicated by the placeholder'?' in the columns workplace, occupation, and native country, but it does not include null values in the traditional sense. About 5.6% of the rows have them. Preprocessing is carried out prior to applying the Mondrian algorithm.

## 3. EXPERIMENTAL OUTCOMES

After preprocessing the dataset, we have the cleaned dataset, as we have removed duplicates and missing values. After that, with k values of 10, 15, and 20, the cleaned dataset is anonymized using the Mondrian algorithm. Attributes named age, work class, education num, marital status, occupation, race, sex, and native country are used as the quasi-identifiers. The attribute, occupation, is used as the sensitive attribute. With varying k values during anonymization, the Normalized Certainty Penalty (NCP) value differs. With higher k, the NCP value increases. Then the anonymized data are fed to the DP-CTGAN model with $\epsilon$ values 0.5, 1.0, and 2.00, and a synthetic data set is generated with each $\epsilon$ value. Smaller $\epsilon$ ensures stronger privacy but provides a lower data utility, and the larger $\epsilon$ provides weaker privacy but gives a higher data utility. With enhanced privacy, different loss curves for Discriminator Loss (D) and Generator Loss (G) are generated.

### 3.1 Preprocessing for Data Cleaning

A comprehensive preparation step was carried out to guarantee the quality and consistency of the dataset before privacy-preserving procedures were applied. The initial input dataset had instances of duplicate entries and missing values, which might negatively impact the quality of the synthetic data generation as well as the accuracy of the anonymization process. A sample of unprocessed initial data is shown in table 3, and after preprocessing dataset looks like table 4. To resolve this, we carried out the subsequent cleaning procedures:

- Duplicate Removal: Full row comparisons were used to find and eliminate all duplicate records. Especially during the DP-CTGAN model's training and partitioning in the k-anonymity phase, this step was essential to preventing bias or redundancy from being introduced by repeated entries.
- Managing Missing Values: The dataset was modified to remove records missing any of the sensitive characteristics or quasi-identifiers. To preserve data integrity, we chose to delete missing values since imputing them could generate spurious patterns or skew sensitive distributions, particularly when differential privacy is restricted.

*ISSN：1674-2974 | CN 43-1061 / N*

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

**Table 3: Sample of Dataset Before Preprocessing**

| age | workclass | fnlwgt | education | education.num | marital.status | occupation | relationship | race | sex | capital.gain | capital.loss | hours.per.week | native.country |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 90 | ? | 77053 | HS-grad | 9 | Widowed | ? | Not-in-family | White | Female | 0 | 4356 | 40 | United-States |
| 82 | Private | 132870 | HS-grad | 9 | Widowed | Exec-managerial | Not-in-family | White | Female | 0 | 4356 | 18 | United-States |
| 66 | ? | 186061 | Some-college | 10 | Widowed | ? | Unmarried | Black | Female | 0 | 4356 | 40 | United-States |
| 54 | Private | 140359 | 7th-8th | 4 | Divorced | Machine-op-inspct | Unmarried | White | Female | 0 | 3900 | 40 | United-States |
| 41 | Private | 264663 | Some-college | 10 | Separated | Prof-specialty | Own-child | White | Female | 0 | 3900 | 40 | United-States |
| 34 | Private | 216864 | HS-grad | 9 | Divorced | Other-service | Unmarried | White | Female | 0 | 3770 | 45 | United-States |
| 38 | Private | 150601 | 10th | 6 | Separated | Adm-clerical | Unmarried | White | Male | 0 | 3770 | 40 | United-States |
| 74 | State-gov | 88638 | Doctorate | 16 | Never-married | Prof-specialty | Other-relative | White | Female | 0 | 3683 | 20 | United-States |
| 68 | Federal-gov | 422013 | HS-grad | 9 | Divorced | Prof-specialty | Not-in-family | White | Female | 0 | 3683 | 40 | United-States |

