Bima Journal of Science and Technology, Vol. 9(2A) Jun, 2025 ISSN: 2536-6041

DOI: 10.64290/bimagombe.v9i2A.1087

Improved Estimation of Attribute Privacy Disclosure Risk Using Machine Learning

Armaya'u Z. Umar[1]*, Muhammed M. Abubakar[2] and Mansir Abubakar[2]

Department of Software Engineering and Cyber Security, Al-Qalam University Katsina Katsina, Nigeria

Department of Computer Science and IT, Al-Qalam University Katsina, Katsina, Nigeria

Corresponding Author: azumar@auk.edu.ng

ABSTRACT

Personal data is widely used in predictive modelling and data analytics across various domains such as healthcare. Privacy-Preserving Data Publishing has emerged as sets of techniques for privacy protection when sharing data with analysts or researchers. It aims to balance the privacy of users with the utility of the dataset. One of the most popular approaches to privacy protection when publishing data is anonymization but it often leads to loss of data utility when applied uniformly across all attributes without considering the specific vulnerabilities of the different attributes. An approach was proposed to use machine learning to estimate the contribution of individual attributes to disclosure attack. However, the main challenge with the approach is its computational intensity and insensitivity to minority cases. This paper proposes to improve the assessment of attribute vulnerability to disclosure attacks and reduce computation overhead using adaptive outlier management, a fusion of data augmentation with Conditional Tabular Generative Adversarial Network, balancing using Synthetic Minority Over-sampling Technique for Nominal Data, and stratified sampling. The proposed approach results in over 38% reduction in computation overhead. Sensitivity to vulnerable attributes was also improved by up to 20.69%.

Keywords: Attribute vulnerability to disclosure attack, Data augmentation, Privacy preserving anonymization; Privacy-Preserving data publishing.

INTRODUCTION

The application of personal data in predictive modelling and data analytics is becoming increasingly common in many domains, including healthcare and national planning. Government agencies and healthcare providers can now collect, analyse, and leverage extensive medical and citizen data to forecast public health outcomes, plan community resources, and understand the needs of different population segments(Abubakar et al., 2022; Saura et al., 2021).

Information that can be used to identify a specific person is called personal data, or Personally Identifiable Information (PII)(Krishnamurthy et al., 2009). This includes direct identifiers (DIs) such as name, and quasi-identifiers (QIs) such as birthday, gender, and race. Some types of PII, like health conditions and personal income, are considered sensitive attributes (SAs). In this context, the person whose information is being collected is called the *data subject*, and the organization that manages the data is called the data *controller*.

Privacy is a fundamental human right(Wachter, 2017). Thus, organizations that handle personal data, whether private companies or government agencies (data controllers), have a crucial responsibility to protect the privacy of the data subject when they share the data with analysts or researchers. Consequently, privacy-preserving data publishing (PPDP) (Fung et al., 2010; Mendes & Vilela, 2017) have emerged as techniques for protecting the privacy of individuals whose records are in the dataset. It balances making valuable insights available with ensuring personal

DOI: 10.64290/bimagombe.v9i2A.1087

details remain undisclosed. PPDP techniques are based on either cryptographic methods or data anonymization.

Cryptographic techniques(Burkhalter et al., 2021; Mustafa et al., 2018), offer better protection but are computationally intensive. Consequently, anonymization techniques, where records in the dataset are transformed into less specific and indistinguishable forms, have also been developed(Abbasi & Mohammadi, 2022; Machanavajjhala et al., 2007; Rodriguez-Garcia et al., 2021; Song et al., 2019; Sweeney, 2002). Specific transformations for anonymization include generalization and suppression. As an example, consider Table 1 as raw data intended for publishing, generalization is the process of converting specific quasi-identifiers (QIs) into broader, less precise ranges to maintain data privacy, as exemplified by turning exact ages into age groups in Table 1. Suppression is the process of replacing original attribute values with the asterisk (*) symbol. This technique partially masks the attributes, making them less significant. For example, in Table 2, the last three digits of the QI attribute, Zip code, were suppressed. In both Table 1 and Table 2, Salary column represents the Sensitive Attribute.

Table 1: Raw data intended for publishing

Name	Age	Marital Status	Zip code	Sex	Salary
Joe	21	Separated	24028	M	>50K
Jill	26	Single	24030	M	>50K
Sue	32	Widowed	24035	F	≤50K
Abe	36	Separated	32035	F	≤50K
Bob	48	Widowed	32038	M	>50K
Alex	58	Married	32042	F	≤50K

Table 2: Generalized Age attribute and suppressed Zip code

Name	Age	Marital Status	Zip code	Sex	Salary
Joe	20-30	Separated	240**ˆ	M	>50K
Jill	20-30	Single	240**ˆ	M	>50K
Sue	30-40	Widowed	240**ˆ	F	≤50K
Abe	30-40	Separated	320**ˆ	F	≤50K
Bob	45-60	Widowed	320**ˆ	M	>50K
Alex	55-60	Married	320**ˆ	F	≤50K

As data controllers handle sensitive information, the challenge is not only to support useful analytics but also to do so while protecting privacy(Guan et al., 2019; Makhdoumi & Fawaz, 2013; Sankar et al., 2010). Current methods often fail because they apply anonymization uniformly without properly assessing which data attributes are most vulnerable to privacy breaches. This oversight can lead to over-anonymized data with reduced utility. Recently, there has been growing attention to anonymization methods centered on attributes. For example, Song *et al* (Song et al., 2019)have introduced an approach that involves introducing noise to numerical attributes and generalizing categorical attributes. Additionally, methods exist for identifying and anonymizing Quasi-Identifiers (QIs) in a manner that preserves utility and privacy (Li et al., 2022; Rodriguez-Garcia et al., 2021). However, a common drawback in many existing approaches is the absence quantifiable assessment of the vulnerability associated with individual attributes. This lack of vulnerability quantification, which could leverage advanced techniques like machine

DOI: 10.64290/bimagombe.v9i2A.1087

learning (ML), often results in higher utility loss and weaker privacy guarantees.

