IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 January 2025, accepted 10 February 2025, date of publication 14 February 2025, date of current version 21 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3542266

RESEARCH ARTICLE

Can Synthetic Data Protect Privacy?

GIDAN MIN AND JUNHYOUNG OH
Department of Information Security, Seoul Women's University, Seoul 01797, Republic of Korea
Corresponding author: Junhyoung Oh (ohjun02@gmail.com)

This work was supported by a research grant from Seoul Women's University under Grant 2024-0043.

ABSTRACT To systematically evaluate the privacy protection performance of synthetic data generation algorithms (Synthpop, CTGAN, RTVAE, TVAE, DataSynthesizer), this study applied various safety metrics. Synthetic data is designed to protect sensitive information while maintaining statistical similarities to the original data, but a high degree of similarity can increase the risk of re-identification. Therefore, privacy protection was measured using metrics such as DCR, NNDR, Identification Risk Indicator, Inference Risk Indicator, CM3, DUPI, and pMSE. The results showed that Synthpop provided high data utility, but its high similarity to the original data posed significant privacy risks. Conversely, DataSynthesizer and CTGAN demonstrated superior privacy protection by balancing utility and privacy effectively. RTVAE and TVAE maintained a clear distinction from the original data, enhancing privacy protection, though some cases showed decreased data utility. These findings suggest the importance of selecting algorithms based on specific privacy and utility requirements, emphasizing the need to consider the trade-off between data utility and privacy protection depending on dataset characteristics.

INDEX TERMS Synthetic data, privacy protection, re-identification risk, data utility, safety metrics.

I. INTRODUCTION

Technological advancements over the past few decades have transformed nearly every scientific field [1]. In modern society, almost all personal information about our lives is recorded and stored digitally [2], and the data market is expected to continue its growth trajectory [3]. Many private and public institutions are also generating, analyzing, and storing data on behalf of stakeholders, users, and customers. As the volume of data increases, privacy concerns, such as the risk of dataset re-identification, have garnered attention. While it may be difficult to re-identify individuals with only a few attributes, the likelihood of successful re-identification increases rapidly as additional attributes are collected [4]. For example, Yves-Alexandre de Montjoye re-identified specific individuals with over 95% accuracy by combining de-identified data, such as hospital and mobility pattern data in the European Union [5]. In 2008, researchers Arvind Narayanan and Vitaly Shmatikov succeeded in re-identifying 84% of users by using just six ratings in combination with IMDb data [6]. This phenomenon, where combining pseudonymous data from different sources can re-identify an individual, is referred to as the problem of re-identification [4]. Moreover, if the public does not trust the organizations collecting and processing data, they may feel uncomfortable sharing their personal information or allowing it to be used for secondary purposes [7]. Several studies conducted in the U.S. have shown that individuals' attitudes toward privacy and confidentiality in relation to the census can predict their participation [8], [9]. For these reasons, it is crucial to ensure a low risk of re-identification in data to build public trust.

An innovative approach to addressing these issues is synthetic [10]. Synthetic data is artificially generated to maintain similar statistical characteristics to the original data while masking individual's attributes [11], [12].

Synthetic data has numerous applications in fields such as privacy protection, fairness, and data augmentation [13]. Models known as synthetic data generators can take various forms, including Generative Adversarial Networks (GANs) [10], Variational Autoencoders (VAEs) [11], agent-based and economic models [12], or a series of (stochastic) differential equations that model physical or economic systems [14]. The primary goal of synthetic data is to preserve statistical

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed Elhoseny.

properties while protecting sensitive data. As synthetic data generation gains popularity in both research and practical applications, more governments and private organizations are adopting this method as a means of protecting confidential information [15].

However, sensitive information can still be inferred from synthetic data, and biases inherent in the original data can persist. Since synthetic data often seeks to retain the distribution of the original data to maintain utility, it can be vulnerable from a privacy perspective [11]. In deep neural network-based approaches, membership inference attacks [16] can identify whether a specific input was included in the training data, which can be used to evaluate the similarity between synthetic and original data. Sensitive attributes, such as skin color, can be inferred from the behavior of deep learning models [17]. Even individual data points, such as a person's medical records or photograph, can be reconstructed [18]. In the case of generative AI models, the combination of the training process and the model's high complexity often leads to the generation of data that is highly similar to the original training data [19]. As a model generates more data, there is an increasing risk that it may produce data that closely resembles the original training data [20]. Additionally, there is the potential to determine whether specific data was used in the training process when using generative models. Synthetic data based on confidential information is relatively safe, but synthetic data that includes quasi-identifiers (e.g., name, birthdate) may pose a re-identification risk when combined with external datasets [14]. Furthermore, if features from the training data are memorized, these patterns may be replicated in synthetic data, potentially leading to personal information leakage [15]. Therefore, generating synthetic data that is both useful and privacy-preserving requires meticulous attention to detail [21].

This study proposes various evaluation metrics for validating the safety of synthetic data, along with methods for setting different thresholds depending on the characteristics of each dataset [22]. Unlike existing approaches, our methodology not only assesses the risk of re-identification but also establishes threshold values tailored to the specific characteristics of each dataset. By implementing a two-tiered data splitting strategy ($O_1, O_2, \ldots, O_{50}$ and $S_1, S_2, \ldots, S_{10}$), we ensure that the evaluation metrics are both robust and unbiased. The proposed safety metrics, including the Originality Index (OI) and Safety Index (SI), provide nuanced insights into the balance between data utility and privacy preservation. By obtaining the distribution of safety indicators when synthetic data is perfectly generated, we calculate threshold values. A guideline for safety is provided by measuring how far the safety indicators calculated from synthetic data generated by various algorithms deviate from the threshold [23]. This allows for a more precise and context-aware assessment of synthetic data's safety.

Through extensive experimentation using real-world datasets and multiple synthetic data generation algorithms, we validate the effectiveness and reliability of our safety verification process. The contributions of this study are twofold: firstly, we enhance the understanding of synthetic data's privacy risks through comprehensive safety metrics; secondly, we provide practical guidelines for organizations to assess and mitigate re-identification risks, thereby fostering greater public trust in data-driven applications.

The results of this study will contribute to strengthening data protection in various fields that handle sensitive information, thereby enhancing public trust. The safety metrics presented in this process can serve as useful tools for preemptively assessing and predicting the re-identification risk of synthetic data. Through this, users can set appropriate safety standards for specific synthetic data generators and datasets, and devise strategies to minimize re-identification risks.

II. LITERATURE REVIEW

A. RE-IDENTIFICATION RISK OF PSEUDONYMIZED INFORMATION

Pseudonymized information is data that is processed by removing identifying details to protect personal privacy, making it impossible to determine the original identity. However, if pseudonymized information is not sufficiently transformed or is combined with other additional information, the possibility of ''re-identification''—where the original identity can be determined—arises. Various studies have developed metrics and techniques to evaluate re-identification risk, with prominent methods including K-anonymity and l-diversity, which assess the level of anonymity in the data. K-anonymity ensures that at least k records in a dataset share the same attribute values, preventing individual records from being uniquely identified. L-diversity complements the limitations of k-anonymity by ensuring that diverse sensitive information exists within groups of records with identical attribute values, providing stronger protection.

Lubarsky [24] demonstrated that pseudonymized information from online shopping data could be re-identified by combining it with individual purchase records, and Ohm [25] highlighted the risk of breaking pseudonymized data's anonymity through correlations between specific variables. Recent studies frequently employ machine learning for re-identification techniques. Culnane et al. [26] conducted research on re-identifying pseudonymized medical data through pattern recognition, revealing the limitations of data protection. The Identification Risk Score (IRS) is used as a metric for assessing re-identification risk, measuring how many individuals in a pseudonymized dataset are likely to be re-identified [27]. These studies expose the security vulnerabilities of pseudonymized data and emphasize the need for stronger protection methods. Additionally, Gong et al. [28] developed a model to evaluate the vulnerabilities of pseudonymized data through adversarial attacks, stressing that without sufficient security measures, re-identification of pseudonymized data is easily possible.

Recent research seeks methods to minimize re-identification risk while maintaining data utility. One

such approach, Differential Privacy, provides a robust mathematical guarantee to balance privacy protection and data utility [29], [30]. Related studies continue to evolve to improve the level of protection for pseudonymized data and meet the stringent standards of data protection laws.

B. SYNTHETIC DATA ALGORITHMS

There are various methods for generating synthetic data, and selecting the appropriate method according to the type of data and intended use is crucial [20]. Common synthetic data generation methods include Sequential Modelling, Simulated Data, and Deep Learning Methods [31]. In this study, we used Synthpop, Conditional GAN (CTGAN), Tabular VAE (TVAE), Recurrent TVAE (RTVAE) and DataSynthesizer, each belonging to different synthetic data generation approaches.

Firstly, Synthpop is an R package that uses the Fully Conditional Specification (FCS) method, a sequential modeling approach. It generates synthetic data by conditionally modeling the relationships between variables in the original data [34]. FCS estimates the conditional distribution of each variable sequentially, preserving the statistical properties of the original data [32], [33]. Synthpop is particularly popular for synthesizing demographic data and data containing sensitive information [34]. CTGAN (Conditional Generative Adversarial Network) is a deep learning-based algorithm from the generative adversarial network (GAN) family. It learns the complex distributions of continuous and categorical data to generate synthetic data. CTGAN performs particularly well in environments where categorical and continuous data are mixed, maintaining data diversity while closely mimicking real data [35]. TVAE (Tabular Variational Autoencoder) is a deep learning method from the Autoencoder family that uses a Variational Autoencoder (VAE) structure to learn and synthesize data [36]. TVAE handles both continuous and categorical data, learning the latent space of the data to generate synthetic data that reflects the patterns of the original data [37]. It effectively learns complex distributions and characteristics of the data. TVAE is also part of MIT's SDV (Synthetic Data Vault) library. RTVAE (Recurrent Tabular Variational Autoencoder) is an extension of the Tabular Variational Autoencoder (TVAE) designed to handle tabular data with mixed continuous and categorical variables. RTVAE learns the temporal dependencies of the data, generating synthetic data that reflects the changing patterns of time-series data, making it useful when temporal changes in the data must be considered [38]. DataSynthesizer is a tool designed to generate synthetic data that preserves privacy while maintaining structural and statistical properties similar to sensitive datasets. This tool is specifically developed to enable secure and effective data generation and sharing without requiring data owners to configure specific parameters. It consists of three core modules: DataDescriber, DataGenerator, and ModelInspector. Particularly beneficial in fields with stringent privacy regulations, such as government, social sciences, and healthcare, DataSynthesizer facilitates data-driven collaboration by providing a simple and intuitive user interface, ensuring high accessibility. As an open-source tool, it allows users to generate and analyze synthetic datasets across a wide range of data applications [39].

C. EVALUATION OF THE UTILITY AND SAFETY OF SYNTHETIC DATA

Synthetic data, artificially generated based on real data, is widely used for privacy protection purposes. It preserves the privacy of data subjects while being utilized for data analysis or research purposes, which has led to extensive studies on its utility.

