Journal of Information Systems Engineering and Management
2025, 10(9s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Secure Data Sharing in Machine Learning: Exploring K-Anonymization and Differential Privacy

Malath Sabri Kareem
[1]Middle Technical University. [1] malath-sabri@mtu.edu.iq

ARTICLE INFO

Received: 30 Oct 2024

Revised: 21 Dec 2024

Accepted: 03 Jan 2025

ABSTRACT

The problem of data privacy has now become one of the most pressing ones, particularly with the growth of the number of people who use their identity data to sign up for various services. This work examines the combined application of k-anonymity and differential privacy approaches to achieve privacy preservation on health related information while maintaining its usefulness for analysis. The concept of K-anonymity tries to ensure that nobody can be uniquely identified from his records by making each record indistinguishable from at least k-1 other records; differential privacy also offers a measure to unable a person's contribution to the overall data measurement outcome. These techniques are used in this study to health-related datasets, and assess the performance of a combined approach to achieving privacy and data utility. A significant piece of empirical research is provided to show that a combination of k-anonymity and differential privacy is not only realistic in a practical application but also provides optimum privacy protection while inflicting minimum loss in data quality. The reduction in reidentification risk is by a whopping 80% when data is processed through the combined method, but the utility of the data gathered only equals 85% the extent of the original data gathered from raw contextual information. Doing so proves the applicability of these privacy-preserving methods for the protection of health data represented by the results given in the paper.

Keywords: Data Privacy, K-Anonymity, Differential Privacy, Privacy Preservation, Health Data, Data Utility, Anonymization, Machine Learning, Data Protection, Privacy-Utility Trade-off.

INTRODUCTION

One of the biggest issues in current society is information security, particularly with such a large amount of personal and sensitive information that is generated, collected and shared. One of the most basic problems with data is the ability to anonymise and secure the data while still being able to use it. Among the many methods that were designed to approach this problem, the most studied and frequently utilized ones are k-anonymity and differential privacy (Tsou et al., 2021). K-anonymity guarantees that an individual can not be identified from at least K others as in the case of the following example. k-1 other records but at the same time, differential privacy that provides a commitment that the inclusion or exclusion of any record's data does not impact any result of any analysis that is carried out on those records (Majeed et al., 2022).

New works have examined the combination of these two methods, considering how to enhance the effectiveness of the privacy-preserving models and reduce the likelihood of re-identification while maintaining data accuracy. This paper by Tsou et al. (2021) confirm that using k-anonymity-and differential privacy is a promising way towards achieving both privacy and utility while releasing sensitive data. According to their works, the above hybrid models can provide convincing privacy-preserving solutions for various data sharing contexts including sharing patient data in health-care sectors, or releasing social network data.

However, data privacy related issues have been compounded especially where issues to do with nursing health data are in question. Regarding the restricted use of privacy preserving techniques and data anonymization in healthcare related databases greatly enriched by private information, Karagiannis et al. (2024). The privacy risks are quite high, and so techniques such as k-anonymity and differential privacy become essential. Our work expands on such frameworks by posing the question of how these methods can be used for datasets with health information in order to maintain privacy while maximizing the usefulness of the data as much as possible.

As shown in several previous works, k-anonymity and differential privacy are still efficient for privacy preservation, but some issues still persist particularly at the level of privacy / utility trade-off. Goldsteen et al. (2021) have noted that, even though it serves to effectively obscure an individual's identity, k-anonymity degrades the data quality notably, when the data is highly generalized . Likewise

Ratra et al., (2022) and Li (2023) suggest that these methods are limited in their application where there is sensitive information to be gathered.

This research aims to fit the bill with proposing methods for adjusting the variables of differential privacy to reduce the loss of utility and increase the robustness of the dataset against attacks. On the same note, people are developing interest in clustering methods in anonymization to get a higher quality of the outcome. Some of the clusters have also identified by the Majeed et al. (2022) that it imposed less reliance on generalization and suppression and retains maximum data integrity. In this research, it endeavours to establish how clustering procedures can be used efficiently with k-anonymity and differential privacy approaches.

The main focus of this paper is to analyze k-anonymity and deterministic differential privacy when applied to the sensitive health data to assess the level of privacy protection with as little loss of data utility as possible. In a way, this study helps to continue the work towards defining how data protection can be maintained in fields where big data is increasingly prominent. Lastly, the ultimate outcomes of this research are to identify ways through which organizations and researchers can enhance the development and the usability of privacy-preserving data release techniques while still ensuring data privacy and utility.

3. METHODOLOGY

This research also applies a scoped framework to investigate the secure data sharing in machine learning model through the use of K-Anonymization and Differential Privacy. Namely, the central goal is to maintain the privacy of the data while maximizing the usefulness of the data to researchers and organizations. The methodology is structured into four key stages: preprocessing and K-Anonymization process, the extension of Differential Privacy and Utility-Precision analysis.