Motivation

The work in (Majeed & Hwang, 2023) proposed the use of machine learning, specifically, Random Forest, to estimate the vulnerability of quasi-identifiers (QIs) in order to decide the level of anonymization that should be applied to them. The researchers built a Random Forest model and checked how accurate it was (reference accuracy). Then, they focused on one quasi identifier (QI) and shuffled its values. With the shuffled values, they ran the model again and got a new accuracy score. The difference between the original and the shuffled accuracy scores indicates how important that quasi-identifier (QI) was. A big difference meant the original data had a lot of unique values. When they shuffled these unique values, it threw off the model's accuracy a lot. This means that quasi identifier (QI) is vulnerable because it could be used to identify someone and their sensitive information even after the data is supposedly anonymized(Zhang et al., 2019). Equations (1) through Equation (5) in Section III provide a theoretical analysis of the approach proposed in(Majeed & Hwang, 2023).

The approach(Majeed & Hwang, 2023) presented improvement over previous attribute-centric privacy-preserving techniques(Abbasi & Mohammadi, 2022; Li et al., 2022; Majeed, 2019; Rodriguez-Garcia et al., 2021; Rogovschi et al., 2022; Song et al., 2019; Srijayanthi & Sethukarasi, 2023). The strength of the approach lies in its robust assessment of dispersion. The use of Random Forest allows for the creation of an ensemble of decision trees to model the relationship between the QIs and the SAs. Random Forest is known for its robustness and ability to handle high-dimensional data. Nonetheless, the approach is computationally intensive, which may lead to scalability issues in practice. In addition, the approach is insensitive to minority quasi-identifiers in the dataset. This paper proposes to reduce the computation overhead, improve sensitivity to vulnerable attributes, and improve the approach's scalability through adaptive outlier management, fusion of data augmentation with data balancing, and stratified sampling.

MATERIALS AND METHODS

This section describes the materials and methods used in this research which include the adaptive outlier management, a fusion of data augmentation with data balancing, and data sampling.

Data Preprocessing

To implement the proposed approach, the *Adult dataset*[1] was used. It is the dataset commonly used as a benchmark in machine learning. It contains anonymized data from the 1994 U.S. Census and is publicly available for research purposes. The preprocessing phase focuses on cleaning and preparing the data for further analysis. The improved structure of the preprocessed table (dataset), T becomes $T[QI, Y]$, where $QI = \{q_1, q_2, \ldots, q_m\}$, and $Y = \{y_1, y_2, \ldots, y_n\}$. Missing values were filled in with the most frequent category for the categorical attributes and with the average value for the numerical attributes using a Simple Imputer - a machine learning tool that belongs to the scikit-learn library[2].

Adaptive Outlier Management and Vulnerability Estimation

Outlier management in privacy-preserving data publishing is particularly important because outliers can significantly skew results if not handled properly. They might represent rare but valid data points in the data collection process. Suppressing outliers completely can lead to biased conclusions against the minority(Currie & Rohren, 2022). On the other hand, giving outliers too much weight can also distort the analysis(Majeed

& Hwang, 2023). Outlier management in this paper is not adaptive. The data publisher sets a threshold percentage for outlier removal. Outliers exceeding the threshold are removed if their impact can be disregarded, but important outliers below the threshold are kept and might be augmented even though they might be rare. The specific outlier analysis and data augmentation for both the numerical and the categorical attributes proposed in this paper are presented in Algorithm 1.

```
Algorithm 1: Outlier Analysis and Data Augmentation
Require: T: Table with columns $qi_1$, qi2... qim
Require: zThreshold: Predefined z-score threshold, e.g., 3
Require: thresholdPercentage: Threshold percentage for outlier removal
Ensure: Outlier-managed table
1: zThreshold ← 3          //Default z-score threshold
2: numeric columns ← select numeric data (T)
3: outliers_dict ← ∅
4: for each column c in numeric columns do
5:    z ← calculate z_scores (T[c])
6:    outliers ← $\{i \mid |z_i| > zThreshold\}$
7:    outliers_dict[c] ← outliers
8: end for
9: total_outliers ← outliers_dict_values ()
10: if numericalOutlierPercentage > thresholdPercentage then
11:    AugmentNumerical(T)
12: end if
13: for each column c in columns of T do
14:    if c is categorical then
15:       minority class ← argmin_size(groupby(T[c]))
16:       minority class data ← T[T[c] = minority class]
17:    end if
18: end for
19: if categoricalOutlierPercentage > thresholdPercentage then
20:    AugmentCategorical(categoricalMinorityData)
21: end if
22: return the augmented dataset
```

The algorithm also identifies minority classes within categorical variables (lines 13-22 in Algorithm 1). More specifically, the minority class is the specific value, V_i, associated with the minimum cardinality, $n(V_i)$, in the specified column, q_i, of the table *T*. For a concrete example, let q_{i1} represent the ’race’ column in the Adult dataset. Minority classes in the ’race’ column are the specific values of race that appeared the least in the column ’race’.