Goncalves [40] evaluated the similarity between real and synthetic data by assessing classification performance and consistency of data distribution to verify whether synthetic data maintains the same patterns as real data. Snoke et al. [41] demonstrated that synthetic data could provide sufficient accuracy to replace real data while reducing the risk of sensitive data leakage. This study emphasized the utility of synthetic data for data analysis, highlighting that synthetic datasets can protect privacy without significantly affecting analytical outcomes. The safety of synthetic data is also a crucial consideration in terms of re-identification risk. Bowen and Snoke [42]. analyzed potential information loss and privacy violations that could occur during the synthetic data generation process, emphasizing that synthetic data can be safely utilized only when appropriate privacy protection mechanisms are applied. They particularly concluded that applying differential privacy techniques during the modeling stage enhances data protection while generating synthetic data similar to real data. Recently, the use of Generative Adversarial Networks (GANs) for synthetic data generation has been actively researched [43]. Shen et al. [44]. used GANs to learn from real data and generate synthetic data with similar patterns, demonstrating the utility of synthetic data by producing high-performance analytical results. However, there are concerns that GAN-based synthetic data still pose a risk of information leakage, necessitating additional security measures during the generation process [45].

In conclusion, while the utility of synthetic data has been proven compared to real data, further research is needed to enhance its safety. By addressing these concerns, synthetic data can be utilized as a reliable and privacy-preserving alternative.

III. METHOD

A. DATA AND PREPROCESSING

In this study, three mixed-type tabular datasets from the UCI Machine Learning Repository were used to evaluate the privacy protection of synthetic data. Each dataset represents typical business scenarios where privacy-sensitive data assets need to be shared for analytical tasks [46]. Each record in the datasets represents an individual, and while the

privacy of these individuals must be protected, the statistical information of the entire dataset should be preserved [47].

- **Adult**: Contains 48,842 records and 15 attributes (6 numerical, 9 categorical).
  Available at: Adult Census Income Dataset
- **Bank-marketing**: Contains 45,211 records and 17 attributes (7 numerical, 10 categorical).
  Available at: Bank Marketing Dataset
- **Credit-default**: Contains 30,000 records and 24 attributes (20 numerical, 4 categorical).
  Available at: Credit Default Dataset

Numerical features were standardized using Z-score normalization, which scales the data to have a mean of 0 and a standard deviation of 1. This approach facilitates the convergence of machine learning models and ensures that each feature contributes equally when calculating distance-based safety metrics, such as Distance to Closest Record (DCR). Categorical features were transformed into binary matrices using One-Hot Encoding. This conversion enables categorical data to be represented as numerical data, allowing synthetic algorithms to process them effectively.

For all synthetic data generation algorithms, the data were split into training and testing sets using a 70:30 ratio. This split was implemented to evaluate the generalization performance of the models. Each algorithm was executed with the following training settings:

- **CTGAN**: Trained for a maximum of 10 epochs with a batch size of 128.
- **RTVAE**: Trained for up to 100 iterations, maintaining a batch size of 128.
- **TVAE**: Similarly trained for a maximum of 100 epochs with a fixed batch size of 128.
- **Synthpop**: Utilized default settings, with the seed set to 20040618 to ensure the reproducibility of results.
- **DataSynthesizer**: Used DataSynthesizer's correlated attribute mode to generate synthetic datasets. Bayesian network degree was set to 2, and differential privacy was controlled using an epsilon value of 1. A threshold value of 20 was used to identify categorical attributes, while the generated synthetic data contained 1000 records per dataset.

B. SYNTHETIC ALGORITHMS

The three datasets (Adult, Bank, Credit) were randomly divided into 50 sets of 3,000 records each, creating datasets assumed to be fully synthetic (from $O_1, O_2, \ldots, O_{50}$, hereafter referred to as 'O datasets'). These O datasets were used to calculate safety metrics (OI). Next, the three datasets were randomly divided into 10 sets of 5,000 records each, creating datasets from $S_1, S_2, \ldots, S_{10}$ (hereafter referred to as 'S datasets').

This two-tiered splitting approach allows for a robust assessment of synthetic data quality. By isolating O datasets for safety metrics and S datasets for synthetic data generation, we ensure that the evaluation metrics are not biased by overlapping data points. This separation enhances the reliability of the safety and utility assessments, as it prevents the artificial inflation of similarity measures that could occur if the same data points were used for both safety metric calculations and synthetic data generation. Synthetic data were then generated by applying each synthetic algorithm to the S datasets. The synthetic data algorithms used include the synthpop R package [34], an open-source generator from Gretel4, CTGAN and TVAE from MIT's SDV (Synthetic Data Vault) library, RTVAE from the synthcity library by vanderschaarlab, and DataSynthesizer [39], which applies differential privacy to protect sensitive data. All synthesizers were run with default settings without parameter tuning [46]. Specifically, synthetic data were generated by applying the synthpop, ctgan, RTVAE, TVAE, and DataSynthesizer algorithms to the S1 dataset, and the same process was repeated for S2 through S10. The generated synthetic data were then evaluated by averaging the values within each algorithm. Finally, SI values were calculated for each synthetic algorithm.

The threshold for the OI was set at the 95th percentile of the O datasets. This threshold aligns with the commonly used 95% confidence level in standard statistical practices, ensuring a balance between data utility and privacy protection [48]. In this context, the 95th percentile guarantees that the synthetic data do not exhibit excessive similarity to the original data. Additionally, this approach offers the flexibility to be adjusted based on specific use cases. For instance, datasets containing sensitive information, such as medical or financial data, may adopt stricter thresholds (e.g., the 90th percentile) to enhance privacy protection. Conversely, datasets with relatively lower privacy risks, such as those used for research purposes, can apply more lenient thresholds to achieve a balanced trade-off between data utility and analytical needs.

To evaluate the safety of synthetic data, SI values are compared against the OI threshold. The interpretation of these comparisons varies depending on the safety metric, as the potential for privacy violations depends on whether the SI value exceeds or falls below the OI threshold. For example, in some metrics (e.g., Inference Risk, DUPI), synthetic data is considered sufficiently distinct from the original data, with a low risk of privacy violations, when the SI value is below the OI threshold. Conversely, in other metrics (e.g., CM3, pMSE), synthetic data is deemed to have a low risk of privacy violations when the SI value exceeds the OI threshold. These interpretive criteria depend on the unique characteristics and objectives of each metric. Through such comparisons, the safety of synthetic data can be quantitatively and precisely evaluated.

However, the O datasets were excluded from the identification risk metric calculations because the identification risk metric is designed to assess the number of matching records between the original and synthetic data. The O datasets consist of data randomly extracted from the original data, resulting in significantly more overlapping instances with the

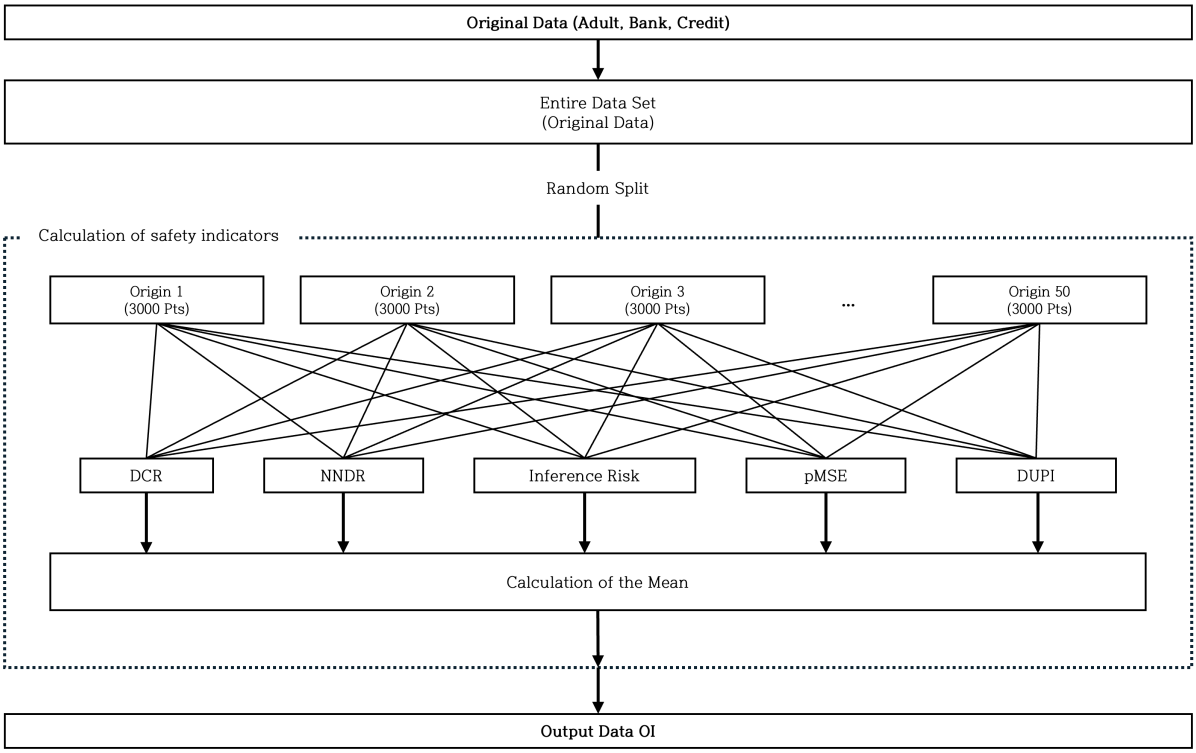


FIGURE 1. Results for splitting the original data, where 50 sets of 3000 data points are created from the original datasets (Adult, Bank, Credit).

original data than actual synthetic data. This overlap may not adequately reflect differences from the original data in the identification risk assessment, potentially distorting the true identification risk of the synthetic data. Therefore, to ensure the accuracy of this metric, the O datasets were excluded.

The selection of Synthpop, CTGAN, TVAE, RTVAE, and DataSynthesizer in this study was due to their effective handling of various data types and complexities. Synthpop generates synthetic data while preserving the statistical properties of data containing sensitive information through sequential modeling [34]. CTGAN, based on GAN, learns a diverse range of distributions, effectively mimicking categorical and continuous data [35]. TVAE uses a VAE structure to effectively learn continuous and categorical data, while RTVAE reflects temporal dependencies, maintaining patterns of time-series data [35]. DataSynthesizer, based on differential privacy, utilizes a Bayesian Network to learn correlations between attributes or employs noise-added histograms to analyze attribute distributions. This approach ensures strong privacy protection while maintaining statistical similarity to the original data [39]. The characteristics of these algorithms align well with the objectives of synthetic data generation in this study.

C. SAFETY METRICS

A total of seven methods are proposed as safety metrics to calculate OI and SI.

1) DCR

DCR is a metric that represents the Euclidean distance between each synthetic record and its nearest real neighbor [50]. When the DCR is zero, it indicates that real data could be exposed through the synthetic record, whereas higher DCR values suggest a reduced likelihood of privacy breaches [51].

$$\mathrm{DCR}(x, D_r) = \min_{y \in D_r} \|x - y\| \tag{1}$$

2) NNDR

Nearest Neighbor Distance Ratio (NNDR) represents the ratio of the Euclidean distance between each synthetic record's nearest neighbor and its second nearest neighbor [52]. For each synthetic data point, the closest and second closest real data points are identified. The NNDR value for each synthetic point is the ratio between the solid line (distance to the nearest neighbor) and the dashed line (distance to the second nearest neighbor).

$$\mathrm{NNDR}(x, D_r) = \begin{cases} \dfrac{\|y^{1st} - x\|}{\|y^{2nd} - x\|}, & \text{if } \|y^{1st} - x\| > 0 \\ 0, & \text{if } \|y^{1st} - x\| = 0 \end{cases} \tag{2}$$

NNDR falls within the range of 0 to 1, where larger values correspond to increased privacy protection.