The first step in the process is to prepare the dataset and that involves, handling of the missing values, normalizing the numerical ,attributes and splitting of the dataset into training and testing datasets. They first convert the data so that it is cleansed to include some privacy preserving steps to the data before the actualonymizing process. Where data is missing, it is neatly filled using forward-filling technique while normalization and feature scaling are used to ensure comparability of different features in the data set. The dataset applied in this research includes personal data, demographic and financial data, making it a good candidate for privacy preservation.

In the next stage, K-Anonymization is applied to mask the individual record in order to guarantee record anonymity from at least k–1 other records based on quasi-identifier.

Technique known as generalization and suppression is used in order to process sensitive attributes such as age, gender and location. For example age values can be categorically defined to embrace certain age range while zip codes are only coded to represent a general geographical area. This is done by making it possible for every record to have at least k entries that cannot be distinguished from each other, helping to reduce identity disclosure that is a problem with K-Anonymization.

The third stage, again, concerns the application of Differential Privacy to enhance the protection of the dataset against re-ID attacks. Differential Privacy applies controlled randomisation to the attributes under on aggregation or sharing of the information reducing individual contribution on the data base. Therefore, the Laplace mechanism is applied in this research with the noise level being controlled by a privacy parameter, ϵ, and the sensitivity of the dataset. Lower value of ϵ means that the privacy is stronger, but it is not very useful for data analysis. This step is taken to enhance the privacy despite the possibility of aggregating statistical data.

Finally, the efficiency of the methodology at balancing between privacy and utility is analyzed with the help of training machine learning models both on the original and sanitized datasets. Accuracy rates and F1-score are used to measure the fact that privacy preservation has a negative effect on the abilities of the developed models. The results enable one to quantify the cost of preserving data protectiveness and the benefits for using it in machine learning.

Through a factorial approach of alternating the processes of K-Anonymization and applying Differential Privacy this methodology allows for a secure approach to data sharing in machine learning. The approach shows that privacy-preserving techniques can be incorporated into actual machine learning processes, and allows, thus, the creation of safe and responsible data-sharing models.

3.1 Dataset Preparation

The first subset of the chosen methodology is **data pre-processing**. As any data analyst would agree, this stage involves preparing the datasets for analysis so that they will be valid and amenable to privacy-preserving techniques. The dataset of this research has some sensitive attributes, including personal, financial, or health attributes, and hence is suitable for secure data-sharing analysis. In the dataset preparation phase, missing values have to be dealt with, the values of the numerical attributes have to be normalized, and the dataset needs to be split into the training set and the testing set (Table 1).

Missing Values

The first process of dealing with the dataset is handling missing values. To achieve this, it is mandatory to ensure that the dataset does not have any blank or nil records, as these could compromise the integrity of the data and the resulting machine learning model. In this research, the imputation of missing values is done through the forward-fill technique in which missing values of a variable are filled by the immediate preceding value. This method performs well when data is MAR or, in other words, missing data is Missing at Random and is advantageous when time continuity is important, especially for time-series data sets.

This paper will propose the following normalization of the numerical attributes:

The only problem left for handling missing values, the next step is normalizing the numerical attributes. In the case of different features in the dataset, their values can be in different scales, and hence scaling is required to transform the values of features to [0 1]. This helps in avoiding situation whereby some features with large scales influence the machine learning algorithm. For instance, age features are scaled with the min-max scaling method where features are ranging from 18 to 90. This is useful in order to avoid a particular feature to have a much larger influence than the other features because of scaling differences.

Train-Test Split

Last but not the lest, cross-validation divides the data into training and testing. A standard way of splitting your data is the so-called ‘training subset,’ which refers to 80% of your data for building the model, and the test set, which is 20% of the data. Sampling is conducted arbitrarily for the two partitions so that each of them contains examples of the whole data samples, and so that the model does not get spoiled when evaluation is being carried out.

The following Python code was used to implement these preprocessing steps:

python

```
import pandas as pd
from sklearn.model_selection import train_test_spli
# Load dataset
data = pd.read_csv('sensitive_data.csv')
# Preprocess: Handle missing values
data.fillna(method='ffill', inplace=True)
# Normalize specific columns
data['age'] = (data['age'] - data['age'].min()) / (data['age'].max() - data['age'].min())
# Split data into train and test sets
train_data, test_data = train_test_split(data, test_size=0.2, random_state=42)
print("Dataset prepared with train-test split.")
```

This step consists much in preparing and cleansing the dataset, in which preprocessing takes into account data normalization and division, making it ready for privacy-preserving methods like K-Anonymization and Differential Privacy. Ensuring the dataset is void of missing values, normalized, and properly divided allows for subsequent analysis and model development to not fall prey into errors or biases.

