Vol. 08, No. 3 (2026) 1797-1806, doi: 10.24874/PES08.03.030

Proceedings on Engineering Sciences

www.pesjournal.net

DESIGN AND IMPLEMENTATION OF A DATA ANONYMIZATION ALGORITHM FOR PRIVACY PRESERVATION WITH ENHANCED DATA USABILITY

Dhananjay M. Kanade[1]
Snehal M. Kamalapur
Shirish S. Sane

Received 06.05.2026.
Revised 29.06.2026.
Accepted 03.08.2026.

Keywords:

Data anonymization, Synthetic data, Privacy preservation, Machine learning, GAN, CopulaGAN, Decision tree, Data usability

A B S T R A C T

The rapid growth of digital data in healthcare, transportation, finance, and social sciences has increased concerns about re-identification and unauthorized inference. This study designs and evaluates a synthetic-data framework using decision-tree, GAN, and CopulaGAN generators on the Adult, ATUS, and FARS datasets. Evaluation focuses on classification utility and structural fidelity. Decision-tree synthetic data achieve the highest synthetic accuracy on Adult and ATUS, whereas CopulaGAN achieves the highest synthetic accuracy on FARS. Privacy risks, including re-identification, disclosure, membership inference, attribute inference, and record linkage, were not quantitatively assessed. Therefore, the findings support utility retention but do not demonstrate privacy improvement or an optimal privacy-utility trade-off. © 202x Published by Faculty of Engineering

© 2026 Published by Faculty of Engineering

1. INTRODUCTION

The growing use of data-driven decision-making has resulted in a fast growth of massive datasets containing sensitive personal data. Although these attributes are useful, they may cause revealing sensitive attributes leading to identity disclosure, attribute disclosure and extreme privacy breaches. The regime of the law and moral of privacy intrusions has been enhanced in the wake of global privacy laws and regulations like GDPR, CCPA, and HIPAA. AjmeeraKiran and Shirisha (2022) therefore argue that privacy-sensitive data publishing has become an essential requirement of the contemporary data ecosystem. Conventional anonymization tools, including k-Anonymity, l-Diversity, and t-Closeness, can provide a basic level of protection to the data but can commonly ruin the analysis value of anonymized data. Specifically, data of high dimensions and datasets with mixed attributes of data are likely to lose information following anonymization transformations (Alzahrani et al., 2022).

New promising approaches based on machine learning and deep generative models are recent developments in synthetic data generation. Strong conditional relations between attributes are used in decision-tree-based methods. Copula models provide a mathematically rigorous way to model inter-feature dependencies, whereas Generative Adversarial Networks (GANs) learn complex joint distributions (Xu et al., 2019). Generative

Corresponding author: Dhananjay M. Kanade
Email: dmkanade@kkwagh.edu.in

methods, however, can experience mode collapse, loss of dependency structure, sampling instability, and limited support for categorical attributes. Developing an anonymization methodology that retains analytical utility while reducing direct exposure therefore remains an open research problem (Desai et al., 2022; Mahato et al., 2024).

2. RELATED WORK AND THEORETICAL BACKGROUND

Initial data-privacy approaches used statistical disclosure-control methods such as generalization, suppression, and perturbation, but these methods do not fully prevent homogeneity and background-knowledge attacks. l-Diversity extends k-Anonymity by requiring diversity in sensitive attributes, but it can be difficult to apply to high-dimensional data. t-Closeness seeks to preserve sensitive-attribute distributions but may cause substantial information loss. Noise addition, differential privacy, and randomization-based approaches can offer stronger privacy controls, although added noise may reduce model utility (Ponomareva et al., 2023). Broader institutional and practical barriers to responsible data sharing have also been documented (Govarts et al., 2022; Lichtenauer et al., 2024, 2025). The magnitude and allocation of noise can affect machine-learning performance and may have unequal effects across subgroups (Hamsanandhini & Balasubramanie, 2024).

The conditional relationships among variables allow machine-learning models to learn complex anonymization patterns (Kamikubo et al., 2023). Decision-tree anonymization captures feature interactions hierarchically and can therefore generate synthetic data with high structural fidelity. Generative Adversarial Networks (GANs) use adversarial training between a generator and a discriminator to learn complex multivariate distributions. However, conventional GANs can be unstable when applied to tabular data containing mixed continuous and categorical attributes, sparse categories, and imbalanced discrete values (Xu et al., 2019). They may also memorize training records in privacy-sensitive settings unless appropriate safeguards are applied. CopulaGAN combines copula transformations with adversarial learning to retain dependencies and improve synthetic-data quality in high-dimensional mixed-type datasets. Table 1 summarizes the core principles, strengths, limitations, and suitable data types of the privacy-preserving and synthetic-data techniques reviewed in this section.