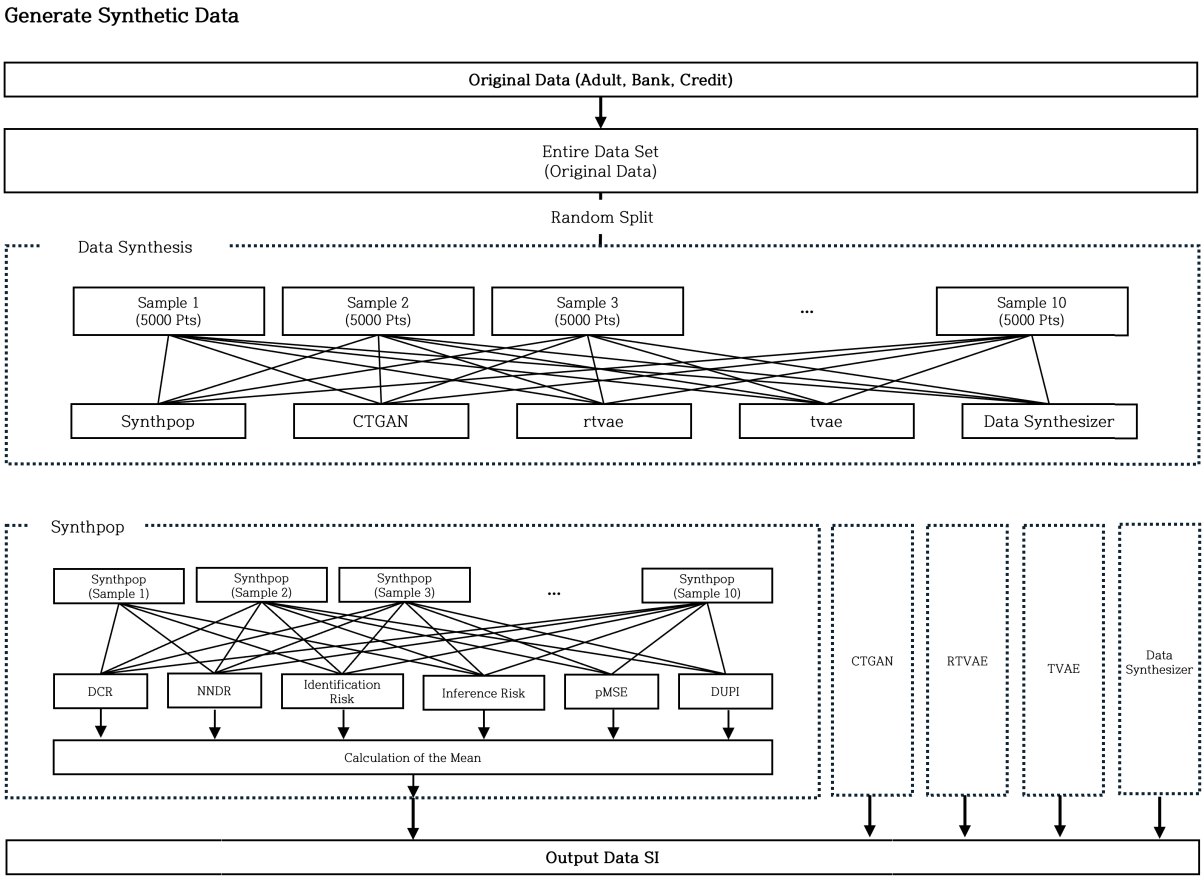


FIGURE 2. Results for generating synthetic data using *S*1 to *S*10 datasets, with different synthetic data generation algorithms applied.

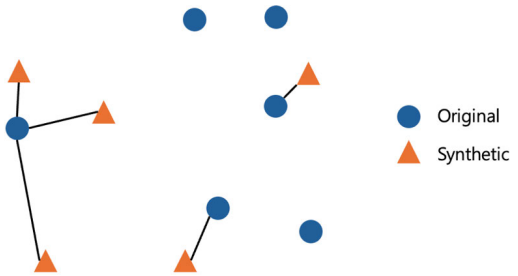


FIGURE 3. Illustration shows blue circles (original data), orange triangles (synthetic data), and black lines for nearest neighbor connections [49].

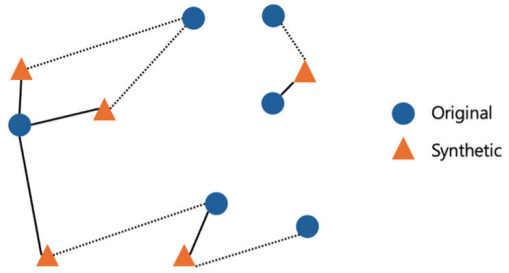


FIGURE 4. Illustration of DCR with solid lines for nearest neighbors and dashed lines for other connections between original and synthetic data points [49].

3) IDENTIFICATION RISK INDICATOR

Since synthetic data may generate more records than actual data points, there is a risk of exact synthetic matches overlapping with the original data [49]. The identification risk indicator calculates the ratio of matching records between synthetic data S_i and original data R_j on a per-record basis.

$$\text{Identification Risk} = \frac{1}{n}\sum I(S_i = R_i) \tag{3}$$

4) INFERENCE RISK INDICATOR

The Inference Risk Indicator measures the risk that the synthetic data contains many points similar to the original data, making it possible to infer the source. This is calculated by comparing the distance d_s between a synthetic data point and its closest original data point with the distance d_0, which is the closest distance between that original data point and any other original data point. The proportion of cases where $d_s < d_0$ is determined. If the distance between a synthetic data point and the nearest original data point is smaller than the distance to the second closest original data point, the risk of inference exposure is considered high. This occurs when synthetic data closely matches specific original data points.

$$A = \frac{1}{n_s}\sum a_i, \quad a_i = I(d_s < d_0) \tag{4}$$

5) CM3

CM3 is a metric for evaluating attribute exposure risk, calculated using Canonical Correlation. A CM3 value close to 1 indicates a low level of attribute exposure risk.

Algorithm 1 Phase for CM3 Calculation

Input: Original Dataset, Synthetic Dataset
Output: CM3: Calculating CM3

Data Preparation
  $X \leftarrow$ Original dataset
  $Y \leftarrow$ Synthetic dataset
Linear Combinations:
  Find linear combination for X:
    $U = a_1X_1 + a_2X_2 + \ldots + a_pX_p$
  Find linear combination for Y:
    $V = b_1Y_1 + b_2Y_2 + \ldots + b_pY_p$
Calculate Canonical Correlation:
  $CCA \leftarrow$ Calculate the correlation coefficient between U and V
Calculate CM3:
  $CM3 = 1 - CCA$
Repeat:
  Repeat steps 2 to 4 for the remaining dimensions to maximize the correlation.

6) pMSE (PROPENSITY MEAN SQUARED ERROR)

Propensity Mean Squared Error (pMSE) is a method for quantitatively evaluating the utility of synthetic data, measuring how similar the synthetic data is to the original data [53], [54]. This method is based on using a binary classifier to distinguish between original and synthetic data, and pMSE is calculated through the following steps. If the synthetic data is very similar to the original data, the pMSE value approaches 0, and as the difference between the two datasets increases, the pMSE value also increases [41].

To better interpret the pMSE results, the standardized pMSE or pMSE ratio can be calculated.

$$\text{Standardized pMSE} = \frac{\text{pMSE} - E|\text{pMSE}|}{\text{StDev}|\text{pMSE}|} \tag{5}$$

$$\text{pMSE Ratio} = \frac{\text{pMSE}}{E|\text{pMSE}|} \tag{6}$$

The standardized pMSE is calculated by dividing the difference between the observed pMSE and the expected pMSE by the standard deviation. It follows a distribution with a mean of 0 and a standard deviation of 1. A value close to 0 indicates a small difference between the original and synthetic data. The pMSE ratio is calculated by dividing the observed pMSE by the expected pMSE. A value close to 1 suggests that the synthetic data is similar to the original data.

7) DUPI (DATA UTILITY AND PRIVACY INDEX)

DUPI (Data Utility and Privacy Index) is a metric that measures the distributional similarity between the original data and the synthetic data to evaluate the quality of the synthetic data. DUPI considers the balance between the statistical similarity of the two datasets and the risk of personal information exposure. Ideally, a DUPI value close to 0.5 indicates the highest quality of the synthetic data.

Algorithm 2 pMSE Calculation

Input: Original Dataset, Synthetic Dataset
Output: pMSE: Calculating Propensity Mean Squared Error

Data Preparation
  $X \leftarrow$ Original dataset
  $Y \leftarrow$ Synthetic dataset
  $c = \frac{n_1}{n_0+n_1}$
  $Z = \{(x_i, 0) \text{ for } x_i \in X\} \cup \{(y_j, 1) \text{ for } y_j \in Y\}$
Train a Binary Classifier:
  The classifier f is trained to predict the label, where $\hat{y}_i$:
  $\hat{y}_i = f(z_i) = P(y_i = 1 \mid z_i)$
Estimate Propensity Scores:
  $\hat{p}_i = f(z_i)$
Calculate pMSE:

$$\text{pMSE} = \frac{1}{n_0 + n_1} \sum_{i=1}^{n_0+n_1} (\hat{p}_i - c)^2$$

$$DUPI^{\langle k \rangle} = \frac{1}{n} \sum_{i=1}^{n} I\left(d_{y_m}^{\langle k \rangle}(X_i) \leq d_{x_{n\setminus i}}^{\langle k \rangle}(X_i)\right) \tag{7}$$

The theoretical result is defined as $DUPI0\langle k \rangle$, and its sample estimator is denoted as $DUPI\langle k \rangle$, defined as above. Choosing a large k value makes the metric sensitive to extreme values or outliers in the data. Since k = 1 is sufficient to compare the characteristics of the data before and after synthesis, calculations are performed with $k = 1$ [55].

When comparing the calculated theoretical probability values with the experimental probability values, a large difference indicates that the synthetic data is not statistically similar to the original data, whereas a small difference suggests that the synthetic data is statistically similar to the original data.