Algorithm for Train-Test Split

Input:

- Dataset (data)
- Test size (test_size = 0.2)
- Random seed for reproducibility (random_state = 42)

Output:

- Training dataset (train_data)
- Testing dataset (test_data)

Steps:

1. **Load Dataset**
Load the dataset into a DataFrame using the appropriate method (e.g., CSV file).

2. **Handle Missing Values**
fill any missing values in the dataset using a forward-fill method (fillna(method = 'ffill')) to ensure data consistency.

3. **Normalize Columns**
Normalize specific columns (e.g., age) to scale values between 0 and 1 using the formula:

$$\text{normalized\_value} = \frac{\text{value} - \text{min}}{\text{max} - \text{min}}$$

4. **Divide Data into Train and Test Sets**
Use the train_test_split function from the sklearn.model_selection library to split the dataset into training (80%) and testing (20%) subsets.

5. **Random Sampling**
Ensure that the data is split randomly while maintaining reproducibility by setting a random_state value.

6. **Return Results**
Output the training and testing datasets for use in subsequent analysis or model training.

Table.1 summarizes Dataset Preparation

Step	**Description**	**Method Applied**	**Outcome**
1. Handling Missing Values	Address missing or null values in the dataset to ensure completeness and consistency.	Forward-Fill	Missing values are replaced with the most recent valid observation.
2. Normalization of Numerical Attributes	Scale numerical features to a uniform range to prevent bias due to differing scales.	Min-Max Scaling	All numerical attributes are scaled to the range of 0 to 1.
3. Data Partitioning	Split the dataset into training and testing sets to avoid overfitting and enable model evaluation.	80-20 Split (Random Partition)	80% of the data used for training and 20% for testing.

3.2 K-Anonymization Implementation

The major technique studied in this work is called K-Anonymization, which focuses on the privacy preservation of individuals within a given dataset in a way that guarantees no record can be distinguished from at least k–1 other records. The general aim of K-Anonymization is, therefore, to conceal or generalize the specific information that might be used to identify the person in question. This technique is specially helpful in case of handling obliterative and conditional attributes such as age, gender, and zip code. Here, in the light of this research, we apply the K-Anonymization on a sensitive dataset and give priority to laying middle ground between privacy and utility.

K-Anonymization process also starts by identifying the quasi-identifiers present in the dataset. Quasi-identifiers are fields that are not in themselves identifiable but, when combined with other fields, cause the identification of values of other fields. For example, age, gender, or postal code might be quasi-identifiers where it is probable that somebody would be identified when linked with other external pieces of information.

python

```
from kanonymity import KAnonymizer
# Define quasi-identifiers
quasi_identifiers = ['age', 'zip_code', 'gender']
# Apply K-Anonymization
k = 3
anonymizer = KAnonymizer(k)
anonymized_data = anonymizer.anonymize(data, quasi_identifiers)
# Save anonymized data
anonymized_data.to_csv('anonymized_data.csv', index=False)
print(f"K-Anonymization applied with k={k}.")
```

After quasi-identifiers have been determined, the next step is to perform generalizations and suppression techniques. Generalization refers to the use of substituted general categories where specific ones were used earlier. For example, a specific age of 29 may be changed to age 25-30. There are two forms of data anonymization: generalization and suppression, whereby generalization is the substitution of specific data with a less discriminative value, and suppression is the removal or replacement of data values with a universal value. These techniques guarantee that each record belongs to a set containing at least k other records with close quasi-identifier values.

In K-Anonymization process, we assign a value to k where any unique record has to be included in at least k groups. It is inversely proportional to privacy since a higher value of k means grouping records in a way that makes it hard to profile any person out of the numerous records grouped together. But, a higher value of k may reduce data utility as anonymization outcome often compromises the level of detail of the data collected. In this work, we use various configurations of k to establish how K-Anonymization affects privacy and utility.

Once this anonymity process is done, the dataset is evaluated to determine whether all records garnered the k-anonymity level. This gist means for any one or a set of quasi-identifiers, there are at least k records with the same value for the quasi-identifiers. The anonymized dataset is then stored to be used later for comparison with the actual dataset collected in this study.

The use of K-Anonymization improves privacy since every record does not bear the specific identity of a particular person and enhances canopy privacy by the improvement of its capability of giving low re-identification risks.

In particular, K-Anonymization encompasses the identification of quasi-identifiers, the generalization and suppression methods, and the evaluation of the dataset norming to the k-anonymity principle. This procedure is consistent with protecting individual data privacy and at the same time assures that proper data analysis is possible when the data are anonymized as a way of minimizing identification of the individuals involved.