For the initial vulnerability estimation using machine learning, this paper implemented, for replication, Equations (1) through Equation (5) of Section 3 as in (Majeed & Hwang, 2023). The implementation was tested with the raw dataset and with the different combinations of augmented, balanced, and sampled datasets. Next section presents our approach to reduce the computation overhead, improve sensitivity to vulnerable attributes, and improve scalability.

DOI: 10.64290/bimagombe.v9i2A.1087

Data Augmentation and Balanced Stratified Sampling (DABS)

Figure 1 represents the high-level architecture of DABS. In the Figure 1, the Conditional Generator is a component of a Conditional Generative Adversarial Network (CTGAN), a variant of Generative Adversarial Networks (GANs)(Gui et al., 2023). GANs are powerful tools for data augmentation, helping to artificially increase the size and diversity of a dataset. This can be useful for improving the performance of various machine learning models, particularly when dealing with limited data and when traditional data augmentation techniques, like rotations or flips(Cirillo et al., 2021), may not be sufficient to capture the true data distribution.

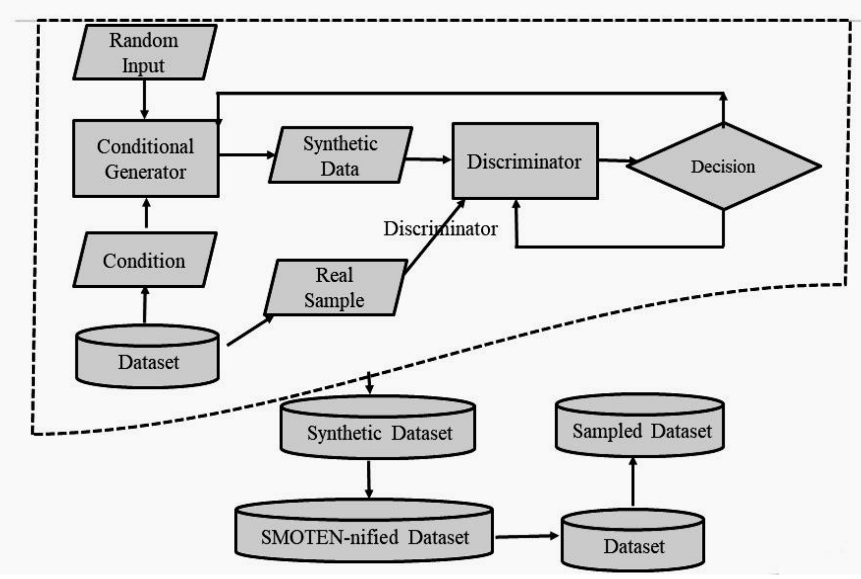


Figure 1: High-level representation of Data Augmentation and Stratified Sampling (DABS).

In particular, Conditional Tabular Generative Adversarial Network (CTGAN)(Xu et al., 2019) extends the conventional GAN framework for the synthesis of structured tabular data, ensuring that the synthetic samples not only resemble the marginal distribution of the original data but also maintain conditional dependencies between attributes. Rather than a simple Generator that produces samples from random noise, CTGAN uses a conditional generator, which takes additional input in the form of conditioning information, such as class labels or attributes from the real dataset. This conditioning information guides the generation process, allowing the generator to produce samples that align with specific conditions or criteria.

Thus, in CTGAN, initially, a sample condition is extracted from the real dataset and fed into the conditional generator along with the random input. This approach enables the preservation of dependency relationships(Xu et al., 2019). In the context of DABS, the random input is fed into the Conditional Generator component, and then the component uses this input along with its knowledge of the real dataset to produce synthetic data that shares similar characteristics with the real data. The discriminator component is trained to distinguish between the real data (from the original dataset) and the synthetic data generated by the Conditional Generator. Thus, it receives, as input, the synthetic data generated by the Conditional Generator and the real sample from the dataset. The discriminator then makes a decision about whether the synthetic data is realistic enough to be indistinguishable from the real data. This process helps the Conditional Generator

DOI: 10.64290/bimagombe.v9i2A.1087

to improve the quality of the synthetic data it generates over time.