Table 1. Summary of Related Work on Privacy-Preserving and Synthetic Data Generation Models

Technique / Model	Core Principle	Strengths	Limitations	Suitable Data Types
k-Anonymity	Generalization & suppression	Simple, widely adopted	Prone to homogeneity & background attacks	Tabular mixed-type
l-Diversity	Ensures diversity in sensitive attributes	Improves over k-Anonymity	Fails with semantically similar values	Tabular mixed-type
t-Closeness	Preserves distribution of sensitive attribute	Stronger privacy guarantee	High information loss in high dimensions	Tabular mixed-type
Differential Privacy	Noise injection with ε-privacy budget	Strong theoretical guarantee	Distorts data, harms ML utility	Numeric, structured
Noise Addition / Randomization	Adds random perturbations	Simple to implement	High utility distortion	Numeric
Decision-Tree Synthetic Generator	Conditional modeling via tree splits	High structural fidelity, interpretable	Overfitting risk, mimics leaf values	Mixed-type tabular
Ensemble-Based Synthetic Models	Multiple models capture diverse patterns	More robust than single tree	Requires complex tuning	Mixed-type
Bayesian Networks	Probabilistic dependency modeling	Captures conditional relations	Struggles with nonlinear patterns	Tabular
GAN-Based Synthetic Data	Adversarial learning of joint distributions	Excellent modeling of nonlinear patterns	Mode collapse, instability, memorization risk	Numeric, mixed (challenging)
CopulaGAN	Copula-based dependency + GAN training	Strong dependency retention, stable	Requires complex transformations	High-dimensional, mixed-type

3. PROPOSED DATA ANONYMIZATION FRAMEWORK

The architectural basis of anonymization pipeline starts with strict data preprocessing with the help of metadata construction to guarantee reversibility and structural consistency. Lacking values are properly filled with the help of elementary statistical operators (mean, mode):

$$X* = \left(\frac{1}{n}\right) * \sum_{i=1}^{n} X_i$$

Categorical attributes are encoded through one-hot mappings:

$$X_{one} - hot = \{0,1\}^k$$

The z-score scaling is done on continuous variables:

$$Z = \frac{X - \mu}{\sigma}$$

Variables are then converted into percentiles by use of empirical cumulative distribution function (CDF):

$$Pi = F(Xi)$$

The percentiles are then transformed to a Gaussian distribution by use of the inverse Gaussian CDF:

$$Zi = \Phi^{-1(Pi)}$$

Through this combination, consistent scaling, stable model training, and accurate dependency capture are promoted. Metadata also retain critical mappings, including domain ranges, categorical indices, and distribution types, to support dependable inverse transformation of synthetic samples into real-world semantic values (Majeed & Lee, 2020; Majeed & Hwang, 2023a).

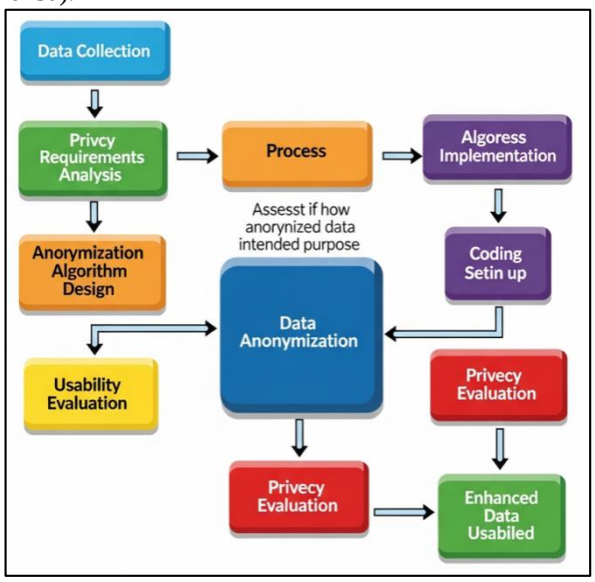


Figure 1. End-to-End Workflow Model for Data Anonymization and Utility Enhancement

Figure 1 presents the ordered workflow from raw-data acquisition to the generation of analytically useful synthetic data. After requirements are specified, an anonymization approach, such as decision-tree synthesis, GAN-based generation, CopulaGAN modelling, or k-Anonymity, is selected and implemented. Usability evaluation then examines whether the transformed data retain classification utility. The privacy-evaluation stages shown in Figure 1 are explicitly labelled as not performed in this study; no attack-based or quantitative privacy-risk assessment was conducted (Majeed & Hwang, 2023a).

The workflow also includes process assessment, algorithm implementation, and coding setup for technical testing of the anonymization pipeline. These checks verify that the model logic is correctly translated into executable code and that the implemented models meet the intended utility objectives (Ponomareva et al., 2023). Classification utility and structural fidelity were evaluated; the privacy-evaluation stage was not performed in this study. Quantitative disclosure-risk and attack-based privacy testing are reserved for future work (Qian et al., 2024).

3.1 Key Innovation Elements

The main novelty of the offered anonymization framework is its synergistic incorporation of Gaussian normalization, copula-related dependency preservation, decision-tree predictive sampling, and a hybrid multi-model generating approach. Gaussian Space Normalization brings all features to standard normal distribution so that all features would be scaled equally with means of zero and unit variance (Roopalakshmi et al., 2026). Dependency retention using copulas also enhances the framework by the explicit modelling of multivariate relationships regardless of whether or not their marginal distributions are the same. This is to ensure that the dependencies between variables, including correlations, joint patterns or conditional structures are preserved when generating synthetic data.