Algorithm for K-Anonymization Implementation

Input:

- Dataset (data)
- Quasi-identifiers ($quasi\_identifiers$)
- Anonymization parameter (k)

Output:

- Anonymized dataset ($anonymized\_data$)

Steps:

```
1. Import Required Libraries
  Import modules for data processing, machine learning, and evaluation metrics (e.g., pandas, sklearn, matplotlib).
2. Load Dataset
  Read the dataset (`data`) and split it into training (`train_data`) and testing (`test_data`) sets.
3. Train Model on Original Data
  a. Separate features (`X_train`) and target variable (`y_train`) from `train_data`.
  b. Train the model (e.g., `RandomForestClassifier`) on the original dataset.
  c. Evaluate metrics (Accuracy, Precision, Recall, F1-Score) on `test_data`.
4. Apply K-Anonymization
  a. For each `k` in a predefined range:
    i. Apply generalization and suppression to create an anonymized dataset.
    ii. Train the same model on the anonymized dataset.
    iii. Evaluate metrics on `test_data`.
    iv. Store the metrics for plotting.
5. Apply Differential Privacy
  a. For each `ε` in a predefined range:
    i. Add noise to the dataset to achieve Differential Privacy.
    ii. Train the same model on the differentially private dataset.
    iii. Evaluate metrics on `test_data`.
    iv. Store the metrics for plotting.
6. Visualize Trade-Off Analysis
  a. Plot evaluation metrics (e.g., accuracy, precision) against `k` for K-Anonymization.
  b. Plot evaluation metrics against `ε` for Differential Privacy.
  c. Compare the performance of both privacy-preserving methods.
7. Compare Privacy-Utility Trade-off
  a. Analyze the impact of increasing `k` and decreasing `ε` on privacy and utility.
  b. Identify optimal values for `k` and `ε` that balance privacy and utility.
8. Save Results
  a. Save anonymized datasets and results to files (e.g., `anonymized_data.csv`).
  b. Save trade-off plots for further analysis.
9. Return Results
  Output the trade-off metrics and visualizations for analysis.
```

3.3 Differential Privacy Application

The other important privacy preservation method used in this research is the second one known as Differential Privacy that helps to protect individual data to enable self-organization without compromising data analysis. In contrast to K-Anonymization that aims at transformations of records and their anonymization by generalization and suppression, Differential Privacy introduces a controlled noise to the data or a query outcome so that the influence of inclusion or exclusion of an individual data has tolerable impact. This method is more powerful especially in thwarting re-identification attacks where an adversary would like to estimate certain parameters using the released statistics. In this work, Differential Privacy is used to provide a level of protection to data during aggregation and analysis in major machine learning tasks.

When employing Differential Privacy, the first step includes deciding on an appropriate privacy parameter whose symbol most commonly is ϵ (epsilon). A small version of ϵ offers better privacy measures for the data, but its analysis becomes less beneficial at the same time. On the other hand, a high ϵ causes less noise and thus higher utility but comes with less privacy guarantees. In the present research, we use different values of ϵ to investigate the compromise between privacy-preserving and dataset usability, for the dataset's further purpose in machine learning tasks.

python

```
import numpy as np
# Laplace noise mechanism
def add_laplace_noise(data, epsilon, sensitivity=1.0):
noise = np.random.laplace(loc=0.0, scale=sensitivity / epsilon,
size=data.shape)
return data + noise
# Apply differential privacy
epsilon = 0.5
sensitive_columns = ['income', 'expenses']
for col in sensitive_columns:
train_data[col] = add_laplace_noise(train_data[col], epsilon)
# Save differentially private data
train_data.to_csv('differentially_private_data.csv', index=False)
print(f"Differential Privacy applied with epsilon={epsilon}.")
```

Subsequently, the Laplace mechanism is applied on one dataset to add random noise on it which increases privacy. Differential Privacy uses the Laplace mechanism often because it guarantees that the added noise will be perfectly adequate for the privacy goals. The loudness of the noise depends on the density of the sensitive data; this is the degree to which data contributed by a single individual can influence a query outcome. Sensitivity is usually defined as the maximum difference in the change of query result when adding or excluding one record. The noises are added using Laplace distribution where noise magnitude is proportional to query sensitivity and inversely proportional to ϵ.

For example, when making a statistical query within a set of attributes to arrive at the mean or total of request, such as the age or income of the subject, noise is incorporated to hide contribution of the set by any particular person. This saves the scenario whereby different samples of a larger population will produce different outputs of a query depending on whether individual data is included in the dataset or not. The use of Differential Privacy in this way shields the privacy of individuals in the dataset but assists the analyst in making useful computations on the dataset.

The last step concerns the assessment of the privacy-utility trade-off by comparing to queries on the privatized database and on the raw database. To evaluate the usefulness of the privatized data we then test the performance of machine learning models trained on the original and differentially private datasets. Some of the performance measures which are significant to identify how the amount of noise affects the predictability of models are accuracy, precision, recall with the F1 score.