Although not shown in the diagram, the synthetic data go through Synthetic Minority Over-sampling Technique for Nominal Data (SMOTEN) to have the dataset balanced. SMOTEN is used in machine learning to address the issue of class imbalance in nominal data, which means data with discrete categories. It uses linear interpolation between nearest neighbors within the same class to generate synthetic samples. This maintains the discrete nature of the data. In addition, it is computationally efficient for large datasets. The choice of SMOTEN and CTGAN was informed by extensive experiments in which SMOTEN was found to be modest in model performance but tends to generate data that does not respect the distribution of the dataset. On the other hand, CTGAN alone tends to generate data that leads to model overfitting. Overall when the data is augmented with CTGAN and balanced with SMOTEN, it tends to produce a large number of data points. Thus, stratified sampling was used to sample a fraction of the dataset that retains its statistical properties. Stratified sampling is a statistical technique for drawing samples from a population that ensures the sample accurately reflects the proportions of different subgroups (strata) within the population. Algorithm 2 represents the concise step from the stratified sampling.

```
Algorithm: 2 Stratified Sampling
Require: $N$: Total population size
Require: $N_i$: Size of stratum $i$
Require: $n$: Total sample size
Ensure: Sample size for each stratum $n_i$ for $i$ = 1 to number of strata
Calculate proportion per stratum:
for each stratum $i$ do
    $fi \leftarrow \underline{N}N\underline{i}$
end for
Calculate sample size per stratum:
for each stratum $i$ do
    $ni \leftarrow n \times fi$
end for
Return Sample Sizes:
Return the list of $n_i$ values
```

Algorithm 2 divides the population into strata, calculates the proportion of each stratum in the total population, determines the sample size for each stratum based on these proportions, and returns the sample sizes for each stratum to ensure a representative sample. The final processed dataset may be subjected to grouping to create k-anonymized records that respect l-diversity. In this work, the processed dataset was grouped using Jaccard Similarity (Equation 7 in Section 3).

Creating equivalence Classes using Jaccard Similarity

In contrast to(Majeed & Hwang, 2023) that used Cosine Similarity (see Equation 8 in Section 3) to ensure maximum similarity between users in creating equivalent classes by grouping similar records, this work utilizes the Jaccard Similarity index (see Equation 7 in Section 3). The Jaccard Similarity score is in the range of 0 to 1. If the two records are identical, the Jaccard Similarity is 1. The Jaccard similarity score is 0 if there are no common QIs between two records. Algorithm 3 is used for grouping records based on Jaccard Similarity. Notice that Algorithm 3 also computes Entropy, which is given in Equation 9. From Equation 9, a value of 0 means everyone in a group has the same label (identical SA values). This makes it very easy to figure out someone's SA value, which is a privacy risk. On the other hand, a value of 1 means there is enough diversity in the label (diverse SA values). This makes it difficult to guess any one SA's value, reducing the privacy risk.

$$\Gamma(\mathrm{Ci}) = \sum_{i=1}^{|YCi|} pilog2pi \text{ -----------------9}$$

Consequently, the vulnerability value, τ, from Equation 5 and the diversity value, Γ, from Equation 9 should derive the adaptive anonymization applicable to each group of k users whose similarity was computed using the Jaccard Index (Equation 7). Intuitively, Jaccard Similarity generally has a lower overhead compared to Cosine Similarity. This is because Jaccard Similarity involves finding the intersection and union of sets, which are relatively simple operations. This translates to lower computational cost.

Algorithm 3: Group Records based on Jaccard Similarity and Calculate Entropy

Require: datasets containing the records tobe grouped

Require: threshold: Jaccard similarity threshold

Ensure: groups: List of dictionaries containing groups and their respective entropies

```
$num\ records$ ← length of records
$groups$ ← empty list
for $i$ ← 0 to $num\ records - 1$ do
    $group$ ← [$i$] // Initialize a group with the current record
    for $j$ ← $i + 1$ to $num\ records - 1$ do
        $similarity$ ← JaccardSim($dataset[i], dataset[j]$)
        if $similarity \geq threshold$ then
            $group$.append($j$)
        end if
    end for
    $group\ data$ ← $data\ sets[group]$
    $probabilities$ ← Probabilities($group\ data$['income'])
    $entropy$ ← CalculateEntropy($probabilities$)
    $groups$.append({'group' : $group$,'entropy' : $entropy$})
```

DOI: 10.64290/bimagombe.v9i2A.1087

```
15: end for
16: return groups
```

In addition, Jaccard Similarity works well with categorical features and most machine learning tasks will likely require encoding of the categorical features. In contrast, as observed by(Khan et al., 2021), Cosine Similarity involves calculating the dot product and vector magnitudes, which require more computations compared to set operations. Furthermore, Cosine Similarity often works better with continuous numerical data, which is not always the case. However, the Jaccard similarity may not be suited for all cases as the index may be a poor metric if there are no positives for some samples or classes. The key here is to be adaptive to the specific dataset to be published.

Evaluation Metrics

1) Expressiveness of τ

Expressiveness of *τ* refers to the level of detail it offers in capturing the vulnerability of an attribute. A highly expressive *τ* would provide a more pronounced score.

2) Computation overhead

Generally, in data processing applications, computational complexity refers to the amount of resources required by an algorithm to process and analyze a dataset. The complexity can be measured theoretically using asymptotic notation and empirically by recording the actual running time of the algorithm on a chosen machine.