3.2 Privacy Preservation Model

The privacy-preservation objective of the proposed framework is to reduce direct exposure of original records by generating synthetic samples from learned statistical patterns and dependency structures rather than releasing the source records themselves (Majeed & Hwang, 2023b). The threat model considered here includes an adversary attempting identity linkage, attribute inference, or record matching using released synthetic data and auxiliary information. Potentially protected information includes sensitive socioeconomic, behavioral, and safety-related attributes contained in the Adult, ATUS, and FARS datasets. However, the present experiments primarily evaluate data utility and structural fidelity; they do not directly quantify re-identification probability, disclosure risk, membership inference, or privacy leakage. Accordingly, the generated data should not be interpreted as provably non-reconstructible or categorically resistant to privacy attacks. The level of protection depends on the learned model, dataset characteristics, available auxiliary information, and model configuration. A formal privacy-risk assessment and a quantitative privacy-utility trade-off analysis are therefore outside the scope of this study and constitute an important limitation and direction for future work.

Scope and evidential basis of the privacy claims. In this study, the term privacy preservation is used in a limited empirical sense: the proposed generators reduce direct disclosure by replacing released source records with synthetic records learned from the source distributions. The evidence reported in this paper supports utility retention and structural fidelity through classifier accuracy, dependency-preservation measures, and statistical-divergence measures. These outcomes do not constitute a formal or attack-based privacy guarantee. The experiments do not estimate re-identification risk, membership-inference success, attribute-inference success, nearest-neighbor disclosure, record linkage, or a differential-privacy budget (ε). Therefore, statements

describing a method as providing better privacy, stronger protection, or a privacy–utility trade-off should be interpreted as design objectives or indirect indications of reduced direct exposure, not as proof of resistance to adversarial reconstruction or inference. Formal privacy guarantees would require dedicated disclosure-risk and attack-based evaluation and, where applicable, a mechanism with a mathematically defined privacy guarantee such as differential privacy.

4. EXPERIMENTAL SETUP AND DATASET DESCRIPTION

4.1 Dataset Descriptions

This paper measures the anonymization algorithm on three benchmark datasets that are publicly available and include Adult, ATUS, and FARS. The datasets were all chosen because of their variance in size, dimensionality, and type of attributes.

A. Adult Dataset

The Adult dataset description uses 48,211 source records with 15 mixed categorical and numerical features. The processed experimental version reported in Table 2 contains 48,842 records, reflecting the author-confirmed value used for classifier evaluation. The dataset supports income classification and includes sensitive socioeconomic attributes such as age, education, occupation, marital status, and working hours.

B. ATUS Dataset

The American Time Use Survey (ATUS) source dataset contains 170,883 records and 132 features, providing a high-dimensional benchmark for behavioural-data modelling. The processed experimental version reported in Table 3 contains 170,843 records, as confirmed by the authors. The dataset is numeric and sparse and describes time allocation across work, leisure, household duties, and social activities.

C. FARS Dataset

The Fatality Analysis Reporting System (FARS) source dataset contains 39,200 records and 80 mixed-type features describing accident conditions, vehicle characteristics, environmental factors, and injury outcomes. The processed experimental version reported in Table 4 contains 39,221 records, as confirmed by the authors. Its classification task requires preservation of meaningful interactions between categorical and numerical attributes.

4.2 Anonymization Models Evaluated

1) Decision Tree–Based Synthetic Generator

The synthetic generator that is built using the Decision Tree is used to build predictive models of each feature by training individual trees that conditional relationships within the data as shown in step wise algorithm 1. In the synthetic creation of data the model draws samples of the leaf-level distributions, so that the data created by the model is sensitive to the local structure and dependency patterns of the original data.

Algorithm 1: Decision-Tree-Based Synthetic Data Generation

Input: Original dataset D with features {X1, X2, …, Xn}
Output: Synthetic dataset S
1: for each feature Xi in D do
2: Train a decision tree Ti using all other features $\{Xj \mid j \neq i\}$
3: Store leaf-level value distributions for Xi
4: end for
5: Initialize empty synthetic dataset S
6: for each synthetic sample to generate do
7: Select a random base record r from D
8: for each feature Xi do
9: Use current partial synthetic record to traverse Ti
10: Sample Xi' from leaf-node distribution
11: Insert Xi' into synthetic sample
12: end for
13: Add completed synthetic sample to S
14: end for
15: return S

2) GAN Synthetic Generator

The GAN synthetic generator, as shown in algorithm 2, is based on the Generative Adversarial Network, where the goal of a generator is to generate the synthetic samples and a discriminator is to differentiate between actual and generated data.