The relationship between experimental and theoretical probabilities serves as an important metric for evaluating the characteristics of synthetic data and the level of privacy protection. If the experimental probability is similar to the theoretical probability, it indicates that the synthetic data appropriately reflects the characteristics of the original data while maintaining good privacy protection. On the other hand, if the experimental probability is lower than the theoretical probability, the synthetic data is less similar to the original data, which may reduce data utility but relatively enhance privacy protection. Lastly, if the experimental probability is higher than the theoretical probability, it suggests that

Algorithm 3 DUPI Calculation

Input: Original Dataset $X_n = \{x_1, x_2, \ldots, x_n\}$, Synthetic Dataset $Y_m = \{y_1, y_2, \ldots, y_m\}$

Output: DUPI Value (Experimental Probability and Theoretical Probability)

Data Preparation

$X \leftarrow$ Original dataset

$Y \leftarrow$ Synthetic dataset

Define the number of observations:

$n = |X|, m = |Y|$

Distance Calculation:

For each data point $x_i \in X$, compute the distances to all points in Y and $X \setminus \{x_i\}$.

$$d_{y_m(k)}(x_i) = \min\{d(x_i, y_j) : y_j \in Y\}$$

$$d_{x_{n\setminus i(k)}}(x_i) = \min\{d(x_i, x_j) : x_j \in X \setminus \{x_i\}\}$$

Calculate Experimental Probability:

Compare the k-th smallest distances for each data point x_i:

$$P\left(d_{y_m(k)}(x_i) \leq d_{x_{n\setminus i(k)}}(x_i)\right) =$$

$$\begin{cases} 1 & \text{if } d_{y_m(k)}(x_i) \leq d_{x_{n\setminus i(k)}}(x_i) \\ 0 & \text{otherwise} \end{cases}$$

Sum these comparisons across all n points in X:

Experimental Probability =

$$\frac{1}{n}\sum_{i=1}^{n} P\left(d_{y_m(k)}(x_i) \leq d_{x_{n\setminus i(k)}}(x_i)\right)$$

Calculate Theoretical Probability:

Use the binomial coefficient to calculate the theoretical probability:

Theoretical Probability =

$$\sum_{S=k}^{2k-1} \frac{1}{\binom{n+m-1}{k}} \cdot \binom{S-1}{k-1} \cdot \binom{n+m-S-1}{m-k}$$

Calculate DUPI

Condition = Compare($P_{\text{experimental}}, P_{\text{theoretical}}$)

the synthetic data is excessively similar to the original data, increasing the risk of personal information exposure.

IV. RESULT

A. DCR

DCR is a crucial metric for measuring how close synthetic data is to real data, evaluating the correlation between privacy and utility. The changes in DCR values for each dataset are as follows:

Among the algorithms, DataSynthesizer consistently demonstrated the highest DCR values across all datasets and training data ratios, starting at 20.28 ($DCR0.01$) for the Adult dataset and reaching 50.02 ($DCR0.9$) in the Credit dataset. This indicates a significant distance between the synthetic and original data, highlighting its strong privacy protection capabilities. Such performance may stem from its use of Bayesian networks and differential privacy mechanisms, which ensure robust privacy guarantees while maintaining the structural integrity of the data. RTVAE followed closely in terms of DCR values, particularly excelling in the Bank dataset ($DCR0.9$: 18.46). This suggests that RTVAE effectively balances privacy and utility, making it suitable for datasets with complex attribute correlations, such as Bank. Interestingly, its performance in the Credit dataset ($DCR0.9$: 23.54) was comparable to CTGAN ($DCR_{0.9}$: 23.71), but lower than DataSynthesizer. This might indicate limitations in handling datasets with highly variable distributions, as seen in Credit. The CTGAN algorithm maintained moderate DCR values across all datasets. Its performance in the Adult dataset ($DCR0.9$: 6.44) and Bank dataset ($DCR0.9$: 5.74) reflects a trade-off between privacy and utility, with a slight emphasis on utility. However, in the Credit dataset, CTGAN's higher DCR values ($DCR_{0.9}$: 23.71) suggest stronger privacy protection, potentially due to its ability to model complex data distributions through GAN architectures. This adaptability allows CTGAN to perform better in datasets like Credit, where the distributions are diverse and less structured compared to Adult or Bank. TVAE displayed inconsistent performance across datasets. While it achieved a sharp increase in DCR values in the Adult dataset ($DCR0.9$: 16.63), indicating improved privacy protection at higher percentiles, its overall values were lower than those of RTVAE or DataSynthesizer. In the Bank dataset, TVAE struggled in earlier percentiles ($DCR0.01$: 0.09) but gradually improved ($DCR0.9$: 2.71). Similarly, in the Credit dataset, TVAE showed moderate DCR values ($DCR0.9$: 15.75), suggesting that its privacy protection performance is highly dependent on the specific data distribution and attribute correlations. Synthpop consistently recorded the lowest DCR values across all datasets and training data ratios, with values such as 0.0001 ($DCR0.01$ in Adult) and 2.16 ($DCR0.01$ in Credit). This indicates that Synthpop generates synthetic data that closely resembles the original data, offering high utility but making it more vulnerable to privacy breaches. Its deterministic approach to replicating data structures and distributions may explain this behavior, as it prioritizes accuracy over privacy.

The differences in DCR values among datasets can be attributed to their inherent characteristics. The Adult dataset, with relatively simple attribute distributions, allowed algorithms like Synthpop and CTGAN to achieve higher utility, while DataSynthesizer excelled in maintaining privacy. The Bank dataset, characterized by stronger attribute correlations,

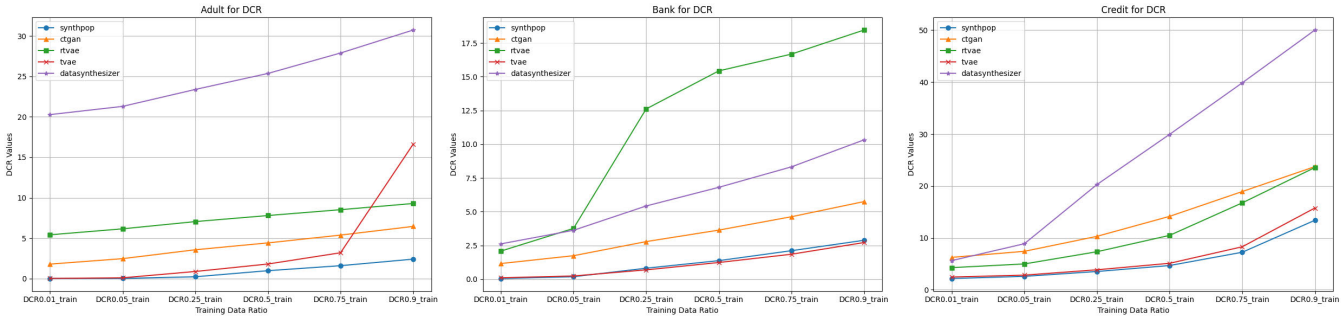


FIGURE 5. DCR values across different training data ratios for Adult, Bank, and credit datasets using synthpop, CTGAN, RTVAE, TVAE and DataSynthesizer.

favored models like RTVAE and DataSynthesizer, which leverage advanced probabilistic models to capture relationships while preserving privacy. The Credit dataset, with its diverse and complex distributions, posed challenges for all algorithms, but DataSynthesizer's adaptability and RTVAE's probabilistic approach allowed them to maintain stronger privacy protection.

Why Certain Algorithms Perform Better The superior performance of DataSynthesizer across datasets can be attributed to its use of Bayesian networks and differential privacy mechanisms, which provide robust protection even in complex datasets like Credit. RTVAE's consistent performance is due to its ability to model attribute correlations effectively, making it particularly strong in datasets like Bank. CTGAN, while versatile, relies on its GAN-based architecture, which performs well in capturing intricate data patterns but may struggle in datasets with simpler structures like Adult. TVAE's variability stems from its reliance on latent variable modeling, which can perform well in high-dimensional datasets but may underperform when attribute relationships are weak. Finally, Synthpop's deterministic approach ensures high utility but limits its ability to protect privacy, especially in datasets with complex attribute relationships or high variability.

B. NNDR

The NNDR metric provides insights into the distinctiveness of synthetic data compared to original data, with higher values indicating stronger privacy protection.

DataSynthesizer consistently achieved the highest NNDR values across all datasets and training scenarios, peaking at 0.998 (*NNDR*0.9 in Adult) and 0.995 (*NNDR*0.9 in Credit). This demonstrates its ability to maintain a clear distinction between synthetic and original data, offering robust privacy protection. In contrast, Synthpop displayed the lowest NNDR values in most scenarios, with 0.0003 (*NNDR*0.01 in Adult) and 0.303 (*NNDR*0.01 in Credit), indicating that its synthetic data closely resembles the original data, prioritizing utility but raising potential privacy concerns. CTGAN and RTVAE performed similarly across most datasets, maintaining high NNDR values at upper percentiles. For example, in the Bank dataset, both algorithms exceeded 0.99 (*NNDR*0.9), indicating strong privacy protection. TVAE, however, showed more variability, with a significant dip at mid-percentiles (*NNDR*0.5 in Adult: 0.67), suggesting that its privacy protection is inconsistent across different training scenarios.

C. IDENTIFICATION RISK METRIC

The Identification Risk Metric evaluates the extent to which synthetic data includes records identical to those in the original data, with higher values indicating a greater risk of data leakage.

The Synthpop algorithm recorded the highest Identification Risk Metric value of approximately 0.025 in the Adult dataset. This indicates that synthetic data generated by this algorithm contains a significant number of records matching the original data, suggesting high risk in terms of privacy protection. In contrast, CTGAN and TVAE achieved values close to 0 across all datasets, demonstrating minimal matching ratios with the original data and stable privacy protection performance. RTVAE exhibited a slightly increased value in the Credit dataset (0.00496), but maintained low values in other datasets, indicating relatively strong privacy protection. Similarly, DataSynthesizer recorded consistently low values across all datasets (e.g., 0.0020 in the Adult dataset), showing stable performance.

These results highlight that the Synthpop algorithm prioritizes data utility but may be vulnerable in terms of privacy protection. On the other hand, CTGAN and TVAE minimize the matching ratio with the original data, ensuring robust privacy protection, while RTVAE and DataSynthesizer also demonstrate stable performance.

D. INFERENCE RISK METRIC

The Inference Risk Metric represents the likelihood of inferring original data from synthetic data and plays a crucial role in evaluating the sensitivity of each dataset. For each dataset, the risk threshold was set using the 95th percentile (top 5 The Synthpop algorithm recorded relatively high Inference Risk Metric values across all datasets (Adult, Bank, Credit). In particular, the Bank dataset exceeded the risk threshold (0.46413), recording the highest value (0.5159). This suggests that the synthetic data generated by Synthpop closely resembles the original data, emphasizing

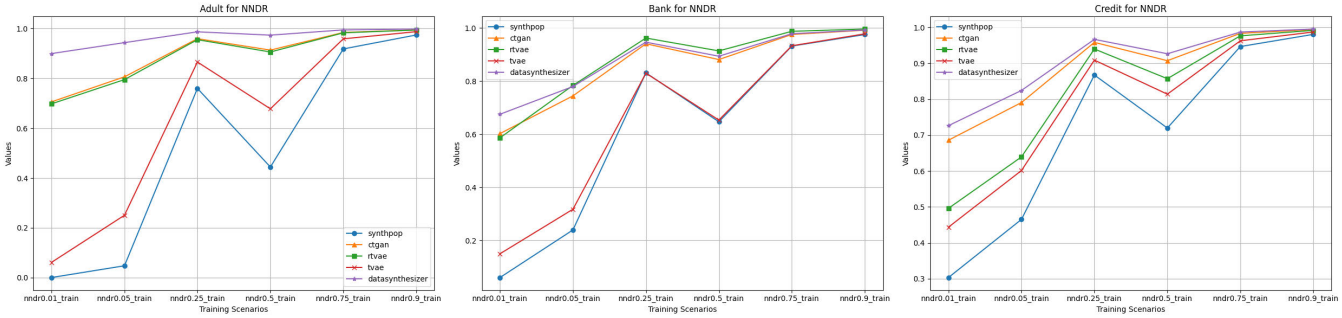


FIGURE 6. NNDR values across different training data ratios for Adult, Bank, and credit datasets using synthpop, CTGAN, RTVAE, TVAE and DataSynthesizer.