3) F1 score

F1 score is used for the assessment of the performance of a classification model by considering both precision and recall. Mathematically, F1 is defined by Equation 10.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (10)$$

The Precision and the Recall are given in Equations (11) and (12) respectively:

$$\text{Precision} = \frac{\text{True Positive Predictions}}{\text{True Positive Predictions} + \text{False Positive Predictions}} \text{-----------------}11$$

$$\text{Recall} = \frac{\text{True Positive Predictions}}{\text{True Positive Predictions} + \text{False Negative Predictions}} \text{-----------------}12$$

THEORY

Using Random Forest Model to Estimate Attributes' Vulnerability to Disclosure Attack

The study in(Majeed & Hwang, 2023) found that shuffling the QI values creates a unique impact on each individual tree (Ti) within a Random Forest. This impact can be represented as E (qi), where E refers to the effect and qi represents the specific QI value for the *ith* tree. The effect E of a particular QI (e.g., qi) in the b can be determined via Eq. 1:

$$E(qj) = \frac{\sum i \in \zeta^{(b)} I(yi = \hat{y}i^{(b)})}{|\zeta\,(b)|} - \frac{\sum i \in \zeta^{(b)} I(yi = \hat{y}^{(b)}_{i,\pi j})}{|\zeta\,(b)|} \text{-----}(1)$$

The first term of Equation 1 calculates the accuracy of tree *b* on the OOB samples where the values of *qj* are not shuffled. The second term of Equation 1 calculates the accuracy of tree *b* on the OOB samples where the values of *qj* are shuffled column-wise.

The value V (E(qj)) in b can lie in two types depending upon the availability:

$$V(E(qj)) = \begin{cases} E^{b}(qj), & if\ qj\epsilon b \\ 0, & otherwise \end{cases} \text{---------------}2$$

The $V\ (E(q_j)) = 0$ if a q_j is not part of *b* or encompasses only identical values in a column. The mean ($\bar{m}_{qi}$) of *E* for each QI from all trees is then determined using Eq. 3.

$$\overline{m}\text{qi} = \frac{\sum_{t=1}^{ntree} E^{2t}(qi)}{ntree} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.3$$

The mean, $\bar{m}_{qi}$, represents the average predictive power across all the trees in the forest. The standard deviation, s_{qi}, and the vulnerability value (henceforth referred to as τ) are calculated using Equations 4 and 5, respectively.

$$sqi = \sqrt{\frac{1}{ntree-1}\sum_{t=1}^{ntree}(E^{t}(qi) - \overline{mqi})^{2}} \text{-----------}4$$

$$\tau = \frac{sqi}{m_{qi}} \text{---}5$$

More specifically, the ensemble structure of Random Forest allows for a robust evaluation of individual attribute importance across multiple trees. To estimate the vulnerability of each QI, for each tree in the forest, the portion of data not used during training—referred to as OOB samples—is used for testing. The central idea is to measure how much the predictive accuracy of each tree is affected when the values of a specific QI are shuffled in these OOB samples.

By comparing the model's accuracy before and after shuffling the values of a QI, this study was able to quantify the QI's influence on the predictions as in(Majeed & Hwang, 2023). A significant drop in accuracy after shuffling indicates that the QI plays an important role in the model's decision-making process, suggesting a higher potential for re-identification or disclosure if that attribute is exposed. The use of Random Forest in this manner is particularly powerful because it provides a systematic, data-driven way to assess risk.

Z-statistics

Z-statistics, or Z-scores, is a way to measure how far a particular data point is from the mean in a standard normal distribution which is useful for finding outliers. Z score is represented by Equation 6.

In Equation (6) above, *X* is the data point, μ is the mean of the distribution and σ is the standard deviation of the distribution.

Jaccard Similarity

$$Z = \frac{X-\mu}{\sigma} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

$$J(A,B) = \left|\frac{A\cap B}{A\cup B}\right| \text{-------------------------}7$$

RESULTS

The box plot in Figure 2 shows a slight reduction in the number of outliers after removing outliers in the age attribute. This is

DOI: 10.64290/bimagombe.v9i2A.1087

evident by the shorter whiskers in the "After Removing Outliers" plot compared to the "Before Removing Outliers" plot. Similarly, the Interquartile Range (IQR) which is the distance between the first and third quartiles, is relatively smaller in the ”After Removing Outliers” plot compared to the ”Before Removing Outliers” plot. This indicates that the overall spread of the data is reduced after removing outliers, making the data more concentrated around the middle quartiles. However, the outlier in age was only removed because it was negligible, in this case, and it may not introduce bias to certain age groups. For the race attribute, to avoid being biased against the minority race, we augmented the minority data with CTGAN even though the ratio of the minority race to the majority is 1:120! Before the data augmentation, the sample of the minority 'race' attribute is printed and shown in Figure 3.