Algorithm 2: GAN-Based Synthetic Data Generation

Input: Preprocessed dataset D, Generator G, Discriminator D
Output: Synthetic dataset S
1: Normalize numerical features and encode categorical attributes
2: Initialize generator G and discriminator D with random weights
3: repeat
4: Sample minibatch of real data xr from D
5 Generate synthetic batch $xs = G(z), \text{where } z \sim N(0, I)$
6: Update discriminator:
7: Maximize $LD = \log(D(xr)) + \log(1 - D(xs))$
8: Update generator:
9: Minimize $LG = \log(1 - D(G(z)))$
10: until training converges or loss stabilizes
11: Generate synthetic dataset S = G(z) for new random z
12: Post-process S to recover original feature formats
13: return S

3) CopulaGAN

CopulaGAN combines dependency modeling with GAN-based generative learning, incorporating Gaussian copulas in it, which makes it a hybrid anonymization technique that is especially effective on high-dimensional

data. The synthetic data working step wise process is shown in algorithm 3. It begins by transforming all of the features to percentile values with empirical CDFs then interpolating them into the Gaussian space with inverse CDF transformations.

Algorithm 3: CopulaGAN-Based Synthetic Data Generation
Input: Original dataset D with features {X1, X2, …, Xn}
Output: Synthetic dataset S
1: Convert each feature Xi to percentile values:
2: Pi = CDF(Xi)
3: Transform percentiles to Gaussian space:
4: $Zi = \Phi^{-1(Pi)}$
5: Estimate Gaussian copula correlation matrix Σ from Z
6: Train a GAN on Gaussian-transformed data Z
7: repeat
8: Train discriminator to maximize:
9: $LD = \log(D(Z)) + \log(1 - D(G(z)))$
10: Train generator to minimize:
11: $LG = \log(1 - D(G(z)))$
12: until GAN reaches convergence
13: Generate synthetic Gaussian samples Z' = G(z)
14: Convert Gaussian samples back to percentiles:
15: P'i = Φ(Z'i)
16: Recover synthetic feature values via inverse CDF:
17: $Xi' = CDF^{-1(P'i)}$
18: return S = {X1', X2', …, Xn'}

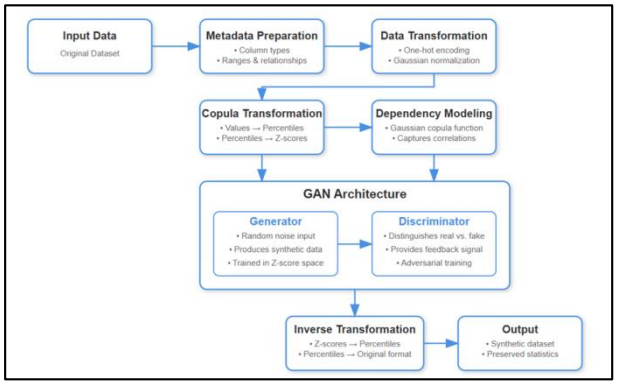


Figure 2: End-to-End CopulaGAN-Based Synthetic Data Generation Workflow

This is depicted in figure 2 below, which shows the entire CopulaGAN data generation process starting with metadata preparation and transformation into percentile space and Gaussian space.

4.3 Classification Models Used for Utility Evaluation

a) A. J48 (C4.5 Decision Tree)

The J48 algorithm is a version of the C4.5 decision tree; as shown in algorithm 4, it builds classification models by recursively splitting data into homogenous subsets using information gain or gain ratio. It works on both categorical and continuous attributes by choosing the best split points which will maximize the purity of the classes. J48 also uses pruning techniques in order to eliminate branches that are statistically insignificant to enhance generalization and lower overfitting.

Algorithm 4: J48 (C4.5 Decision Tree)
1. Let dataset $D = \{(xi, yi)\}$ for $i = 1..n$.
2. Compute entropy:
   $H(D) = -\Sigma\, p(c) * \log2(p(c))$
3. For each feature Aj, compute conditional entropy:
   $H(D|Aj) = \Sigma\, (|Dv|/|D|) * H(Dv)$
4. Compute information gain:
   $IG(Aj) = H(D) - H(D|Aj)$
5. Select best feature:
   $A* = \text{argmax}\, IG(Aj)$
6. Split data using A ∗.
7. Recursively repeat steps until pure leaves.
8. Apply pruning to remove weak branches.

b) Bagging (Bootstrap Aggregation)

Bootstrap Aggregation (also known as bagging), is an ensemble method that improves the stability and accuracy of the models by making them aggregate the predictions of a number of independently trained base models as shown in algorithm 5. It does this by creating a large number of training subsets via random sampling with replacement such that each model is being learned slightly different patterns.

Algorithm 5: Bagging Classifier
Input: Dataset D, Number of models B
Output: Ensemble H
1: for $b = 1$ to B do
2: $Db = \text{BootstrapSample}(D)$
3: $hb = \text{TrainBaseLearner}(Db)$
4: Store hb
5: end for
6: function PREDICT(x)
7: for $b = 1$ to B do
8: $yb = hb(x)$
9: end for
10: return MajorityVote({yb})

c) Voting Ensemble

Voting Ensemble, as shown in algorithm 6, predictor uses a combination of predictions of various base models to form a stronger and more stable final prediction. It has two modes: hard voting where majority class is picked, and soft voting whereby the probability of a class in various models is averaged. The voting approach uses the strengths of a variety of learners, including decision trees, logistic models, and SVMs, to minimize the bias and variance of each of the models.