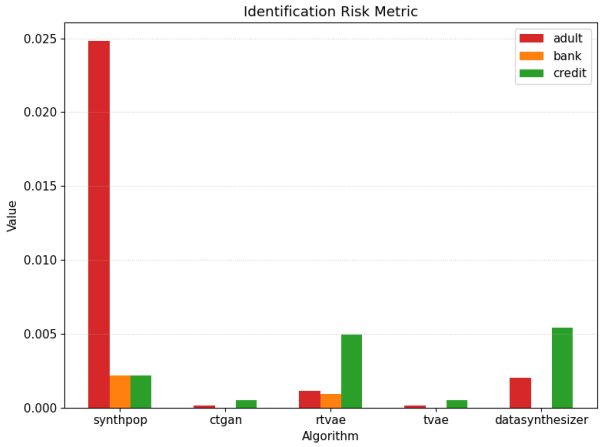


FIGURE 7. Identification risk metric values across different training data ratios for Adult, Bank, and credit datasets using synthpop, CTGAN, RTVAE, TVAE and DataSynthesizer.

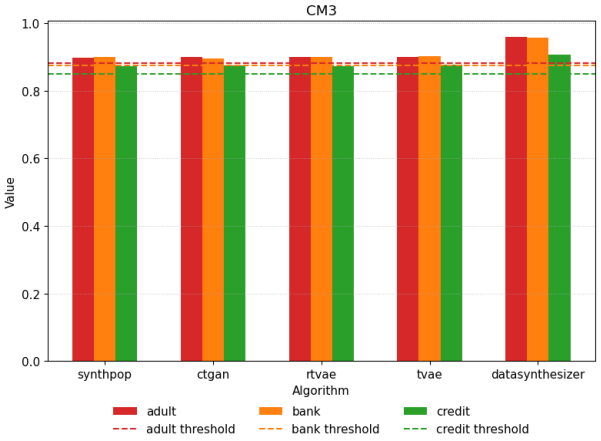


FIGURE 9. CM3 values across different training data ratios for Adult, Bank, and credit datasets using synthpop, CTGAN, RTVAE, TVAE and DataSynthesizer.

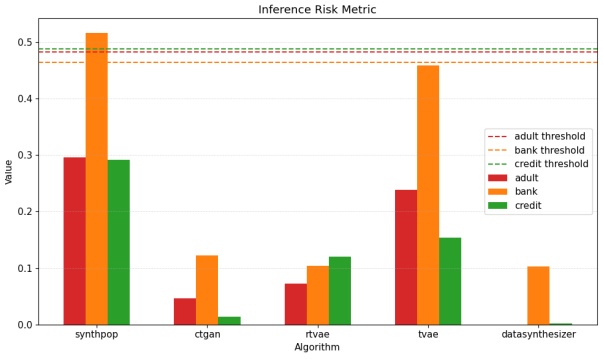


FIGURE 8. Inference risk metric values across different training data ratios for Adult, Bank, and credit datasets using synthpop, CTGAN, RTVAE, TVAE and DataSynthesizer.

the need for caution when using this algorithm with sensitive data. The TVAE algorithm recorded values close to the risk threshold in the Bank dataset (0.458), indicating a potential increase in privacy risk. In the Adult and Credit datasets, TVAE demonstrated moderate risk levels, providing relatively stable privacy protection. However, the elevated risk in the Bank dataset highlights the need for careful consideration when using the TVAE algorithm. The CTGAN algorithm recorded the lowest Inference Risk Metric values across most datasets, particularly in the Adult (0.0459) and Credit (0.0138) datasets. These results suggest that CTGAN is a reliable choice for privacy protection when dealing with sensitive data. The RTVAE algorithm showed moderate Inference Risk Metric values for the Adult (0.0723) and Credit (0.1202) datasets but recorded a relatively low value in the Bank dataset (0.1039), indicating strong privacy protection performance in this context.

The DataSynthesizer recorded the lowest Inference Risk Metric values across all datasets. It demonstrated stable performance below the threshold for the Adult (0.0001), Bank (0.1027), and Credit (0.0017) datasets. This highlights that the Bayesian network and differential privacy mechanisms applied in DataSynthesizer offer robust privacy protection.

In conclusion, CTGAN, RTVAE, and DataSynthesizer consistently provided stable privacy protection performance, while Synthpop and TVAE showed relatively higher risks depending on the dataset. These findings underscore the importance of considering dataset characteristics and sensitivity when selecting a synthetic data generation algorithm.

E. CM3

The CM3 metric evaluates the correlation between original and synthetic data using Canonical Correlation Analysis (CCA) to assess the risk of attribute exposure. A CM3 value closer to 1 indicates a lower risk of attribute exposure, suggesting that the synthetic data provides excellent protection by making it difficult to distinguish between the original and synthetic data. This graph compares CM3 values for the Synthpop, CTGAN, RTVAE, TVAE, and DataSynthesizer algorithms across the Adult, Bank, and Credit datasets. The threshold for each dataset was calculated by dividing the original data into 50 subsets of 3,000 records each, with the 95th percentile of calculated values set as the threshold.

DataSynthesizer recorded the highest CM3 values across all datasets, achieving 0.9583 in the Adult dataset, 0.9577 in the Bank dataset, and 0.9075 in the Credit dataset, significantly exceeding the thresholds. This demonstrates exceptionally low attribute exposure risk and outstanding data protection performance. DataSynthesizer consistently exhibited stable and reliable performance across all datasets. The Synthpop algorithm also recorded high CM3 values across all datasets, particularly excelling in the Adult dataset with a value of 0.8976. This indicates that distinguishing between the original and synthetic data is challenging, resulting in low attribute exposure risk. CTGAN demonstrated stable performance across datasets, achieving 0.8989 in the Adult dataset, 0.8963 in the Bank dataset, and 0.8745 in the Credit dataset. These results suggest that CTGAN provides consistent attribute protection across diverse datasets. RTVAE showed stable and high CM3 values in the Adult (0.8989) and Bank (0.8998) datasets but recorded a comparatively lower value of 0.8715 in the Credit dataset. This suggests that the complexity of the Credit dataset may have impacted the attribute protection performance of RTVAE. The TVAE algorithm performed exceptionally well in the Bank dataset, recording the highest CM3 value of 0.9023, indicating effective attribute exposure prevention. It also maintained high CM3 values in the Adult dataset and a stable performance in the Credit dataset with a value of 0.8761.

Additionally, it is noteworthy that the CM3 values of all four synthetic data generation algorithms exceeded the threshold for all three datasets. This indicates that each algorithm provides safe performance in preventing attribute exposure.

In conclusion, the CTGAN, TVAE, Synthpop, and DataSynthesizer algorithms demonstrated their effectiveness in minimizing attribute exposure risk. DataSynthesizer, in particular, exhibited consistent and superior performance across all datasets, making it a suitable choice for data protection. RTVAE showed some variability in the Credit dataset but recorded relatively high CM3 values in the Bank dataset.

F. PMSE

The Propensity Mean Squared Error (pMSE) measures the utility of synthetic data by quantifying the similarity between the original and synthetic datasets. A higher pMSE value indicates greater differences between the two datasets, suggesting that the synthetic data is more distinguishable from the original data. Conversely, a pMSE value closer to 0 implies that the synthetic data closely resembles the original data.

The graph visualizes the pMSE Mean, pMSE Ratio, and pMSE Standardized values for the Synthpop, CTGAN, RTVAE, TVAE, and DataSynthesizer algorithms across the Adult, Bank, and Credit datasets. The threshold for each dataset was set based on the 95th percentile of values calculated from 50 subsets of 3,000 records each from the original data.

In pMSE Mean, the DataSynthesizer algorithm recorded the highest values across all datasets, particularly exceeding 0.11 in the Credit dataset, indicating that it generates synthetic data that is clearly distinguishable from the original data. CTGAN followed with the second-highest values, while RTVAE exhibited mid-level pMSE Mean values, maintaining stable performance. TVAE recorded the lowest pMSE Mean values, particularly in the Credit dataset, suggesting a high similarity to the original data. Synthpop consistently recorded the lowest pMSE Mean values, indicating that it generates synthetic data highly similar to the original data.

For pMSE Ratio, DataSynthesizer exhibited the highest values across all datasets, showing clear distinctions between the original and synthetic data. CTGAN also recorded high pMSE Ratio values, demonstrating similar trends. RTVAE and TVAE maintained stable pMSE Ratio values, with TVAE showing the lowest value in the Credit dataset. Synthpop consistently recorded the lowest pMSE Ratio values, maintaining high similarity to the original data.

In pMSE Standardized, DataSynthesizer showed the highest values across all datasets, with the most pronounced differences observed in the Credit dataset. CTGAN followed with the second-highest values, maintaining stable performance. RTVAE and TVAE demonstrated stable performance in the Bank dataset, with some variability observed in the Credit dataset. Synthpop recorded the lowest pMSE Standardized values across all datasets, indicating high similarity to the original data.

In conclusion, DataSynthesizer achieved the highest pMSE values across all metrics, making the distinction between the original and synthetic data most evident, demonstrating superior performance in terms of data utility. Conversely, Synthpop consistently exhibited low pMSE values, maintaining a close resemblance to the original data. CTGAN demonstrated stable utility, ranking second after DataSynthesizer, while RTVAE and TVAE maintained consistent performance across datasets, with TVAE showing high similarity to the original data in specific datasets.

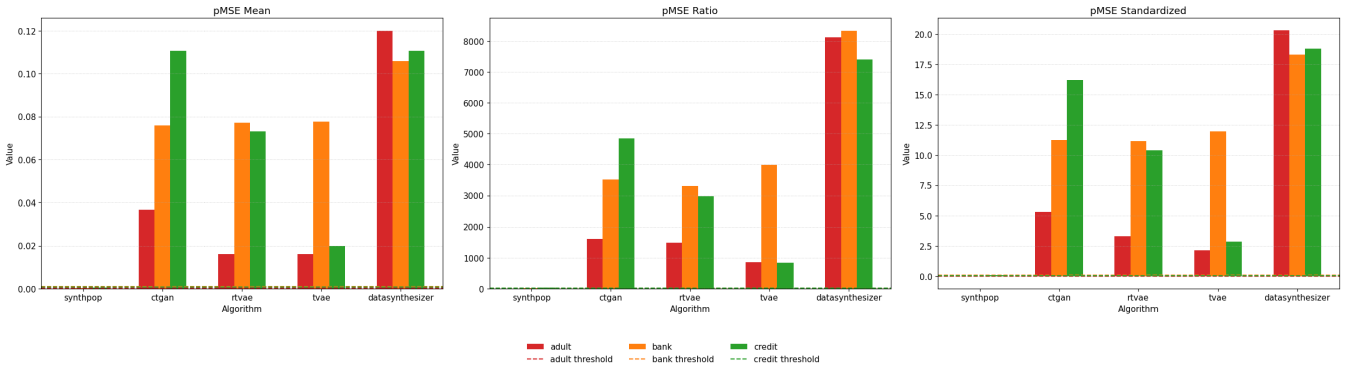


FIGURE 10. pMSE values across different training data ratios for Adult, Bank, and credit datasets using synthpop, CTGAN, RTVAE, TVAE and DataSynthesizer.