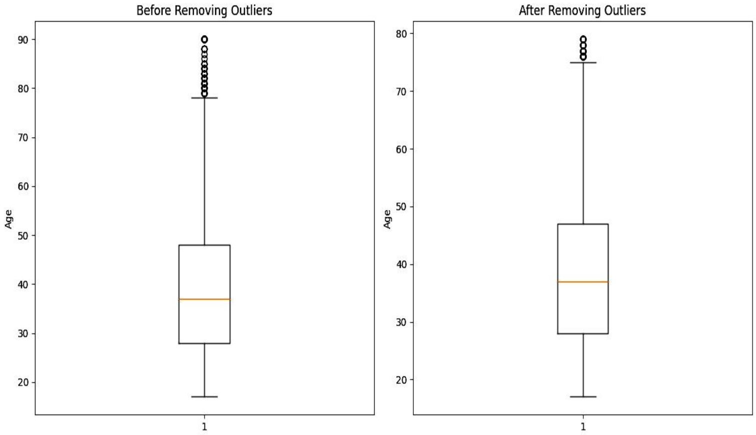
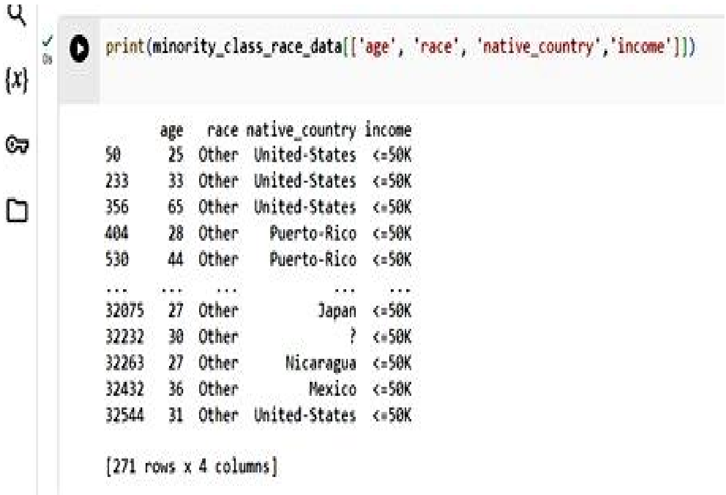


Figure 2: Adult Dataset with outliers in age vs when the outliers in age have been removed.



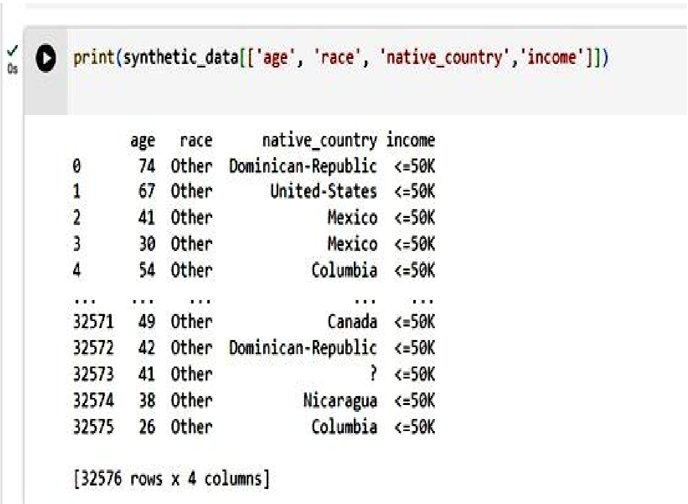
```
print(minority_class_race_data[['age', 'race', 'native_country','income']])

       age   race native_country income
50      25  Other  United-States  <=50K
233     33  Other  United-States  <=50K
356     65  Other  United-States  <=50K
404     28  Other    Puerto-Rico  <=50K
530     44  Other    Puerto-Rico  <=50K
...    ...    ...            ...    ...
32075   27  Other          Japan  <=50K
32232   30  Other              ?  <=50K
32263   27  Other      Nicaragua  <=50K
32432   36  Other         Mexico  <=50K
32544   31  Other  United-States  <=50K

[271 rows x 4 columns]
```

Figure 3: Sample of minority 'race' in the Adult dataset.

The snapshot data in Figure 4 is totally synthetic but notice how it resembles the original data. As a matter of fact, the evaluation of how well the distributions of individual columns in the synthetic data match those in the real data, using Kolmogorov-Smirnov (KS) statistic and Total Variation (TV) complement, suggested high quality of some of the generated data in some attributes such as age. The next section presents the results of how our approach improves the sensitivity to the detection of vulnerable attributes through more expressive τ values.



```
print(synthetic_data[['age', 'race', 'native_country','income']])

       age   race      native_country income
0       74  Other  Dominican-Republic  <=50K
1       67  Other       United-States  <=50K
2       41  Other              Mexico  <=50K
3       30  Other              Mexico  <=50K
4       54  Other            Columbia  <=50K
...    ...    ...                 ...    ...
32571   49  Other              Canada  <=50K
32572   42  Other  Dominican-Republic  <=50K
32573   41  Other                   ?  <=50K
32574   38  Other           Nicaragua  <=50K
32575   26  Other            Columbia  <=50K

[32576 rows x 4 columns]
```

Figure 4: Sample of the synthetically generated data based on minority ‘race’ in the Adult dataset.

Sensitivity to Vulnerable Attribute

Looking at Figure 5, it is clear that even with a smaller dataset DABS, shows more sensitivity to vulnerability than the existing approach(Majeed & Hwang, 2023). This is because, DABS has incorporated data augmentation, balancing, and stratified sampling, leading to a potentially more expressive τ. Notice that in both approaches, age appeared to be more vulnerable to disclosure attacks. In DABS, the next more vulnerable attribute is race, thanks to the augmentation using CTGAN, otherwise it could not have been detected as vulnerable to disclosure since the minority races would have been diluted or wiped in the inference. This underscores the utility of our approach compared to the status quo. Both approaches rank sex as more vulnerable than native country. However, the vulnerability of ‘Native Country’ attribute in DABS is more pronounced as a result of the cascading effect of augmenting the ‘race’ attribute.