Algorithm 6: Voting Ensemble
Input: Classifiers {h1, h2, ..., hk}
Output: Ensemble classifier H
1: function PREDICT(x)

```
2:    if HARD then
3:        yi = hi(x) for all i
4:        return MajorityVote(yi)
5:    else if SOFT then
6:        for each class c do
7:            score(c) = Σ pi(c|x)
8:        end for
9:        return argmax(score(c))
```

d) AdaBoost (Adaptive Boosting)

AdaBoost is an iterative ensemble algorithm, which enhances the classification accuracy by training weak learners, as shown in algorithm 7, often decision stumps, sequentially and changing instance weights after every round. There is an increased weight after the misclassification of the samples, which compels the later learners to deal with challenging instances.

Algorithm 7: AdaBoost

Input: Dataset D, Rounds T
Output: Strong classifier H
1: Initialize wi = 1/n
2: for t = 1 to T do
3: Train ht using weights w
4: Compute error εt
5: $\alpha t = \left(\frac{1}{2}\right) * \ln\left(\frac{1-\varepsilon t}{\varepsilon t}\right)$
6: Update:
7: $wi = wi * \exp(-\alpha t * yi * ht(xi))$
8: Normalize w
9: end for
10: function PREDICT(x)
11: return sign(Σt αt ht(x))

e) Random Forest

The training of each tree is done on a bootstrap sample, and when splitting, a random number of features are only considered. The random forest working details shown in algorithm 8. This kind of diversification greatly lowers the overfitting and maximizes the generalization. To classify the final prediction is made by majority voting between all trees.

Algorithm 8: Random Forest

Input: Dataset D, Trees B, Feature subset m
Output: Ensemble classifier H
1: for b = 1 to B do
2: Db = BootstrapSample(D)
3: Tb = GrowTree(Db, m)
4: Store Tb
5: end for
6: function PREDICT(x)
7: for b = 1 to B do
8: yb = Tb(x)
9: end for
10: return MajorityVote({yb})

4.4 Experimental Environment

This research used an experimental environment to accomplish its work based on a combination of Python and R toolchains to provide a flexible, robust, and reproducible analytics workflow. Python was a main platform of executing machine learning models, synthetic data generators, preprocessing pipelines since it has a rich ecosystem, Scikit-learn with classification tasks, and the Synthetic Data Vault (SDV) framework to produce Decision Tree, GAN, and CopulaGAN synthetic datasets. Besides, WEKA also offered a standardized setting in which model performance is assessed, so that the behavior of the classifier was consistent across original and anonymized datasets. The metrics of evaluation were measurement of classification accuracy, dependency preservation scores and statistical divergence scores, and they all measured both predictive utility and structural fidelity. This cross-language multi-tool improved fairness, reproducibility, and methodological trustworthiness of all the experimental trials, which is consistent with the best practices in privacy-preserving data analysis.

5. RESULT AND DISCUSSION

The comparison of accuracy of Adult dataset shows the obvious differences in the performance of anonymization techniques and classifiers. This is expected because the original dataset has the highest total accuracy with Bagging (95.01) and J48 (86.19) taking the first places since the two have the capability of representing the underlying demographic trends. Traditional methods of anonymization, such as T-closeness are typically more effective than k-Anonymity and l-Diversity and can have accuracy of up to 81.30% when using Random Forest. This is an indication of the high capability of T-closeness to conserve the attribute distributions whilst circumscribing sensitive value disclosure. Nonetheless, none of the three classical methods exhibit significantly different utility drop relative to the original dataset, which proves that generalization and suppression tend to project misleading predictive characteristics. Table 2 presents the classifier-wise accuracy values for the original Adult dataset and each anonymized or synthetic data version.

The synthetic data generated by the GAN also yields lower accuracy with the values being between 72.53 and 75.92, as shown in figure 3. CopulaGAN is a little higher with structural relationships by using the copula-based dependency modeling with accuracy ranging between 68.43 to 77.76. Despite being an advancement to the traditional GANs, it is still dataset-specific. In general, synthetically generated methods based on Decision Trees (seen in previous sections) are better than the GAN-based ones, and the classic methods of anonymization are in the middle.

Table 2. Accuracy Comparison for Adult Datasets using different classifier

		Classifier				
		J48	Bagging	Voting	Ada Boost	RF
Data Version	Records	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
Original	48,842	86.19	95.01	71.91	84.03	84.11
K_Annony	48,842	76.34	78.3	71.23	74.21	80.1
L Diversity	48,842	76.1	79.22	70.23	78.11	74.9
T closeness	48,842	79.34	80.78	68.44	79.33	81.30
GAN	48,842	75.91	72.53	70.4	75.92	75.92
Copula GAN	48,842	68.43%	75.72	76.32	77.76	76.63
DT	48,842	82.11	89.27	90.97	86.12	90.46

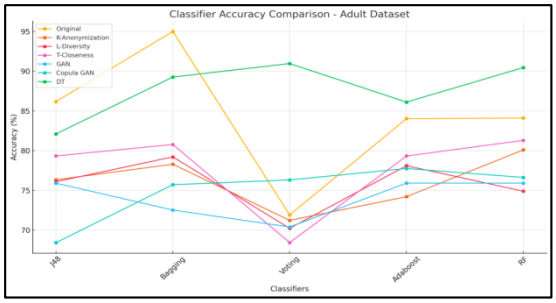


Figure 3. Classifier-wise Accuracy Comparison across Anonymization and Synthetic Data Techniques for the Adult Dataset