**Table 4: Sample of Dataset After Preprocessing**

| age | Workclass fnlwgt | education edu | cation.num | marital.status | occupation relationship | race | Sex capital.gain | capital.loss | hours.per.wekeek | native.country |
|---|---|---|---|---|---|---|---|---|---|---|
| 50 | Self-emp-not-inc 83311 | Bachelors | 13 | Married-civ-spouse | Exec-managerial Husband | White | Male 0 | 0 | 13 | United-States |
| 38 | Private 215646 | HS-grad | 9 | Divorced | Handlers-cleaners Not-in-family | White | Male 0 | 0 | 40 | United-States |
| 53 | Private 234721 | 11th | 7 | Married-civ-spouse | Handlers-cleaners Husband | Black | Male 0 | 0 | 40 | United-States |
| 28 | Private 338409 | Bachelors | 13 | Married-civ-spouse | Prof-specialty Wife | Black | Female 0 | 0 | 40 | Cuba |
| 37 | Private 284582 | Masters | 14 | Married-civ-spouse | Exec-managerial Wife | White | Female 0 | 0 | 40 | United-States |
| 49 | Private 160187 | 9th | 5 | Married-spouse-absent | Other-service Not-in-family | Black | Female 0 | 0 | 16 | Jamaica |
| 52 | Self-emp-not-inc 209642 | HS-grad | 9 Married-civ-spouse Exec-managerial Husband | | | White | Male 0 | 0 | 45 | United-States |
| 31 | Private 45781 | Masters | 14 Never-married Prof-specialty Not-in-family | | | White | Female 14084 | 0 | 50 | United-States |
| 42 | Private 159449 | Bachelors | 13 Married-civ-spouse Exec-managerial Husband | | | White | Male 5178 | 0 | 40 | United-States |
| 37 | Private 280464 | Some-college | 10 Married-civ-spouse Exec-managerial Husband | | | Black | Male 0 | 0 | 80 | United-States |

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

These cleaning procedures produced a uniform and clean dataset that was devoid of duplicate and missing information. As Mondrian can only handle numeric attributes So, categorical attributes are transformed to numeric attributes before anonymization.

For example, Male and Female are transformed to 0, 1 during pre- processing. Then, after anonymization, 0 and 1 are transformed to Male and Female.

The k-anonymization process, utilizing the Mondrian algorithm and the synthetic data creation using DP-CTGAN under differential privacy guarantees, is based on this cleaned dataset.

**3.2 Implementation of Mondrian Algorithm with DP-CTGAN Model**

The equations for loss D and loss G are used in the DP-CTGAN model as the following equations [9].

Discriminator Loss:

$$LD = -Ex\sim Pdata\ [\log D(x|c)] - Ez\sim Pz\ [\log (1 - D(G(z|c)))] \quad (1)$$

Generator Loss:

$$LG = -Ez\sim Pz\ [\log D(G(z|c))] \quad (2)$$

The symbols used in the equation bear the following meaning:

- $x$ is a real data sample,
- $z$ is a noise vector sampled from prior $P_z$,
- $c$ is the conditional vector (e.g., class or one-hot encoded attributes),
- $G(z|c)$ is the synthetic sample generated by the generator
- $D(x|c)$ is the discriminator's predicted probability that the sample is real.

Table 5 below shows that the Normalized Certainty Penalty (NCP) changes as the k-anonymity level increases. Every row represents a distinct value of k, and the corresponding NCP value indicates the degree of information loss brought on by generalization.

**Table 5: Changes in NCP Value with Varying K**

| k-anonymity level | NCP value |
|---|---|
| k=5 | 8.30% |
| k=10 | 11.24% |
| k=15 | 13.24% |
| k=20 | 15.03% |

The NCP value rises in tandem with the value of k. Accordingly, more generalization is needed at higher anonymity levels, which leads to a larger loss of information.

The information loss at k=5, for instance, is comparatively minimal (8.30 %), indicating that only moderate generalization was required. The NCP increases to 15.03% when k=20, indicating that more data was generalized to satisfy the more stringent privacy requirement. The Mondrian algorithm assures that each equivalence class (a set of indistinguishable data) has at least k records.