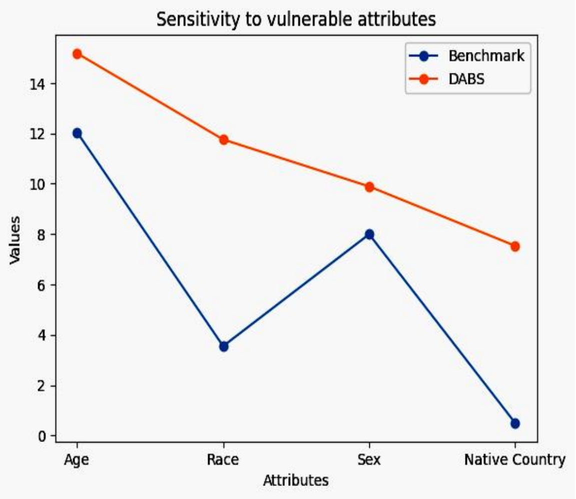


Figure 5: Sensitivity to Vulnerable Attributes: DAS vs Status Quo(Majeed & Hwang, 2023).

Computation Overhead

We evaluated the computational overhead of DABS by measuring and comparing its execution time with that of the benchmark(Majeed & Hwang, 2023) when computing the τ values of four attributes (age, race, sex, and native country) in the adult

DOI: 10.64290/bimagombe.v9i2A.1087

dataset. We used Jupiter Notebook running on a Windows machine with an Intel(R) Core(TM) i7-4600M CPU 2.90GHz, 2.90 GHz, and 16GB of RAM. The Benchmark took 3,664.780069589615 seconds, which translates to approximately 1.02 hours. However, DABS took only 2,284.66241312027 seconds, or roughly 0.63 hours, representing a reduction in execution time of 38.24%. This improvement is attributed to our approach's utilization of data augmentation and stratified sampling to reduce the dataset size while preserving its statistical properties.

F1 Score

As shown in Fig. 6, DABS achieves the highest F1 score (0.81) indicating maintains the best performance. CTGAN + SMOTEN shows improvement over CTGAN alone, suggesting that data augmentation with synthetic data and balancing the class distribution can be beneficial. Raw and Sampled Raw have lower F1 scores, which means that directly using the raw data or simply sampling it might not be the most effective approach for this task.

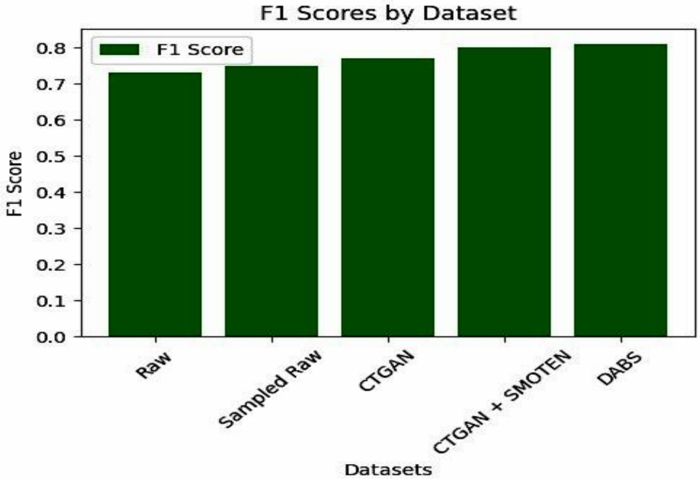


Figure 6: F1 score.

Discussion

The findings of age as the most vulnerable attribute in the Adult dataset are consistent with the finding in the benchmark work. Albert in our approach, the vulnerability is relatively more pronounced in the benchmark work, the vulnerability of attributes related to minority races might be underestimated due to a lack of data points. Thus, patterns observed in the majority group might be misinterpreted as representing the entire population. Consequently, the detection of race as vulnerable next to age in our approach underscores the utility of the proposed adaptive outlier management. Specifically, the utilization of CTGAN to augment the data for minority races was proved useful.

More clearly, DABS help to identify race as the next most vulnerable attribute, likely due to the effectiveness of CTGAN in addressing the minority data imbalance. The implication is that without augmentation, race vulnerability might have been missed. Overall DABS shows increased sensitivity in detecting vulnerable attributes, even with a smaller dataset. This is attributed to DABS' techniques like data augmentation, balancing, and stratified sampling. DABS achieves a substantial reduction (over 38%) in execution time compared to the benchmark approach(Majeed & Hwang, 2023) This is expected as the dataset was sampled in strata to preserve the statistical properties of the dataset. However, while data augmentation and stratified sampling improve efficiency,

DOI: 10.64290/bimagombe.v9i2A.1087

there might be a slight decrease in the accuracy of τ scores compared to using the entire dataset.

CONCLUSION

This paper proposes an approach to enhance privacy preservation and data utility in personal data management. Through the development of ML-based adaptive outlier management and the introduction of Data Augmentation and Balanced Stratified Sampling (DABS), significant strides have been made in reducing computation overhead and improving accuracy in vulnerability estimation. By replacing Cosine Similarity with Jaccard Similarity, computation overhead has been reduced by over 38%, while sensitivity to vulnerable attributes has been improved by up to 20.69%. Moving forward, this research calls for the implementation of reusable and flexible libraries, consideration of attribute vulnerability in privacy preservation frameworks, and the integration of fairness awareness in machine learning models. External validation and refinement of evaluation metrics are equally essential for ensuring the reliability and effectiveness of privacy preservation techniques across diverse contexts.