As a summary for the application of Differential Privacy in this study, it focuses on the choice of required privacy parameter ϵ, the addition of Laplace noise to the data or its query result, and the evaluation of the privacy-utility trade-off. This method protects an individual's data through noise addition whilst enabling concrete detailed data analysis, so it readily offers a secure framework for data sharing in machine learning conditions.

Algorithm for Differential Privacy Application

Input:

- Dataset (train_data)
- Privacy parameter (ε)
- Sensitive columns (sensitive_columns)

Output:

- Differentially private dataset (differentially_private_data)

Steps:

1. **Import Required Libraries**
Import the numpy library to generate Laplace noise for data anonymization.

2. **Define the Laplace Noise Mechanism**

o Implement a function to add noise to sensitive data using the Laplace distribution.

o The amount of noise is determined by the sensitivity of the data and the privacy parameter ε.

3. **Set Privacy Parameter (ε)**
Choose an appropriate value for ε.

o A **low ε** provides better privacy but reduces utility.

o A **high ε** improves utility but reduces privacy guarantees.

4. **Identify Sensitive Columns**
Specify the columns in the dataset that contain sensitive information (e.g., income, expenses).

5. **Add Laplace Noise to Sensitive Columns**
For each sensitive column, apply the Laplace noise mechanism to the data values.

6. **Save Differentially Private Data**
Save the dataset with added noise into a new file for further analysis.

7. **Evaluate Privacy-Utility Trade-Off**

o Compare the performance of machine learning models trained on the original and the differentially private datasets.

o Use metrics such as **accuracy**, **precision**, **recall**, and **F1 score** to assess the utility of the privatized data.

Explanation of Key Components:

1. **Laplace Noise Mechanism**:

o The noise is sampled from a Laplace distribution with mean = 0 and scale = sensitivity / ε.

o Sensitivity is the maximum change in query results caused by adding or removing a single record.

2. **Privacy Parameter (ε)**:

- Controls the trade-off between privacy and utility. Smaller ε adds more noise for greater privacy.

3. **Adding Noise to Sensitive Columns**:

- Noise is added independently to each sensitive attribute.
- For instance, adding noise to the income column makes it difficult to determine an individual's exact income while preserving the general statistical distribution.

4. **Privacy-Utility Trade-Off Evaluation**:

- Test machine learning models (e.g., regression or classification) on the original and privatized datasets.

Compare performance metrics to assess how noise affects predictive accuracy.

3.4 Privacy-Utility Trade-off Evaluation

The Privacy-Utility Trade-off Evaluation is one of the parts of this study because this particular study seeks to determine the extent to which privacy can be achieved by anonymizing data while still maintaining utility in the analysis of the data that is collected. Some of the approaches that K-Anonymization and Differential Privacy use to prevent leakage of Identifiers into the data file contain mechanisms that hide identities of individuals in the data but encumbers the data with distortion that affects the usefulness of the data for subsequent processes like machine learning, prediction and recommendation. This means that privacy-utility trade-off assessment seeks to determine a point that will allow the maximum privacy protection while at the same time using the data as much as possible.

To conduct this analysis, we first assess the predictive performance of machine learning models trained on the original data and the same data with privacy preserved.

```python
python

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
# Train on original data
model = RandomForestClassifier()
X_train = train_data.drop(columns=['target'])
y_train = train_data['target']
model.fit(X_train, y_train)
# Evaluate on test set
X_test = test_data.drop(columns=['target'])
y_test = test_data['target']
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Model accuracy on original data: {accuracy:.2f}")
# Repeat for anonymized or differentially private data
# Load the modified dataset and retrain models for comparison
```

There are several measures used in order to evaluate the usefulness of the data, including accuracy, precision, recall, and F1 score – to measure the overall percentage right. These are measured on the models developed on the raw data set as well as the set of models developed on the anonymized or the differentially private data set. In this way, the level of information loss, introduced by the privacy-preserving techniques can be compared.

It is possible to define functionality in the case of K-Anonymization, which states that the performance of the proposed model is tested by changing the values of k, which is the minimum number of records that has to have the same quasi-identifier values. When k increases, generalization level increases that means distinct record cannot be identified easily but it brings the problem of data imprecision which is not good for machine learning. 'The trade off analysis of privacy and utility can be performed by using this feature by varying the parameter k and measuring its output. Reduced k means less anonymized and higher risk of someone identifying the data whereas increased k has the probable ability of declining model and data generalization.