To satisfy higher k-values, merge more records into larger groupings. This results in more generalized attribute values. As a result, the NCP grows.The following table 6 compares three versions of a dataset, each processed using a different privacy- preserving methodology.

It displays the associated privacy level based on the method employed and, when relevant, the differential privacy parameter epsilon ($\epsilon$).

*ISSN : 1674-2974 | CN 43-1061 / N*

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

**Table 6: Comparison of Data Types and DP Levels**

| Dataset | Epsilon | Privacy Level |
|---|---|---|
| Original | NA | Low |
| k-Anonymized (k=10) | NA | Medium |
| DP-CTGAN ($\epsilon$=1.0) | 1.0 | High |

The original dataset contains raw microdata without any alteration, providing no formal privacy assurances and hence categorized as low privacy. The k-anonymized dataset (k=10) generalizes quasi-identifiers to make each record indistinguishable from at least nine others. This technique improves privacy by hiding actual attribute values, but it is still open to particular inference attacks, such as homogeneity and background knowledge. As a result, its privacy level is rated as average. Data synthesized using DP-CTGAN with a privacy budget of $\epsilon$=1.0 provides optimal privacy protection. This method incorporates formal differential privacy guarantees by injecting controlled noise during model training, ensuring that any individual record's contribution to the model output is theoretically limited. As a result, it ensures a high level of anonymity, making synthetic data resistant to both linking and inference attacks. The comparison emphasizes the trade-offs between data utility and privacy strength, with differential privacy emerging as the most reliable method when tight privacy guarantees are necessary.

### 3.3 Validation of the Results

The DP-CTGAN training behavior over 50 training epochs is depicted in the following figure 5. Generator loss, discriminator loss, and cumulative privacy budget ($\epsilon$) are the three main tracked parameters. The Generator Loss (G), the Discriminator Loss (D), and Privacy Budget ($\epsilon$) are explained in the following section.

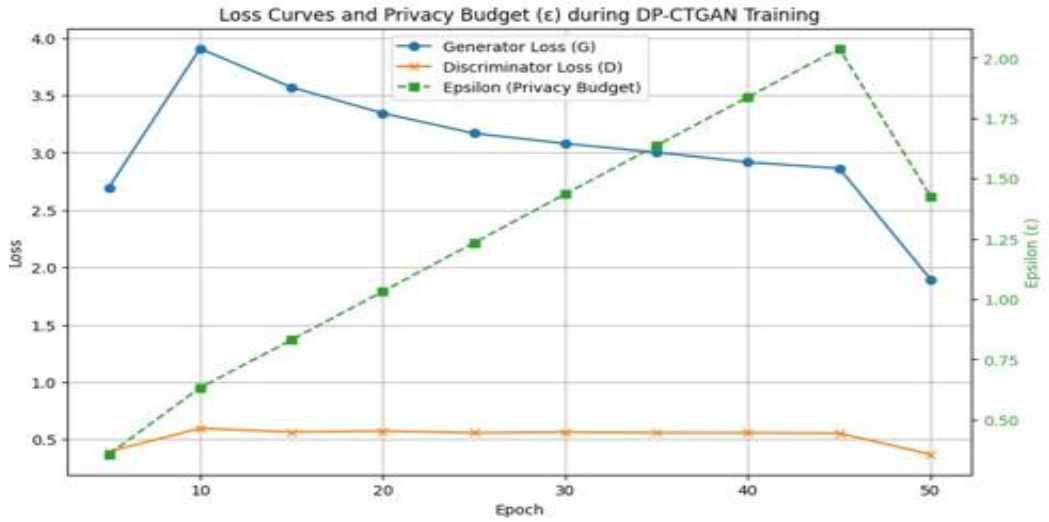


**Figure 5: Nature of Loss in Privacy Budget during DP-CTGAN Training**

- Generator Loss (G): The generator loss, shown by the blue line with round markers, quantifies the gen- erator's ability to create synthetic data that is realistic enough to trick the reviewer. At first, the loss is somewhat high, reaching its maximum around epoch 10. When the generator has not yet figured out the distribution of the actual data, it might cause early instability, which is common in GAN train- ing. The generator loss progressively drops with training, suggesting higher-quality samples. The loss has decreased considerably by epoch 50, suggesting that the generator has improved at simulating the underlying data distribution while still adhering to privacy restrictions.