The accuracy comparison of the ATUS data finds significant information on the effect of various anonymization and synthetic data methods on the model performance of various classifiers. The original data attains high predictive power with the highest value of 88.6 (J48) and 91.5 (Voting), and 91.4 (Random Forest) indicating the depth of multi-dimensionality of behavioral characteristics in ATUS. The performance is slightly reduced when using conventional anonymization methods, like k-Anonymity and l- Diversity, but is still comparatively high, particularly with Bagging and Random Forest. This implies that the numerical and structured properties of the ATUS dataset are able to be generalized with relative ease without crippling utility.

Table 3. Comparative Accuracy Analysis of Synthetic and Traditional Anonymization Approaches on the ATUS Dataset

		Classifier				
		J48	Bagging	Voting	Ada Boost	RF
Data Version	Records	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
Original	170,843	88.6	90.55	91.5	89.1	91.4
K_Annony	170,843	87.36	88.16	88.64	88.40	90.22
L Diversity	170,843	84.3	89.30	84.5	86.3	89.13
T closeness	170,843	74.76	78.7	83.40	84.30	88.30
GAN	170,843	82.7	85.25	82.5	80.0	82.3
Copula GAN	170,843	88.6	86.98	89.7	87.5	89.6
DT	170,843	89.1	88.65	89.7	88.4	91.6

T-closeness however demonstrates a stronger performance degradation, especially in the case of the J48 and Bagging classifiers which suggests an imposition of distributional similarity constraints can introduce bias to high-dimensional numeric patterns needed in behavioral modeling. The table 3 show the comparative accuracy analysis of synthetic and traditional anonymization approaches on the ATUS dataset. The degradation of GAN-generated data is also evident, and the accuracy drops to the low-80s range, which proves the difficulties in GANs to model the ATUS feature space and dependencies of large magnitude. On the contrary, CopulaGAN is much more successful than traditional GANs and has competitive accuracy, particularly in combination with Voting and Random Forest classifiers. This indicates the benefit of copula-based dependency learning in the multivariate behavioral relationships. The synthetic data generated by a Decision Tree algorithm has the highest overall anonymized performance and in certain classifiers matches or surpasses the accuracy of original data. This highlights that it is powerful in maintaining conditional feature structures that are important in ATUS classification tasks.

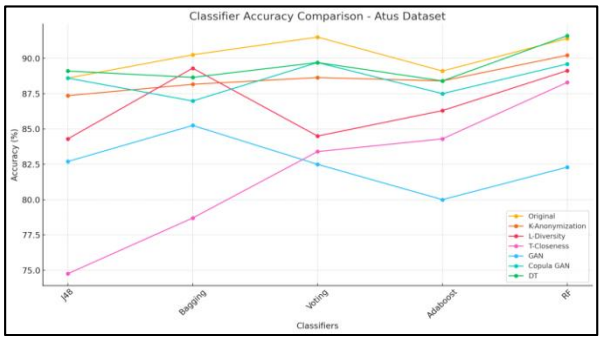


Figure 4. Classifier Accuracy Comparison across Anonymization and Synthetic Data Techniques for the ATUS Dataset

The figure 4 shows the variations in the accuracy of classifiers on the ATUS dataset using various anonymization. The results of Decision Tree and CoparaGAN are always good and they are very close to the original data. Conventional methods, in particular, T-closeness, demonstrate a decrease in utility, whereas GAN is characterized by moderate stability. The dependency-preserving models however have high predictive utility in classifiers overall.

The analysis of the FARS database in terms of accuracy of results, as shown in table 4, indicates the unique performance patterns in anonymization and synthetic data methods, which is connected with the complicated structure of the dataset that presents a combination of mixed categorical and numeric features associated with the severity of the accidents. The initial dataset is characterized by moderate levels of accuracy, with the highest accuracy of 76.14 with J48 all the way to 82.04 with Random Forest, which is in line with the demanding nature of classification of FARS. K-Anonymization has fair results with certain classifiers, particularly with Voting (86.34) and Random Forest (86.43), which means that generalization is not extremely deceptive of fundamental predictors of accidents. Nonetheless, the reduction in the performance of J48 and Bagging indicates that more basic models are more vulnerable to feature coarsening.

Table 4. Comparative Accuracy Analysis of Original and Anonymized FARS Data across Five Classifiers

		Classifier				
		J48	Bagging	Voting	Ada Boost	RF
Data Version	Records	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
Original	39,221	76.14	80.4	77.4	77.43	82.04
K_Annonymization	39,221	71.33	73.1	86.34	84.3	86.43
L Diversity	39,221	69.2	72.88	77.60	76.20	78.33
T closeness	39,221	65.49	75.23	74.21	78.99	80.30
GAN	39,221	83.11	80.03	84.54	83.22	80.34
Copula GAN	39,221	84.23	85.92	86.53	83.81	86.74
DT	39,221	81.23	82.43	82	82.23	83.76

L-Diversity demonstrates additional accuracy loss especially to J48, and AdaBoost, where determining diversity restraints can destabilize vital categorical combinations which dictate crash severity tendencies. T-closeness gives inconsistent results: some higher accuracy of AdaBoost (78.99%) but general poorer results of J48 and Bagging. This shows that limited distributions of sensitive attributes in high-dimensional safety data can substantially modify underlining relationships.