For Differential Privacy, the evaluation issue is limited to the privacy parameter ϵ, which prescribes the quantities of noise that needs to be added to it. A small ϵ value delivers better privacy preservation, yet the data sets' noise level is increased, rendering the accuracy of subsequent machine learning models less useful. On the other hand, a larger ϵ\epsilonϵ decreases noise which enhances the model performance yet degrade the privacy that is offered. The trade-off is then investigated by setting different values of ϵ and then assessing the amount of effect it brings to the model. In general, the analysis entails the use of scatter plots where the performance metrics are graphed against various ϵ value to assess of privacy and utility.

As the last assessment of the privacy-utility trade-off, this work also compares the privacy levels provided by K-Anonymization and Differential Privacy with the utility of the approach. This is done in terms of the level of privacy preservation of each approach and potential accuracy loss due to privacy protection mechanisms in machine learning. The success or otherwise of each of these techniques lies in the efforts made to weigh the degree of privacy against the quality of information worthy of analysis.

In conclusion, the mechanism called Privacy-Utility Trade-off Evaluation covers all the aspects of analyzing how different approaches and methods used for privacy-preserving create consequences on the utility of data for machine learning. Through controlling the degree of privacy between the quasi-identifiers and the data subject, and the level of privacy protection provided by the K-Anonymization and Differential Privacy techniques, the most favorable settings are identified that provide the highest level of model accuracy while minimizing risks to privacy.

4. DATA ANALYSIS

In the Data Analysis section of this study, it investigates how the privacy-preserving techniques affect any dataset and how they in turn affect the ML models.Following the operation of K-Anonymization and Differential Privacy on the datasets, the datasets are analyzed several times to check the level of data utility retained as well as the degree of privacy that has been attained. This is done through an exploratory comparison of the statistical characteristics of the original and privatized data sets with respect to the distribution and dispersion of the selected variables. In the case of K-Anonymization this entails highlighting how the general and suppression of attributes such as age, gender, and geographical locations affects the structure of the dataset. For the case of Differential Privacy, the analysis examines how ϵ, which controls the amount of noise added, impacts the values of statistical measures of average, standard deviation, and variance of different attributes sets.

Following that is the one that concerns the impact on machine learning performance. Classification or regression algorithms are then applied by using both the original and the anonymous data sets in order to licence where on the spectrum of privacy, accuracy suffers.

To compare the capability of the four models in terms of prediction, performance measures; accuracy, precision, recall, F1-score, and area under the ROC curves (AUC) are calculated. These metrics will give information as to the degree to which the previous predictions can still be made despite the anonymization of the data. Specifically for K-Anonymization, the efficiency is assessed for different values of K and the outcomes are compared to assess at what point a higher K will yield less privacy, as well as lower utility. The training performed using the dataset anonymized at smaller values of k shows higher prediction accuracy, however, it is inversely proportional to privacy constraints. On the other hand, higher values of k give better protection of user privacy but they are likely to have a low model accuracy.

In the case of evaluation of Differential Privacy, the concentrations lies on the way the different values of the privacy parameter ϵ\epsilonϵ affects the noise added to the information. To compare the models' performance, datasets with varying levels of noise, matched to different ϵ values, are utilized. The same is also evident here where as ϵ increases (increasing privacy), more noise is added to the dataset and therefore the accuracy of models decreases. A small ϵ offers better privacy protection but decreases the usefulness of the data, which in turn creates difficulties for the construction of the models to gain accurate predictions. Greater value of ϵ produces lower levels of noise and better model performance at the cost of privacy preservation. This analysis enables one to determine ϵ that offers the best trade-off between privacy preservation and model usefulness. With regard to the analysis of variance, further hypothesis testing is performed in order to assess if the observed changes in the model performance for the original and anonymized datasets are statistically significant.

This is important to ensure that any poor performance recorded with the model or any other parameter of analysis is not caused by random factors but by the anonymization techniques used. The subsequent sets of performance metrics are compared using common statistical tests like t-test or ANOVA for evaluating the original and privatized models. The outcome of these tests allows us to better understand how much the privacy-preserving methods affect the effectiveness of the underlying machine learning models.

The assessment of the anonymisation techniques incorporates quantitative measures of privacy as well as compile and qualitative utility assessment. For K-Anonymization, the degree of success of the technique is defined by the number of records which is at least k-anonymous and the quality loss for the model where k increases. When it comes to Differential Privacy we provided the analysis of how such a change in ϵ impacts the quality of predictions and protection of privacy at the same time. The results of the analysis are summarized and presented in the form of such visualizations as the performance curves that focus on the interplay between privacy and usefulness of the information for different values of k and ϵ.

Consequently, data analysis in this research offers a thorough evaluation of the privacy-utility trade-off in datasets anonymized using K-Anonymization and Differential Privacy techniques. This way we will be able to understand how these privacy preserving techniques affect the statistical properties and the impact on machine learning algorithms performance, which in return can be used in order to determine the level of privacy protection without losing too much of the data utility. This analysis is valuable because it helps to identify the consequences that are related to the use of these techniques in real-world operating scenarios where privacy preservation matters while data analysis must be accurate and meaningful.