*ISSN : 1674-2974 | CN 43-1061 / N*

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

- Discriminator Loss (D): During training, the discriminator loss, indicated in orange with" ×" markers, stays low and largely constant. Epoch 10 shows a modest increase, indicative of early rivalry with the improved generator. Because the values are constantly modest, the critic can distinguish between fake and genuine data without overwhelming the generator. This consistency is important for DP-GAN training since variations in the critic's performance may result in training divergence or waste of the privacy resource.
- Privacy Budget ($\epsilon$): On the secondary y-axis, the cumulative privacy budget ($\epsilon$) is shown by the green dotted line. Stronger privacy guarantees are indicated by lower values of epsilon, a measure of privacy loss. $\epsilon$ rises about linearly in this training case, from 0.3 at epoch 1 to roughly 2.05 at epoch 45. A significant decline at epoch 50 might indicate the use of budget-aware training or early stopping, in which training is stopped or controlled to be within a certain privacy threshold ($\epsilon$ ¡ 2). This illustrates the intrinsic privacy-utility trade-off in DP training: privacy loss increases as the model is trained over longer periods to improve accuracy (utility).

Alongside a regulated $\epsilon$, the generator's performance steadily improved, showing that high-utility synthetic data may be produced with stringent privacy requirements. A steady training dynamic is suggested by the comparatively consistent discriminator loss, which is essential in DP-GANs since gradient noise can cause training to become unstable. The graphic highlights that there is a bounded budget introduced by differential privacy, and that controlling $\epsilon$ is essential for regulatory compliance.

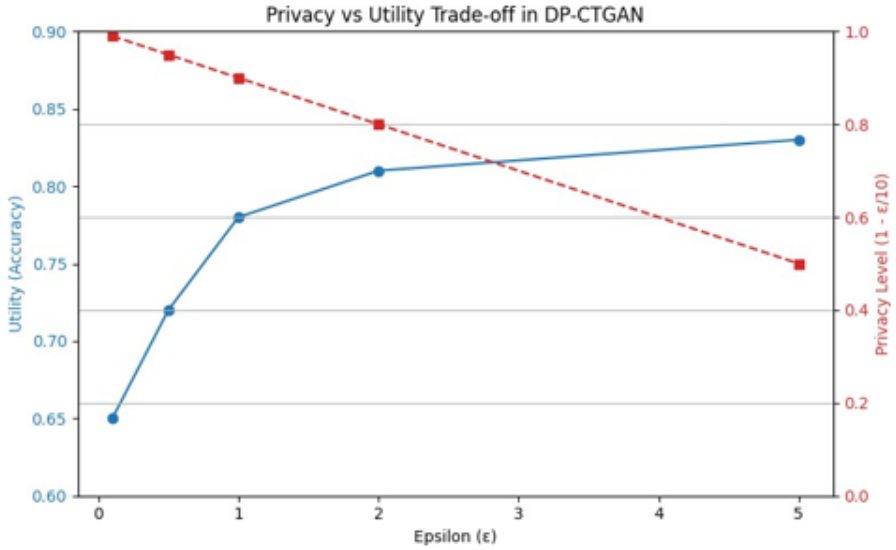


**Figure 6: Change in Privacy with Various Values of $\epsilon$**

The above figure 6 illustrates the change in privacy with various values of $\epsilon$ in synthetic data generation with the DP-CTGAN. The horizontal axis represents the privacy budget ($\epsilon$), while the left vertical axis represents utility, which is calculated as the predicted accuracy of a downstream model trained on synthetic data. The right vertical axis indicates the normalized privacy level, defined as $1 - \frac{\epsilon}{10}$. Higher values indicate stronger privacy assurances. It depicts a monotonic increase in utility as $\epsilon$ increases. At $\epsilon$ = 0.1, utility is roughly