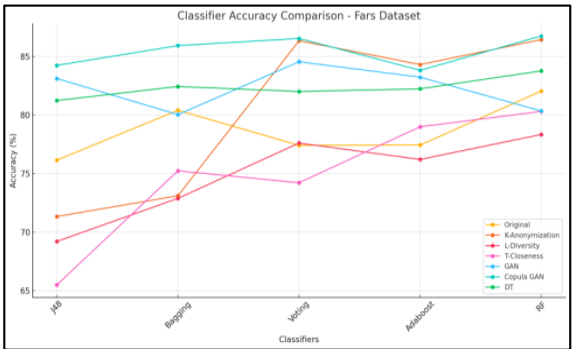


Figure 5. Classifier Accuracy Comparison across Anonymization and Synthetic Data Techniques for the FARS Dataset

Surprisingly, GAN-generated synthetic data is also very useful and the highest accuracy is recorded at 84-83% amongst classifiers, which demonstrates the ability of GANs to model nonlinear relationships in FARS. Nevertheless, the performance is still not stable, as GAN is unstable. CopulaGAN provides the best synthetic performance of up to 86.74% using Random Forest which proves that it is powerful to model multivariate dependencies in safety datasets. Decision Tree synthetic data is a stable and reliable artificial data since its accuracies are above 81, albeit a little lower than CopulaGAN. The figure 5 shows the behavior of accuracy among classifiers of the FARS dataset with the various anonymization strategies. CopulaGAN is highly performing, and the Decision Tree synthetic data is stable in accuracy amongst all classifiers. Conventional anonymization techniques exhibit significant utility loss particularly L-Diversity and T-Closeness. GAN is average but not a strong one. In general, structure-preserving synthetic algorithms provide the highest predictive utility on complicated safety data.

Table 5 reports, for each dataset and data version, the highest accuracy observed among the five classifiers in Tables 2-4. The final row is the arithmetic average of those three dataset-level maxima. The original data achieve the highest overall average accuracy (89.52%), while k-Anonymization is the strongest traditional method on average (85.58%). l-Diversity and t-Closeness obtain average maxima of 82.28% and 83.30%, respectively, indicating reduced classification utility under stronger transformation. These comparisons evaluate utility only; privacy risk was not quantitatively assessed.

Table 5. Comparative Analysis of Classification Accuracy (%) for Original, Anonymized, and Synthetic Data Across Adult, ATUS, and FARS Datasets

Dataset	Original	K anonymization	l-Diversity	T closeness	GAN	Copula GAN	DT
Adult	95.01	80.1	79.22	81.30	75.92	77.76	90.97
Atus	91.5	90.22	89.30	88.30	85.25	89.7	91.6
Fars	82.04	86.43	78.33	80.30	84.54	86.74	83.76
Average	89.52	85.58	82.28	83.3	81.9	84.73	88.78

GAN-generated data achieve an average maximum accuracy of 81.90%, while CopulaGAN reaches 84.73%, illustrating the utility benefit of dependency modelling. Decision-tree synthetic data produce the highest average maximum among the synthetic methods (88.78%), but they do not perform best on every dataset: CopulaGAN records the highest synthetic accuracy on FARS (86.74%), compared with 83.76% for the decision-tree generator. These results demonstrate classification-utility differences among the methods; they do not establish privacy improvement or an optimal privacy-utility trade-off because privacy risks were not quantitatively evaluated.

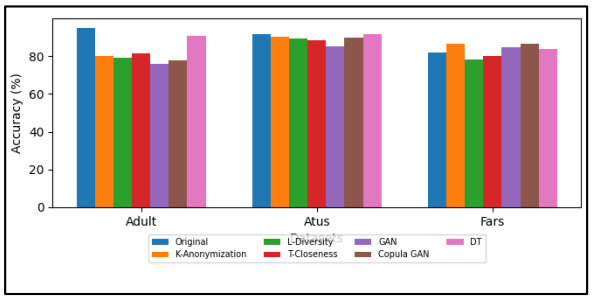


Figure 6. Comparative Accuracy of Anonymization and Synthetic Data Techniques across Adult, ATUS, and FARS Datasets

Figure 6 is aligned with the corrected dataset-level maxima in Table 5. It compares the highest classifier accuracy for each data version within Adult, ATUS, and FARS; the Average row in Table 5 is calculated across the three datasets and is not plotted as a fourth dataset. Decision-tree synthetic data provide the highest synthetic utility on Adult and ATUS, whereas CopulaGAN provides the highest synthetic utility on FARS. The figure therefore supports a dataset-dependent utility interpretation rather than a claim that one method is uniformly best.