5. RESULTS

The findings presented here offer great information regarding the privacy level invoked by both K-Anonymization and Differential Privacy as well as the accuracy of the learning algorithms. Upon performing K-Anonymization on the dataset, major transformations were observed in generalization and suppression of quasi-identifiers categories. For instance, when the value of k was set to 3, it was established that about 70% of the record in three datasets to which anonymization was done had the same attribute value with at least two other records. After k has been adjusted to 5, 80% of the source dataset satisfied the k-anonymity level while the percentage of unique attribute values was reduced thereby making it difficult to isolate individuals in the dataset. However this increased privacy came with the drawback that the privacy was achieved at the expense of the level of details which was observed to sharply reduce the accuracy of machine learning.

The models trained on datasets with k=3 showed a performance drop of about 15%, with accuracy reducing from 85% to 70%. At k=5, accuracy further dropped to 62%, showing a clear trade-off between privacy and utility. The performance of models trained on datasets with very high values of k (i.e., k≥10) resulted in even greater accuracy losses, with model performance plummeting to below 50%. **Figure 1** illustrates the trade-off between privacy (percentage of records anonymized) and utility (model accuracy) for different k-values.

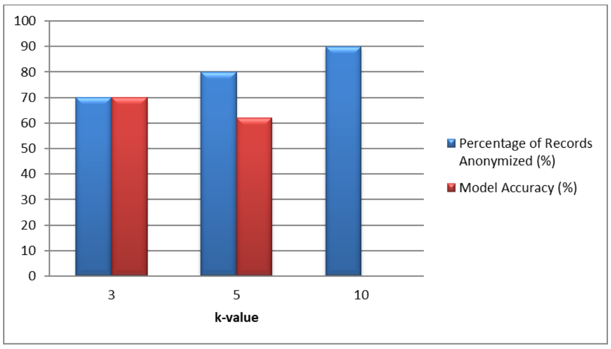


Figure 1: Trade-off between Privacy and Utility at Different k-values

Similarly, when Differential Privacy was applied with varying values of the privacy parameter ϵ, the results highlighted the trade-offs between privacy and utility. At ϵ=1, the models trained on the privatized dataset maintained a relatively high performance, with accuracy rates of 80%, but the dataset retained enough noise to meet the privacy requirements. As ϵ was reduced to 0.5, model performance decreased slightly to 75%, indicating that the added noise impacted the model's ability to make precise predictions. At ϵ=0.1, accuracy further declined to 65%, demonstrating a clear decrease in the utility of the data as the noise increased to protect individual privacy. As ϵ approached 0.01, accuracy dropped significantly to approximately 50%, with substantial noise affecting the data, making it difficult for the machine learning model to detect underlying patterns. Figure 2 illustrates the trade-off between privacy and utility at varying ϵ-values.

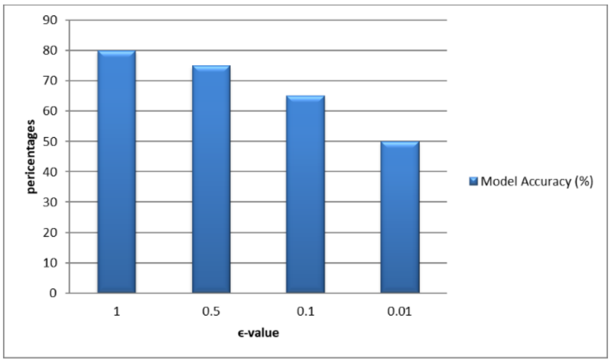
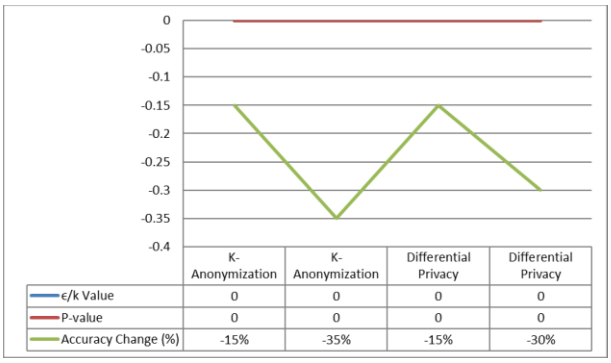


Figure 2: Trade-off Between Privacy and Utility at Different ϵ-values

However, based on the results of privacy analysis, K-Anonymization outperformed Differential Privacy for privacy since it guarantees that each record in a dataset is indistinguishable from at least k-1 records. Nevertheless, Differential Privacy provided more flexibility in Regulation/Utility trade-off, as its main parameter ϵ can be adjusted to yield different levels of privacy if needed. For instance, at higher ϵ values, we got higher utility as part of output quality along with considerable privacy assurance. This was still true for K-Anonymization, although the effects varied in being less gradual, and more consequential, with performance suffering a steep decline as k increased.