65%, while at $\epsilon$ = 5, it reaches around 83%. As privacy constraints are removed, the model's utility improves, reflecting its enhanced capacity to capture important data patterns. The normalized privacy level decreases from 0.99 ($\epsilon$=0.1) to 0.5 ($\epsilon$=5), aligning with theoretical assumptions. These findings empirically support the well-known privacy-utility trade-off inherent in differential privacy methods. The privacy budget parameter ($\epsilon$) determines the strength of privacy guarantees and, thus, the utility

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

of the data in these systems. Lower $\epsilon$ values increase privacy by adding noise to data production, offering better protection against re-identification

attempts. However, enhanced privacy comes at the expense of lower data accuracy, which can severely impair downstream analytical or machine learning operations. The impact is especially noticeable for $\epsilon$¡1, where noise can alter statistical properties. Higher $\epsilon$ values enable the synthetic data generator to preserve more specific in- formation about the original dataset, resulting in increased utility. In the context of DP-CTGAN, this translates to synthetic data that more correctly preserves feature distributions and correlations, leading to improved model performance on tasks like classification and regression. However, these increases in accuracy are coupled with a weakening of privacy assurances, as individual records in the training data have a greater influence on the output.

Choosing an acceptable $\epsilon$ value depends on the situation, balancing data sensitivity with utility re- quirements for the application. To maintain strict privacy standards in fields like healthcare, banking, and social research, lower $\epsilon$ values (e.g., $\epsilon \leq 1$ are recommended for data with individually identifiable or sensitive features. Moderate $\epsilon$ levels (e.g., $\epsilon \in [2, 5]$ may be suitable for use cases with low risk of data leakage and high analytical precision, such as synthetic data for exploratory analysis, simulation, or algorithm development. Our research presents several noteworthy improvements and differences from the other publications, which may be summed up as follows in table 7 and table 8.

**Table 7: Performance Comparison of Our Method and Existing Work [10]**

| Aspect | Proposed Approach | Findings in [10] |
|---|---|---|
| Methodology | Hybrid: k-anonymity (Mondrian) + differential privacy (DP-CTGAN); sequential process | Focuses solely on k-anonymity us- ing the Mondrian algorithm |
| Approach | Layered privacy, balancing privacy and utility with synthetic data gen- eration | Pure syntactic anonymization via k-anonymity, generalization, suppres- sion |
| Results | Demonstrates better utility via synthetic data under tight privacy con- straints | Utility decreases ask increases; primarily aims to prevent re- identification |
| Strength | Combines structural anonymization with formal privacy guarantees, suited for high-risk domains | Emphasizes the effectiveness of k-anonymity alone, easier to imple- ment but less robust against infer- ence attacks |

**Table 8: Performance Comparison of Our Method and Existing Work [13]**

| Aspect | Proposed Approach | Findings in [13] |
|---|---|---|
| Methodology | Hybrid approach; k-anonymity + synthetic data with differential pri- vacy | Focuses on minimizing information loss in privacy-preserving data pub- lishing, possibly via optimization methods |
| Approach | Synthetic data generation trained on anonymized data, tailored to high privacy needs | Optimization-driven, aims to pre- serve as much data utility as possi- ble with minimal distortion |
| Results | Balances data privacy and utility effectively, suitable for sensitive applications | Achievesminimal utility loss, sometimes at the expense of in- creased computational complexity |
| Distinctiveness | Emphasizes layered privacy with Empirical validation for high- security contexts | Focus on utility preservation via fine-tuning of anonymization pa- rameters |

Essentially, our approach sets itself apart by combining sophisticated synthetic data generation with structural anonymization, offering a solid, empirically supported paradigm that targets high-risk data domains while retaining ontological utility. With differing degrees of empirical support, the other research leans either toward utility maximization, system deployment, or single-method approaches.