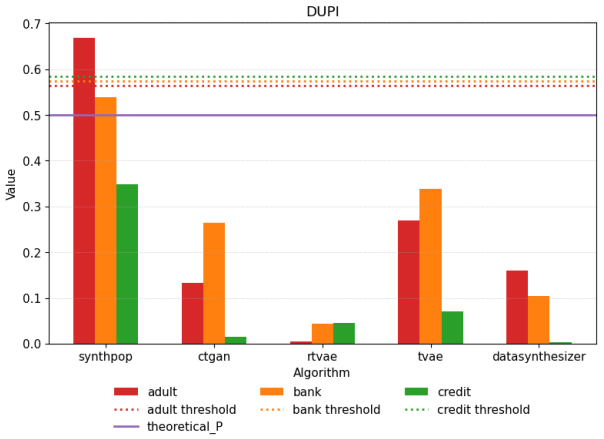


FIGURE 11. DUPI values across different training data ratios for Adult, Bank, and credit datasets using synthpop, CTGAN, RTVAE, TVAE and DataSynthesizer.

G. DUPI

The Data Utility and Privacy Index (DUPI) is a metric used to evaluate how well synthetic data reflects the statistical properties of the original data while effectively controlling the risk of personal information exposure. A DUPI value close to the theoretical probability (theoretical P) indicates an appropriate balance between data utility and privacy protection. If the DUPI value is lower than the theoretical probability, it suggests reduced data utility but strong privacy protection. Conversely, a DUPI value higher than the theoretical probability indicates increased similarity to the original data, which may lead to a higher risk of privacy exposure.

The updated graph visualizes the DUPI values for the Synthpop, CTGAN, RTVAE, TVAE, and DataSynthesizer algorithms across the Adult, Bank, and Credit datasets. The theoretical probability for each dataset is shown as a solid line, and the 95% thresholds are represented as dashed lines. This theoretical probability is calculated under the assumption that the original and synthetic data are independently drawn from the same population. Values close to the theoretical probability indicate high similarity between the datasets, suggesting that appropriate privacy protection is maintained while preserving data utility.

The Synthpop algorithm recorded the highest DUPI value in the Adult dataset, indicating that the synthetic data may be overly similar to the original data. While this provides high data utility, it also increases the risk of privacy exposure. In contrast, the CTGAN algorithm consistently recorded low DUPI values across all datasets, with the lowest value in the Credit dataset. This demonstrates excellent privacy protection by maintaining a clear distinction between the original and synthetic data. The RTVAE algorithm recorded the lowest DUPI value in the Bank dataset, performing close to the theoretical probability. This indicates that RTVAE effectively balances data utility and privacy protection. The TVAE algorithm exhibited a relatively high DUPI value in the Bank dataset, suggesting increased similarity to the original data and a potential rise in privacy exposure risk. However, it maintained moderate performance in the Credit dataset. The DataSynthesizer algorithm recorded the lowest DUPI value in the Credit dataset, demonstrating its ability to maintain a clear distinction between the original and synthetic data. It also showed stable performance in the Adult and Bank datasets, maintaining an appropriate balance between data utility and privacy protection.

In conclusion, the CTGAN and DataSynthesizer algorithms consistently recorded low DUPI values, demonstrating excellent performance in privacy protection. Conversely, the Synthpop and TVAE algorithms exhibited higher similarity in certain datasets, indicating high data utility but an increased risk of personal information exposure. The RTVAE algorithm performed close to the theoretical probability, maintaining a balanced performance between data utility and privacy protection.

V. DISCUSSION

In this study, we evaluated the privacy protection performance of synthetic data generated using various algorithms

(Synthpop, CTGAN, RTVAE, TVAE, and DataSynthesizer) and analyzed how each algorithm maintains the balance between data utility and privacy protection. These results provide important insights into how well synthetic data preserves the utility necessary for analytical tasks while maintaining data privacy.

One of the most critical aspects of synthetic data generation is the trade-off between data utility and privacy protection. High data utility ensures that synthetic data maintains the statistical properties required for meaningful analysis, while robust privacy protection minimizes the risk of re-identification and other privacy breaches.

A. COMPARISON WITH PREVIOUS STUDIES

The Synthpop algorithm demonstrated high similarity to the original data across most metrics, leading to a relatively high re-identification risk, thus showing poor performance in terms of privacy protection. While this result highlights the advantage of high data utility, it also aligns with previous studies indicating that Synthpop does not fully achieve the primary goal of privacy protection.

According to a study by Fry et al. (2021), Synthpop was assessed as having excellent data utility but relatively low privacy protection performance. This issue arises because Synthpop excessively preserves the statistical properties of the original data, where high utility can, paradoxically, lead to privacy breaches. The study particularly emphasized that for data with high re-identification potential, such similarity could increase re-identification risks, suggesting that privacy enhancements are necessary when using Synthpop [56]. Similarly, Snoke et al. [57] found comparable results, noting that while Synthpop generates synthetic data highly similar to the original data by maximizing utility, it compromises privacy protection, thus increasing re-identification risks. This study cautioned that despite the high reliability of analysis results due to the close similarity between Synthpop's synthetic and original data, the method might not sufficiently suppress re-identification risks [57].

These findings are consistent with our study, underscoring the need for additional privacy enhancement techniques. For instance, integrating Differential Privacy (DP) methods or data masking techniques could help mitigate re-identification risks.

The CTGAN algorithm showed relatively superior privacy protection performance across most metrics, indicating that it strikes a good balance between privacy and utility in synthetic data. These results are consistent with the findings of Nguyen et al. [58], which highlighted that CTGAN's flexibility enables balanced performance across various datasets. The study praised CTGAN's ability to achieve both privacy and utility by recording low values in identification and inference risk metrics while accurately reflecting the statistical properties of the data.

However, the results still suggest that some similarity to the original data is not entirely eliminated in certain situations, highlighting the need for further research to strengthen privacy protection performance. For example, fine-tuning the hyperparameters of the CTGAN model or introducing additional noise during the training process to enhance privacy protection could be considered [59].

B. IMPLICATIONS FOR ALGORITHM SELECTION

Synthpop prioritized data utility by exhibiting high similarity to the original data; however, this approach increased the risk of re-identification. In contrast, CTGAN and DataSynthesizer employed mechanisms to reduce the similarity between synthetic and original records, thereby enhancing privacy protection while maintaining reasonable data utility. This balance is influenced by the characteristics of the dataset, such as the ratio of categorical to numerical features. For instance, in datasets like Bank, which exhibit strong attribute correlations, DataSynthesizer and RTVAE achieved high DCR and NNDR values, providing robust privacy protection. However, additional fine-tuning optimized for handling categorical data may be necessary. Conversely, for simpler datasets like Adult, Synthpop maintained high data utility but posed relatively high privacy risks, highlighting the unique challenges posed by categorical features in preserving privacy. These results emphasize the need for tailored approaches to specific dataset characteristics, underscoring the importance of fine-tuning to achieve a balance between utility and privacy across diverse datasets.

In datasets with simple structures and low attribute correlations, such as Adult, algorithms like Synthpop and CTGAN are particularly effective. Synthpop maximizes data utility by accurately replicating data distributions through its deterministic approach. Similarly, CTGAN's GAN-based architecture excels at capturing simple data patterns effectively without managing complex interdependencies. However, this high accuracy increases privacy risks as synthetic data closely resembles the original data. On the other hand, datasets like Bank and Credit, characterized by high complexity and strong attribute correlations, present unique challenges. In these contexts, RTVAE and DataSynthesizer demonstrated superior performance by leveraging advanced probabilistic models and Bayesian networks, respectively. RTVAE effectively models complex attribute relationships, maintaining privacy without significantly compromising data utility. DataSynthesizer, integrating differential privacy mechanisms, preserves intricate interdependencies within the data while providing robust privacy guarantees. These algorithms excel at managing the subtle relationships inherent in more complex datasets, making them preferable choices when privacy protection is critical. The Credit dataset, characterized by diverse distributions and high variability in both categorical and numerical features, posed significant challenges for synthetic data generation. DataSynthesizer effectively managed this complexity through its Bayesian network structure and differential privacy techniques, achieving superior performance. CTGAN also demonstrated similar

TABLE 1. Summary of metrics across datasets and algorithms.

Dataset	Algorithm	DCR ↑	NNDR ↑	Identification Risk ↓	Inference Risk ↓	CM3 ↑	DUPI ↓
Adult	DataSynthesizer	**50.02**	**0.998**	0.0020	0.0001	**0.9583**	**Low**
	RTVAE	18.46	0.99	0.00496	0.0723	0.8989	Moderate
	CTGAN	6.44	0.99	**0.0000**	**0.0459**	0.8989	Moderate
	TVAE	16.63	0.67	**0.0000**	0.458	0.9023	High
	Synthpop	0.0001	0.0003	0.0250	0.46413	0.8976	High
Bank	DataSynthesizer	**50.02**	**0.998**	0.0020	0.1027	**0.9577**	**Low**
	RTVAE	18.46	0.99	0.00496	**0.1039**	0.8998	Moderate
	CTGAN	6.44	0.99	**0.0000**	0.45	0.8963	Moderate
	TVAE	16.63	0.67	**0.0000**	0.458	0.9023	High
	Synthpop	0.0001	0.0003	0.0250	**0.5159**	0.8976	High
Credit	DataSynthesizer	**50.02**	**0.998**	0.0020	**0.0017**	**0.9075**	**Low**
	RTVAE	23.54	0.99	0.00496	0.1202	0.8715	Moderate
	CTGAN	23.71	0.99	**0.0000**	0.0138	0.8745	Moderate
	TVAE	15.75	0.67	**0.0000**	0.45	0.8761	High
	Synthpop	2.16	0.303	0.0250	0.5159	0.8976	High

effectiveness due to its adaptable GAN architecture capable of modeling diverse data patterns. In contrast, Synthpop struggled in such complex environments as its deterministic replication approach failed to adequately capture the variability and diverse distributions, thereby compromising privacy protection.

Algorithm selection inherently involves balancing data utility and privacy protection. In sensitive domains such as healthcare and finance, where data privacy is paramount, selecting an appropriate synthetic data generation algorithm is essential. CTGAN and DataSynthesizer have proven suitable for these domains through their balanced approaches to privacy and utility. On the other hand, Synthpop, despite its high data utility, poses higher re-identification risks, making it more appropriate for scenarios where privacy is less critical, such as initial data exploration or educational purposes.

For example, consider the development of a patient readmission prediction model in the healthcare domain using synthetic electronic health records (EHR). Synthpop reproduces the original data with high precision, thereby providing high data utility that can substantially enhance the performance of the predictive model. However, this high degree of similarity may also increase the risk of re-identifying individual patient information [60]. In contrast, algorithms such as CTGAN or DataSynthesizer, which emphasize privacy protection, tend to reduce the resemblance to the original data, significantly lowering the risk of re-identification. In fact, the effectiveness of these approaches in safeguarding the privacy of medical data has been reported [61], suggesting that they are the preferable choice in settings where patient privacy is paramount.