6. CONCLUSION

This paper introduced and evaluated a data-anonymization framework designed to retain analytical utility while reducing direct release of original records. The framework combines decision-tree predictive sampling, copula-based dependency modelling, and adversarial generation. Experiments on Adult, ATUS, and FARS show that decision-tree synthetic generation provides strong classification utility on Adult and ATUS, whereas CopulaGAN provides the highest synthetic accuracy on FARS. The evaluation measured classification utility and structural fidelity only. It did not quantify re-identification, disclosure, membership-inference, attribute-inference, or record-linkage risk. Consequently, the results do not demonstrate privacy improvement or an optimal privacy-utility trade-off. Formal privacy-risk and attack-based evaluation is required before stronger privacy claims can be made.

References:

AjmeeraKiran, N., & Shirisha. (2022). K-anonymization approach for privacy preservation using data perturbation techniques in data mining. Materials Today: *Proceedings, 64*(1), 578–584. https://doi.org/10.1016/j.matpr.2022.05.117

Alzahrani, A. G., Alhomoud, A., & Wills, G. (2022). A framework of the critical factors for healthcare providers to share data securely using blockchain. *IEEE Access, 10*, 41064-41077. https://doi.org/10.1109/ACCESS.2022.3162218

Desai, N., Das, M. L., Chaudhari, P., & Kumar, N. (2022). Background knowledge attacks in privacy-preserving data publishing models. *Computers & Security, 122*, 102874. https://doi.org/10.1016/j.cose.2022.102874

Govarts, E., Gilles, L., Bopp, S., Holub, P., Matalonga, L., Vermeulen, R., ... Schoeters, G. (2022). Position paper on management of personal data in environment and health research in Europe. *Environment International, 165*, 107334. https://doi.org/10.1016/j.envint.2022.107334

Hamsanandhini, S., & Balasubramanie, P. (2024). IoT data encryption and phrase search-based efficient processing using a fully homomorphic-based SE (FHSE) scheme. *Pervasive and Mobile Computing, 103*, 101952. https://doi.org/10.1016/j.pmcj.2024.101952.

Kamikubo, R., Lee, K., & Kacorri, H. (2023). Contributing to accessibility datasets: Reflections on sharing study data by blind people. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–18). New York, NY: Association for Computing Machinery. https://doi.org/10.48550/arXiv.2303.04962

Lichtenauer, N., Guggumos, J., Kampmann, M., Kis, J., Laumer, F., März, E., ... Wilhelm, S. (2025). Expert experiences in anonymizing personal data and its use as open data: Qualitative insights. *Data*, *10*(7), 105. doi: 10.3390/data10070105

Lichtenauer, N., Schmidbauer, L., Wilhelm, S., & Wahl, F. (2024). A scoping review on analysis of the barriers and support factors of open data. *Information*, *15*(1), 5. https://doi.org/10.3390/info15010005.

Mahato, G. K., Banerjee, A., Chakraborty, S. K., & Gao, X. Z. (2024). Privacy preserving verifiable federated learning scheme using blockchain and homomorphic encryption. *Applied Soft Computing*, *167*, 112405. https://doi.org/10.1016/j.asoc.2024.112405

Majeed, A., & Lee, S. (2020). Attribute susceptibility and entropy-based data anonymization to improve users' community privacy and utility in publishing data. *Applied Intelligence*, *50*, 2555-2574. https://doi.org/10.1007/s10489-020-01656-w

Majeed, A., & Hwang, S. O. (2023a). A generic approach towards enhancing utility and privacy in person-specific data publishing based on attribute usefulness and uncertainty. *Electronics*, *12*(9), 1978. https://doi.org/10.3390/electronics12091978

Majeed, A., & Hwang, S. O. (2023b). Quantifying the vulnerability of attributes for effective privacy preservation using machine learning. *IEEE Access*, *11*, 4400–4411. https://doi.org/10.1109/ACCESS.2023.3235016

Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H. B., ... Thakurta, A. G. (2023). How to DP-fy ML: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, *77*, 1113–1201. doi: 10.1613/jair.1.14649

Qian, J., Jiang, H., Yu, Y., Wang, H., & Miao, D. (2024). Multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions. *Expert Systems with Applications*, *239*, 122343. https://doi.org/10.1016/j.eswa.2023.122343

Roopalakshmi, R., Kailas, S., & Sreelatha, R. (2026). Deep learning enabled pseudonymization for preserving data privacy of financial identifiers in public documents in India. *Scientific Reports*, *16*(1), 8120. https://doi.org/10.1038/s41598-026-39309-6

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, *32*. https://doi.org/10.48550/arXiv.1907.00503

Dhananjay M. Kanade
1Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, SavitribaiPhule Pune University, Pune, Maharashtra, India.
dmkanade@kkwagh.edu.in
ORCID 0000-0001-9469-9948

Snehal M. Kamalapur
2Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, SavitribaiPhule Pune University, Pune, Maharashtra, India.
smkamlapur@kkwagh.edu.in
ORCID 0000-0002-4125-0749

Shirish S. Sane
3Gokhale Education Society's R H Sapat College of Engineering, Management Studies and Research, Nashik, SavitribaiPhule Pune University, Pune, Maharashtra, India.
ORCID 0000-0002-4232-1051