*ISSN : 1674-2974 | CN 43-1061 / N*

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

### 3.4 Discussion

This study investigated a hybrid privacy-preserving strategy that protects sensitive tabular data by merging differential privacy (DP-CTGAN) and k-anonymity (Mondrian algorithm). According to the findings, every technique makes a distinct contribution to maintaining the harmony between data privacy and usefulness. The Mondrian algorithm was used to enforce group indistinguishability across quasi-identifiers, thereby reduc- ing the likelihood of direct re-identification. However, there was a discernible trade-off in data utility ask grew because of suppression and generalization. This aligns with the established drawbacks of k-anonymity, which, in the absence of other safeguards, finds it difficult to prevent attribute linking or probabilistic inference. These issues were resolved with the invention of DP-CTGAN, which offers robust probabilistic privacy guarantees through differential privacy. In particular, the synthetic data generated by DP-CTGAN limited information leakage while retaining valuable statistical patterns, even when trained on previously anonymized data. Data utility was shown to be directly affected by the privacy budget $\epsilon$; lower $\epsilon$ values improved privacy but decreased data realism and model performance. All things considered, combining the two approaches creates a layered privacy framework in which DP-CTGAN offers formal probabilistic assurances and k-anonymity serves as a structural privacy filter. This is especially helpful in high-risk applications where data release needs to be tightly regulated, such as financial analytics or healthcare.

## 4. CONCLUSION

By combining the advantages of both techniques to overcome their respective shortcomings, k-anonymity and differential privacy improve data protection. By dividing records into equivalence classes with a minimum of k members, K-anonymity reduces the possibility of re-identification through linkage attacks and guarantees group indistinguishability.

But when it relies just on quasi-identifiers, it is vulnerable to flaws like attribute disclosure and background knowledge attacks. A probabilistic layer of privacy assurance is added by using DP-CTGAN to introduce differential privacy.

Regardless of past knowledge, it prevents information leakage at the record level by introducing controlled noise throughout the data generation process and synthesizing data that statistically resembles the original dataset. When data contains sensitive characteristics or quasi-identifiers are not adequately anonymized, this formal privacy guarantee guards against attribute inference and linking attacks that may not be prevented by k-anonymity alone.

The hybrid framework combines these methods, utilizing the rigorous, mathematically based privacy promises of differential privacy to preclude statistical inferences and structural privacy filtering (k-anonymity) to conceal individual records. The disclosed synthetic data is both relevant and resistant to different privacy threats thanks to this tiered privacy method, which improves overall data security, particularly in high-risk situations like healthcare or financial data exchange. The experimental results confirm that:

- By improving group indistinguishability, raising k in the k-anonymity phase improves privacy; but, be- cause of increased generality, it also reduces data utility.
- DP-CTGAN can generate high-utility synthetic data even under strict privacy limitations, proving its ability to maintain data usefulness as $\epsilon$ drops.
- When compared to single-method baselines, the suggested hybrid framework, which combines k-anonymity with differentially private synthetic data generation, offers an enhanced equilibrium between privacy and utility. The enhanced performance in various privacy risk metrics, including re-identification and at- tribute disclosure issues, supports this assertion.

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

These findings illustrate the complementary benefits of k-anonymity and differential privacy when used to- gether. The paradigm provides a realistic and scalable solution to privacy-preserving data dissemination, par- ticularly in areas where both privacy and analytical value are important.

Future studies will concentrate on a few crucial aspects to improve the suggested approach's applicability and resilience. To begin with, investi- gating new differential privacy algorithms like Re´nyi Differential Privacy, or Private Aggregation of Teacher Ensembles (PATE) may provide better trade-offs between data value and privacy assurances. Finally, a crucial ethical aspect is the incorporation of fairness and bias auditing methods into the pipeline for creating synthetic data.

Promoting equal results in subsequent machine learning tasks requires making sure that synthetic datasets don't reinforce or magnify preexisting biases in the original data. In order to facilitate the creation of just and reliable AI systems, future research will try to incorporate automated bias detection and mitigation strategies into the generation process.