REFERENCES

Abbasi, A., & Mohammadi, B. (2022). A clustering-based anonymization approach for privacy-preserving in the healthcare cloud. *Concurrency and Computation: Practice and Experience*. https://doi.org/10.1002/cpe.6487

Abubakar, M. M., Armaya'u, Z. U., & Abubakar, M. (2022). Personal Data and Privacy Protection Regulations: State of compliance with Nigeria Data Protection Regulations (NDPR) in Ministries, Departments, and Agencies (MDAs). *2022 5th Information Technology for Education and Development (ITED)*, 1–6.

Burkhalter, L., Küchler, N., Viand, A., Shafagh, H., & Hithnawi, A. (2021). Zeph: Cryptographic enforcement of end-to-end data privacy. *Proceedings of the 15th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2021*.

Cirillo, M. D., Abramian, D., & Eklund, A. (2021). WHAT IS THE BEST DATA AUGMENTATION FOR 3D BRAIN TUMOR SEGMENTATION? *Proceedings - International Conference on Image Processing, ICIP*. https://doi.org/10.1109/ICIP42928.2021.9506328

Currie, G., & Rohren, E. (2022). Social Asymmetry, Artificial Intelligence and the Medical Imaging Landscape. In *Seminars in Nuclear Medicine*. https://doi.org/10.1053/j.semnuclmed.2021.11.011

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*. https://doi.org/10.1145/1749603.1749605

Guan, Z., Lv, Z., Du, X., Wu, L., & Guizani, M. (2019). Achieving data utility-privacy tradeoff in Internet of Medical Things: A machine learning approach. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2019.01.058

Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2023). A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, *35*(4), 3313–3332. https://doi.org/10.1109/TKDE.2021.3130191

Khan, M. S., Anjum, A., Saba, T., Rehman, A., & Tariq, U. (2021). Improved Generalization for Secure Personal

DOI: 10.64290/bimagombe.v9i2A.1087

Data Publishing Using Deviation. *IT Professional*. https://doi.org/10.1109/MITP.2020.3030323

Krishnamurthy, B., on, C. W.-P. of the 2nd A. workshop, & 2009, undefined. (2009). On the leakage of personally identifiable information via online social networks. *Dl.Acm.Org*. https://dl.acm.org/doi/abs/10.1145/1592665.1592668

Li, S., Schneider, M. J., Yu, Y., & Gupta, S. (2022). Reidentification Risk in Panel Data: Protecting for k - Anonymity . *Information Systems Research*. https://doi.org/10.1287/isre.2022.1169

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). ℓ-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1). https://doi.org/10.1145/1217299.1217302

Majeed, A. (2019). Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2018.03.014

Majeed, A., & Hwang, S. O. (2023). 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

Makhdoumi, A., & Fawaz, N. (2013). Privacy-utility tradeoff under statistical uncertainty. *2013 51st Annual Allerton Conference on Communication, Control, and Computing, Allerton 2013*. https://doi.org/10.1109/Allerton.2013.6736724

Mendes, R., & Vilela, J. P. (2017). Privacy-Preserving Data Mining: Methods, Metrics, and Applications. *IEEE Access*. https://doi.org/10.1109/ACCESS.2017.2706947

Mustafa, G., Ashraf, R., Mirza, M. A., Jamil, A., & Muhammad. (2018). A review of data security and cryptographic techniques in IoT based devices. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3231053.3231100

Rodriguez-Garcia, M., Balderas, A., Sciences, J. D.-A., & 2021, undefined. (2021). Privacy preservation and analytical utility of E-learning data mashups in the web of data. *Mdpi.ComM Rodriguez-Garcia, A Balderas, JM DoderoApplied Sciences, 2021•mdpi.Com*. https://doi.org/10.3390/app11188506

Rogovschi, N., Bennani, Y., & Zouinina, S. (2022). Data Anonymization Through Multi-modular Clustering. In *Studies in Big Data*. https://doi.org/10.1007/978-3-030-95239-6_6

Sankar, L., Rajagopalan, S. R., & Poor, H. V. (2010). A theory of utility and privacy of data sources. *IEEE International Symposium on Information Theory - Proceedings*. https://doi.org/10.1109/ISIT.2010.5513684

Saura, J. R., Ribeiro-Soriano, D., & Palacios-Marqués, D. (2021). From user-generated data to data-driven innovation: A research agenda to understand user privacy in digital markets. In *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2021.102331

DOI: 10.64290/bimagombe.v9i2A.1087

Song, F., Ma, T., Tian, Y., & Al-Rodhaan, M. (2019). A New Method of Privacy Protection: Random k-Anonymous. *IEEE Access*. https://doi.org/10.1109/ACCESS.2019.2919165

Srijayanthi, S., & Sethukarasi, T. (2023). Design of privacy preserving model based on clustering involved anonymization along with feature selection. *Computers and Security*. https://doi.org/10.1016/j.cose.2022.103027

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*. https://doi.org/10.1142/S0218488502001648

Wachter, S. (2017). Privacy: Primus Inter Pares — Privacy as a Precondition for Self-Development, Personal Fulfilment and the Free Enjoyment of Fundamental Human Rights. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2903514

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, *32*.

Zhang, C., Wu, S., Jiang, H., Wang, Y., Yu, J., & Cheng, X. (2019). Attribute-enhanced de-anonymization of online social networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-34980-6_29