In the case of customer behavior analysis for the formulation of initial marketing strategies, high data utility is imperative. Synthpop's ability to accurately replicate the original data ensures that detailed statistical characteristics—such as customer behavior patterns and purchasing trends—are well preserved. These features play a crucial role in predicting consumer responses prior to a new product launch, in market segmentation, and in the development of tailored marketing strategies. In early-stage marketing analyses, where data utility is the primary consideration and where the target data has already been anonymized or stripped of personally identifiable information, the risk of re-identification may be considered relatively low [62], [63]. Thus, in situations demanding high data utility, selecting an algorithm like Synthpop—which generates synthetic data with a high degree of similarity to the original—can be a rational strategy to maximize the predictive accuracy of analytical models. This example demonstrates that in scenarios where the protection of sensitive information is relatively less critical—such as during preliminary exploratory stages, for educational purposes, or in environments where the data has already been anonymized—a lower level of privacy protection may be acceptable, thereby significantly contributing to strategic decision-making.

Additionally, the ability to fine-tune algorithm parameters to achieve desired levels of privacy and utility allows practitioners to customize synthetic data generation to specific use cases. For example, adjusting the epsilon value in DataSynthesizer can provide a tailored balance between privacy and data utility, enabling compliance with various regulatory requirements and privacy standards. As datasets continue to grow in size and complexity, the scalability and flexibility of synthetic data generation algorithms become increasingly important. Future research should focus on developing scalable solutions that can adapt to the growing complexity and volume of data, ensuring consistent performance across diverse datasets. This scalability is essential for real-world applications where data characteristics continuously evolve and vary over time.

C. LIMITATIONS AND FUTURE DIRECTIONS

This study focused on specific datasets and algorithms, which may limit the generalizability of the results. Additionally, the safety metrics used in this study may not fully capture

all potential privacy breach scenarios due to their inherent limitations.

Distance-based metrics such as DCR and NNDR evaluate privacy protection performance by measuring the distance between synthetic records and original records. However, these metrics may not fully reflect the actual risk of re-identification. Since these metrics assess privacy based solely on distance, they may overlook the potential for privacy breaches caused by attribute-specific or contextual factors. The Identification Risk metric assesses privacy risk based on exact record matches, which may underestimate privacy risks in scenarios involving partial matches or attribute inference attacks. While CM3 and pMSE are useful for evaluating correlation and classification-based similarities, they do not account for more sophisticated attack vectors that exploit subtle statistical patterns in the data. To address these limitations, future research should incorporate a variety of safety metrics, such as those based on Differential Privacy principles or Membership Inference Attack-based evaluations, to offer a more thorough evaluation of privacy protection.

Furthermore, this study focused on specific datasets and algorithms, which may limit the generalizability of the results. Various types of data, such as images, text, and time-series data, present unique challenges that require specialized synthetic data generation techniques. For instance, in the case of time-series data, specialized modeling techniques tailored to time series are required to effectively capture the temporal dependencies and sequential patterns among the data. If the algorithms employed do not sufficiently reflect this temporal continuity, the quality and utility of the synthetic data may be compromised. Moreover, unstructured data exists in various forms such as text, images, and audio, and entails additional complexities in the data preprocessing and feature extraction stages. Consequently, there is a risk that the evaluation metrics or models used in this study may not fully capture the unique characteristics of unstructured data. Algorithms like CTGAN and TVAE are effective for tabular data, whereas Generative Adversarial Networks (GANs) such as StyleGAN and BigGAN are better suited for capturing the complex visual patterns of image data. Moreover, the performance of synthetic data generation algorithms can vary significantly across different domains. Future research should explore the applicability of these algorithms across diverse data types and domains to validate their effectiveness and identify domain-specific optimizations.

Although these models are recognized as powerful tools for privacy-preserving data sharing, they have predominantly been applied in the image domain since the current experiments focus on synthesizing text CSV files. In future research, it will be necessary to integrate these advanced generative models with differential privacy mechanisms to explore methods for mitigating privacy risks while maintaining data utility.

Additionally, algorithm optimization is necessary to address overfitting issues and enhance robustness. In the current study, models optimized for specific groups may produce inaccurate results on certain data due to overfitting. To mitigate this issue, it is essential to use more extensive datasets that include a diverse range of ages, genders, and cultural backgrounds, and to combine normalization techniques with various data augmentation methods during the training process. Future research should also consider the latest advancements in privacy-preserving AI models. Specifically, considering the potential emergence of new forms of privacy breaches, ongoing research is needed to enhance the responsiveness of algorithms. Future research must take into account the latest advancements in privacy-preserving AI models. In particular, continuous efforts are needed to enhance algorithmic resilience in anticipation of emerging forms of privacy breaches. Technologies such as Federated Learning and Differential Privacy play a critical role in fostering data sharing and collaboration while providing robust privacy guarantees. Federated Learning, in particular, minimizes privacy risks by enabling model training across distributed data sources without the need to transmit raw data [64]. This approach establishes a reliable foundation for generating synthetic data or analyzing sensitive data without centralized data sharing. In this context, FedGAN (Federated GAN) and Google's Federated Analytics stand out as notable examples of the latest advancements in privacy-preserving technologies.

Currently, there is no internationally standardized institutional framework for evaluating synthetic data. The use and evaluation of synthetic data is a relatively new field, and efforts toward standardization are still in their early stages. As the field of synthetic data evolves, the need for institutional standards is increasing [65].

Future research should not only explore additional privacy protection techniques and algorithm optimization methods to further reduce the exposure risk of sensitive information in synthetic data generation algorithms but also maximize utility. Such research will make significant contributions to maintaining the balance between privacy protection and data utility and will play a crucial role in the safe utilization of synthetic data.

VI. CONCLUSION

This study evaluated the safety of synthetic data generated using various algorithms (Synthpop, CTGAN, RTVAE, TVAE, and DataSynthesizer). Synthetic data is designed to protect sensitive information while maintaining similar statistical properties; however, a high similarity to the original data can increase the risk of re-identification. To assess this risk, the study comprehensively analyzed the performance of each algorithm using various privacy protection metrics (DCR, NNDR, Identification Risk Indicator, Inference Risk Indicator, CM3, DUPI).

The results showed that the Synthpop algorithm exhibited low DCR and NNDR values across most metrics, indicating that the synthetic data is highly similar to the original data, posing a high privacy risk. The algorithm also recorded

high values in the Identification Risk and Inference Risk Indicators, suggesting a relatively high re-identification risk and underscoring the need for caution when using Synthpop to protect sensitive data. Thus, while Synthpop offers high utility, it should be used cautiously from a privacy protection perspective.

The CTGAN algorithm demonstrated relatively superior privacy protection performance, recording low values in the Identification Risk and Inference Risk Indicators. However, its DCR and NNDR values remained at moderate levels, indicating some similarity to the original data without being excessive, thus balancing privacy and utility well. This suggests that CTGAN is suitable for situations where a balance between privacy protection and data utility is required.

The DataSynthesizer algorithm demonstrated superior privacy protection performance across all evaluated metrics. It consistently recorded the highest values in the NNDR metric, indicating a clear distinction between synthetic and original data, and the lowest values in the Identification Risk and Inference Risk metrics, showcasing its ability to minimize re-identification risks. Furthermore, its stable performance in the CM3 metric highlighted its strong protection against attribute exposure. These results indicate that DataSynthesizer is particularly suitable for applications requiring stringent privacy guarantees, such as sensitive datasets in the medical or financial sectors. Despite its emphasis on privacy, DataSynthesizer also maintains an acceptable level of data utility, making it a well-balanced choice for secure data utilization.

The RTVAE and TVAE algorithms showed high DCR and NNDR values, indicating that the synthetic data maintained a substantial distinction from the original data. This is positive from a privacy protection perspective but it may reduce data utility. Notably, these algorithms recorded relatively high values in the CM3 metric, confirming a low risk of attribute exposure. However, some datasets showed higher Inference Risk Indicator values, suggesting caution in certain applications.

This study emphasizes that each synthetic data generation algorithm may exhibit different privacy protection performance depending on the characteristics of the dataset, analyzing the trade-off between privacy protection and utility. While Synthpop provides high utility at the expense of privacy, algorithms like CTGAN, RTVAE, TVAE, and DataSynthesizer demonstrate stronger privacy protection capabilities. These findings provide valuable insights into selecting appropriate synthetic data generation algorithms based on varying data protection requirements. Researchers often face a trade-off between data utility and privacy protection when selecting a synthetic data generation algorithm. This study quantitatively evaluates the strengths and weaknesses of each algorithm, aiding researchers in choosing the most suitable algorithm for specific contexts. For instance, in studies involving sensitive data, algorithms like CTGAN, RTVAE, TVAE or DataSynthesizer may be advantageous as they effectively reduce re-identification risks in areas where data protection is crucial.

As the data economy grows, the ability to safely share and utilize data becomes a key driver of innovation across society. By evaluating the safety and utility of synthetic data, this study can contribute to enhancing data accessibility needed to address various societal challenges. For example, the safe use of medical data can accelerate disease research and new drug development, while data-driven decision-making in innovative public projects such as smart cities can be conducted more securely. Future research directions include integrating Differential Privacy techniques with synthetic data generation methods to provide additional layers of privacy protection. Such combinations can potentially mitigate the re-identification risks associated with high similarity to original datasets while maintaining data utility. Additionally, expanding the research scope to encompass large-scale empirical datasets with complex privacy risks, such as those found in the medical and financial sectors, would allow for the validation of these synthetic data generation algorithms in more realistic and demanding contexts. These advancements would further enhance the applicability and reliability of synthetic data in sensitive and high-stakes environments.

In conclusion, this study systematically evaluates the privacy protection performance of synthetic data generation algorithms, providing crucial guidance for all stakeholders seeking to enhance the safety of data utilization. Ultimately, this can improve the reliability of data use and contribute to better social and economic outcomes.

REFERENCES

[1] F. Emmert-Streib, "From the digital data revolution toward a digital society: Pervasiveness of artificial intelligence," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 1, pp. 284–298, Mar. 2021.

[2] B. Lubarsky, "Re-identification of 'anonymized data,'" *Georgetown Law Technol. Rev.*, vol. 1, p. 202, May 2017.

[3] Korea Data Agency. (2023). *2023 Data Industry White Paper*. [Online]. Available: https://www.kdata.or.kr

[4] M. Guillaudeux, O. Rousseau, J. Petot, Z. Bennis, C.-A. Dein, T. Goronflot, N. Vince, S. Limou, M. Karakachoff, M. Wargny, and P.-A. Gourraud, "Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis," *npj Digit. Med.*, vol. 6, no. 1, p. 37, Mar. 2023.

[5] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, no. 1, p. 1376, Mar. 2013.

[6] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2008, pp. 111–125.

[7] F. K. Dankar, K. E. Emam, A. Neisa, and T. Roffey, "Estimating the re-identification risk of clinical data sets," *BMC Med. Informat. Decis. Making*, vol. 12, no. 1, p. 66, Dec. 2012. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-12-66

[8] T. S. Mayer, "Privacy and confidentiality research and the U.S. census bureau: Recommendations based on a review of the literature," U.S. Census Bur., Stat. Res. Division, Washington, DC, USA, Tech. Rep. 2002-01, 2002.

[9] E. Singer, J. Van Hoewyk, and R. J. Neugebauer, "Attitudes and behavior: The impact of privacy and confidentiality concerns on participation in the 2000 census," *Public Opinion Quart.*, vol. 67, no. 3, pp. 368–384, 2003.