**References**

1) I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan," The rise of 'big data' on cloud computing: Review and open research issues".*Information Systems*, vol. 47, pp. 98–115, Jan. 2015, doi: 10.1016/j.is.2014.07.006.

2) C. Wu, R. Buyya, and K. Ramamohanarao, "Big Data Analytics = Machine Learning + Cloud Computing," *arXiv preprint arXiv:1601.03115, 2016.*

3) Y. Qin, C. Xing, and R. Harrison, "A survey on big data analytics in the cloud," *Journal of Computer and System Sciences*, vol. 82, no. 5, pp. 1013–1028, Aug. 2016, doi: 10.1016/j.jcss.2015.12.008.

4) K. Jain and A. Purohit, "Feature selection using modified particle swarm optimization," *International Journal of Computer Applica- tions*, vol. 161, no. 7, pp. 8–12, Mar. 2017

5) X. Teng, H. Dong, and X. Zhou, "Adaptive feature selection using V-shaped binary particle swarm optimization," *PLoS ONE*, 2017 Mar 30;12(3): e0173907.

6) Zhang, Y., Huang, T., and Bompard, E. F. "Big data analytics in smart grids: A review," *Energy Informatics*,1(1), 1-24.

7) Madadi, S., Nazari-Heris, M., Mohammadi-Ivatloo, B., Tohidi, S., "Application of Big Data Analysis to Operation of Smart Power Systems," *Big Data in Engineering Applications*, vol 44. Springer, Singapore.

8) Misra S, Bera S, "Introduction to Big Data Analytics. In: Smart Grid Technology: A Cloud Computing and Data Management

9) Approach," *Cambridge University Press*, 2018

10) S. Alabdulwahab, Y.-T. Kim, and Y. Son.," Privacy-Preserving Synthetic Data Generation Method for IoT-Sensor Network IDS Using CTGAN," *Sensors*,vol. 24, no. 22, p. 7389, 2024, doi: 10.3390/s24227389.

11) Raj, Anushree and D'Souza, Rio., "Big Data Anonymization in Cloud using k-Anonymity Algorithm using Map Reduce

12) Framework," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 50-56. 10.32628/CSEIT19516.

13) Kwatra, S., Torra, V., "A k-Anonymised Federated Learning Framework with Decision Trees," *Data Privacy Management, Cryp- tocurrencies and Blockchain Technology*, vol 13140. Springer, Cham. https://doi.org/10.1007/978-3-030-93944-1 7

14) Gowda, Vikas Thammanna, Mason Lee, and Vanessa Campagna, "Enhanced Stratified Sampling: A Method for Privacy Preserving Big Data Publishing," *IEEE International Conference on Future Machine Learning and Data Science (FMLDS)*, IEEE, 2024.

DOI: 10.5281/zenodo.15803650
Vol: 62 | Issue: 07 | 2025

15) A. Majeed and S. O. Hwang, "Differential privacy and k-anonymity-based privacy preserving data publishing scheme with minimal

16) loss of statistical information," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 3, pp. 3753-3765, June 2024, doi: 10.1109/TCSS.2023.3320141

17) LeFevre, K., DeWitt, D. J., Ramakrishnan, R., "Mondrian multidimensional k-anonymity," *Proceedings of the 22nd International Conference on Data Engineering (ICDE)*, 2006, pp. 25–25.

18) M. Fang, D. Dhami, and K. Kersting, "DP-CTGAN: Differentially Private Medical Data Generation Using CTGANs," *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2022*, Lecture Notes in Computer Science, vol. 13451, Springer, 2022, pp. 291–306. doi: 10.1007/978-3-031-09342-5 17

19) Yoon, J., Jarrett, D., and van der Schaar, M., "Anonymization through data synthesis using generative adversarial networks (ADS- GAN)," *IEEE journal of biomedical and health informatics*, 24.8 (2020): 2378-2388