[10] J. Jordon, D. Jarrett, E. Saveliev, J. Yoon, P. Elbers, P. Thoral, A. Ercole, C. Zhang, D. Belgrave, and M. van der Schaar, "Hide-and-Seek privacy challenge: Synthetic data generation vs. patient re-identification," in *Proc. NeurIPS Competition Demonstration Track*, vol. 133, Aug. 2021, pp. 206–215.
[11] E. De Cristofaro, "Synthetic data: Methods, use cases, and risks," 2023, *arXiv:2303.01230*.
[12] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller, "Synthetic data—What, why and how?" 2022, *arXiv:2205.03257*.
[13] Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu, and W. Wei, "Machine learning for synthetic data generation: A review," *J. LaTeX Class Files*, vol. 14, no. 8, pp. 1–20, 2021.
[14] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnádi, "A unified framework for quantifying privacy risk in synthetic data," *Proc. Privacy Enhancing Technol. (PoPETs)*, vol. 2023, no. 2, pp. 312–328, 2023, doi: 10.56553/popets-2023-0055.
[15] J. Hu and C. M. Bowen, "Advancing microdata privacy protection: A review of synthetic data methods," in *Wiley Interdisciplinary Reviews: Computational Statistics (WICS)*, Aug. 2023.
[16] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 11s, pp. 1–37, Sep. 2022, doi: 10.1145/3523273.
[17] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Francisco, CA, USA, May 2019, pp. 691–706.
[18] T. Xu, C. Liu, K. Zhang, and J. Zhang, "Membership inference attacks against medical databases," in *Proc. Int. Conf. Neural Inf. Process.*, in Communications in Computer and Information Science (CCIS), vol. 1963. Cham, Switzerland: Springer, Nov. 2023, pp. 15–25.
[19] S. S. Sengar, A. B. Hasan, S. Kumar, and F. Carroll, "Generative artificial intelligence: A systematic review and applications," *Multimedia Tools Appl.*, vol. 10, pp. 1–40, Aug. 2024.
[20] X. Guo and Y. Chen, "Generative AI for synthetic data generation: Methods, challenges and the future," Mar. 2024, *arXiv:2403.04190*.
[21] O. A. Fdal. (Jul. 27, 2021). *How to Manage Re-Identification Risks With Synthetic Data*. Statice Blog. [Online]. Available: https://www.statice.ai/post/how-manage-reidentification-risks-personal-data-synthetic-data
[22] H. Teng, H. Lu, M. Ye, K. Yan, Z. Gao, and Q. Jin, "Applying of adaptive threshold non-maximum suppression to pneumonia detection," in *Bio-Inspired Computing: Theories and Applications (BIC-TA)* (Communications in Computer and Information Science (CCIS)), vol. 1160. Cham, Switzerland: Springer, 2020, pp. 518–528.
[23] F. K. Dankar and M. K. Ibrahim, "A new PCA-based utility measure for synthetic data evaluation," Nov. 2022, *arXiv:2212.05595*.
[24] B. Lubarsky, "Re-identification of 'anonymized' data," *Georgetown Law Technol. Rev.*, vol. 1, p. 202, Apr. 2017.
[25] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA Law Rev.*, vol. 57, p. 1701, Aug. 2010.
[26] C. Culnane, B. Rubinstein, and V. Teague, "Health data in an open world: A report on re-identifying patients in the MBS/PBS dataset and the implications for future releases of Australian government data," School of Computing and Information Systems, Univ. Melbourne, Melbourne, VIC, Australia, Tech. Rep., Dec. 2017.
[27] Y. Jiang, L. Mosquera, B. Jiang, L. Kong, and K. E. Emam, "Measuring re-identification risk using a synthetic estimator to enable data sharing," *PLoS ONE*, vol. 17, no. 6, Jun. 2022, Art. no. e0269097.
[28] Y. Gong, Z. Zeng, L. Chen, Y. Luo, B. Weng, and F. Ye, "A person re-identification data augmentation method with adversarial defense effect," Jan. 2021, *arXiv:2101.08783*.
[29] Insighture Technol. Team. (Jan. 17, 2024). *Differential Privacy: Balancing Data Utility and User Privacy in Machine Learning*. [Online]. Available: https://url.kr/346z6d and https://medium.com/insights-by-insighture/differential-privacy-balancing-data-utility-and-user-privacy-in-machine-learning-2282e51be9bf
[30] C. Kurz, "Understanding differential privacy," *Significance*, vol. 18, no. 3, pp. 24–27, Jun. 2021.
[31] Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, and W. Wei, "Machine learning for synthetic data generation: A review," Jun. 2024, *arXiv:2302.04062*.
[32] K. J. Lee and J. B. Carlin, "Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation," *Amer. J. Epidemiol.*, vol. 171, no. 5, pp. 624–632, Mar. 2010.
[33] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–12, 2011.
[34] B. Nowok, G. M. Raab, and C. Dibben, "Synthpop: Bespoke creation of synthetic data in R," *J. Stat. Softw.*, vol. 74, no. 11, pp. 1–26, 2016.
[35] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2019, pp. 7335–7345, Paper 659.
[36] L. V. H. Vardhan and S. Kok, "Synthetic tabular data generation with oblivious variational autoencoders: Alleviating the paucity of personal tabular data for open research," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1–11.
[37] V. Shulakov, "High-quality tabular data generation using post-selected VAE," Dept. Computer Science and System Analysis, Cherkasy State Technological Univ., Cherkasy, Ukraine, Tech. Rep., Jul. 2024, doi: 10.48550/arXiv.2407.13016.
[38] H. Akrami, S. Aydore, R. M. Leahy, and A. A. Joshi, "Robust variational autoencoder for tabular data with beta divergence," 2020, *arXiv:2006.08204*.
[39] H. Ping, J. Stoyanovich, and B. Howe, "DataSynthesizer: Privacy-preserving synthetic datasets," in *Proc. 29th Int. Conf. Sci. Stat. Database Manage.*, Jun. 2017, pp. 309–1117.
[40] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC Med. Res. Methodol.*, vol. 20, no. 1, p. 108, Dec. 2020.
[41] J. Snoke, G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic, "General and specific utility measures for synthetic data," *J. Roy. Stat. Soc. A, Statist. Soc.*, vol. 181, no. 3, pp. 663–688, Jun. 2018.
[42] C. M. Bowen and J. Snoke, "Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge," *J. Privacy Confidentiality*, vol. 11, no. 1, pp. 1–26, Feb. 2021.
[43] I. H. Rather and S. Kumar, "Generative adversarial network based synthetic data training model for lightweight convolutional neural networks," *Multimedia Tools Appl.*, vol. 83, no. 2, pp. 6249–6271, Jan. 2024.
[44] T. Shen, G. Zhao, and S. You, "A study on improving realism of synthetic data for machine learning," *Proc. SPIE*, vol. 12529, pp. 270–277, Mar. 2024.
[45] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnadi, "A unified framework for quantifying privacy risk in synthetic data," *Proc. Privacy Enhancing Technol. (PoPETs)*, vol. 2023, no. 2, pp. 312–328, 2023, doi: 10.56553/popets-2023-0055.
[46] M. Platzer and T. Reutterer, "Holdout-based empirical assessment of mixed-type synthetic data," *Frontiers Big Data*, vol. 4, pp. 1–15, Jun. 2021.
[47] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," *SpringerPlus*, vol. 4, no. 1, p. 694, Dec. 2015.
[48] R. A. Fisher, *Statistical Methods for Research Workers*. Edinburgh, U.K.: Oliver & Boyd, 1925.
[49] I. Riiser, "Privacy and utility evaluation of synthetic data for multi-state time-to-event applications," Master's thesis, Dept. Mathematics, Univ. Oslo, Oslo, Norway, 2023.
[50] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," 2018, *arXiv:1806.03384*.
[51] Q. Liu, M. Khalil, J. Jovanovic, and R. Shakya, "Scaling while privacy preserving: A comprehensive synthetic tabular data generation and evaluation in learning analytics," in *Proc. 14th Learn. Anal. Knowl. Conf.*, Mar. 2024, pp. 620–631.
[52] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[53] G. M. Raab, B. Nowok, and C. Dibben, "Assessing, visualizing and improving the utility of synthetic data," 2021, *arXiv:2109.12717*.
[54] B. Nowok, G. M. Raab, and C. Dibben, "Providing bespoke synthetic data for the U.K. longitudinal studies and other sensitive data with the synthpop package for R1," *Stat. J. IAOS*, vol. 33, no. 3, pp. 785–796, Aug. 2017.

[55] Korea Insurance Res. Inst. (2023). *Measuring the Utility and Disclosure Risk of Synthetic Data*. [Online]. Available: https://www.kiri.or.kr/pdf/%EC%97%B0%EA%B5%AC%EC%9E%90%EB%A3%8C/%EC%97%B0%EA%B5%AC%EB%B3%B4%EA%B3%A0%EC%84%9C/nre2023-07_4.pdf

[56] S. E. Kababji, N. Mitsakakis, X. Fang, A.-A. Beltran-Bless, G. Pond, L. Vandermeer, D. Radhakrishnan, L. Mosquera, A. Paterson, L. Shepherd, B. Chen, W. E. Barlow, J. Gralow, M.-F. Savard, M. Clemons, and K. E. Emam, "Evaluating the utility and privacy of synthetic breast cancer clinical trial data sets," *JCO Clin. Cancer Informat.*, vol. 7, pp. 1–11, Sep. 2023.

[57] J. Snoke, G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic, "General and specific utility measures for synthetic data," *J. Royal Stat. Soc., Ser. A (Statist. Soc.)*, vol. 181, no. 3, pp. 663–688, Jun. 2018, doi: 10.1111/rssa.12358.

[58] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," Oct. 2019, *arXiv:1907.00503*.

[59] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "CTGAN: Synthetic tabular data using generative adversarial networks," in *Proc. 33rd Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 1–12.

[60] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete electronic health records using generative adversarial networks," in *Proc. 2nd Mach. Learn. Healthcare Conf.*, vol. 68, Mar. 2017, pp. 286–305.

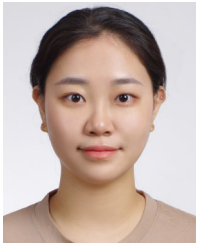
[61] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, and C. S. Greene, "Privacy-preserving generative deep neural networks support clinical data sharing," *Circulat., Cardiovascular Quality Outcomes*, vol. 12, no. 7, pp. 1–11, Jul. 2019.

[62] Y. Xiao, H. Xiong, and L. Wang, "Privacy-preserving data publishing: A review of techniques and applications," *ACM Comput. Surveys (CSUR)*, vol. 42, no. 4, pp. 1–53, Jun. 2010, doi: 10.1145/1749603.1749605.

[63] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," *PLOS One*, Dec. 2011, doi: 10.1371/journal.pone.0028071.

[64] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Netw.*, vol. 33, no. 5, pp. 156–165, Sep. 2019.

[65] S. E. Kim, "A study on the disclosure risk assessment in synthetic data," Master's thesis, Chonnam Nat. Univ., Gwangju, South Korea, p. 62, 2024.



GIDAN MIN is currently pursuing the B.S. degree in information security (primary major) and personal information protection (secondary major) with Seoul Women's University. Her research interests include cybersecurity, data protection, and privacy.

JUNHYOUNG OH received the Ph.D. degree from Korea University. He is currently a Tenure-Track Assistant Professor with the Division of Information Security, Seoul Women's University, where he has been leading the AI Privacy Protection Laboratory, since 2024.