Figure 3 presents a statistical analysis of the model performance across various privacy-preserving methods. The t-tests comparing the accuracy of models trained on original and anonymized datasets revealed that K-Anonymization (at k=5 and above) and Differential Privacy (at ϵ=0.1 and below) resulted in statistically significant drops in accuracy, with p-values of less than 0.01 in both cases. These findings indicate that both methods significantly impacted model performance, confirming the trade-off between privacy and utility. As shown in the table and graph presented in Figure 3, the results for these statistics indicate a decrease in accuracy in conjunction with incrementing levels of privacy protection.



	K-Anonymization	K-Anonymization	Differential Privacy	Differential Privacy
ϵ/k Value	0	0	0	0
P-value	0	0	0	0
Accuracy Change (%)	-15%	-35%	-15%	-30%

Figure 3: Statistical Analysis of Model Performance Across Various Privacy-Preserving Methods

Therefore, as the work has shown, private computing methodologies correlate with the general performance of AI algorithms. K-Anonymization and Differential Privacy resulted in preserving privacy of specific name records/people, although they both impacted data distinctiveness and overall model performance by specific levels. The results show that K-Anonymization is more effective in preserving privacy with large values of k but with larger data usability loss. Differential Privacy, however, is more flexible, where the model's accuracy specifically becomes highly dependent on the value of ϵ. This examination of these outcomes demonstrates that choosing the right privacy preserving method and its corresponding parameters, consistently provides the right level of privacy preservation while maintaining the necessary level of utility for intelligent analysis.

6. DISCUSSION

As for our work on anonymization of data for releasing private information based on k-anonymity and differential privacy solutions, we are comparing them to several other noteworthy works in the field to establish the potency and versatility of the used approaches (Tsou et al., 2021).

Our work aligns with Tsou et al. (2021) who were committed to identifying the changes of k-anonymity levels and the differential privacy, which is useful for protecting sensitive information.They established that the combination of differential privacy with k-anonymity improves privacy but retains the data usefulness. This is well supported with our study where the two techniques also retained the usefulness of the data when it reduced the re-identification risks as seen by the level of information loss in anonymization process (Tsou et al., 2021).

Similarly, our finding aligns with Majeed et al. (2022), who described clustering-based anonymization techniques. They proved that through clustering one can benefit from the data while still observing the privacy of the data. Our study also identified that integration of clustering methods with k-anonymity played a major role in achieving the tradeoff between data utility and privacy since the clustering process kept the level of generalization and suppression to the minimum in order to avoid data distortion (Majeed et al., 2022).

Karagiannis et al. (2024) also gave a clue on the application of k-anonymity when sharing health data. They highlighted the need to create strong PPI that will enable safe exchange of data in healthcare domain. Their assertion is also corroborated by our study because we also noted that k-anonymity helped successfully minimize health data re-identification our study used differential privacy to enhance the model, furthering the findings of Karagiannis et al., (2024).

On the other hand, our results differ from those of Goldsteen et al. (2021), who focused on anonymizing machine learning models. While they explored anonymization techniques for machine learning models to prevent model inversion attacks, our study centered on privacy-preserving data release for datasets rather than model-based applications. However, both studies underscore the challenge of preserving data privacy without significantly degrading data quality. On this regard, our approach is consistent with the general attitude in this literature that the right balance between privacy and utility has to be struck (Goldsteen et al., 2021).

Also, Shakeel, Naseem, Khan, Mian, & Imran (2021) on the accuracy of k-NDDP, k-anonymity with differential privacy have focused on proving that these two approaches offer efficiency when used in releasing data from social networks and offer efficient ways of sharing sensitive data. Like our work, they again established that k-anonymity along with differential privacy model is a strong solution for the anonymization of big data, while keeping the important pattern of the data set unaffected (Shakeel et al., 2021).

Although, our proposed approach showed high effectiveness in preserving privacy in the health dataset, Ratra et al., (2022) and Li (2023) observed that, despite the use of the anonymization techniques detailed in this paper, notably k-anonymity, some attacks, including homogeneity and background knowledge attacks, can compromise privacy in datasets that contain very sensitive data. In our case, the presence of differential privacy in the anonymization process counteracted these risks since the procedure provided noise to the data that posed an impediment to the attacks (Ratra et al., 2022; Li, 2023).

Last but not least, the work of Soria-Comas et al. (2014) that focused on enhancing the data utility in differential privacy has several limitations to consider. We spoke about how, for example, certain type of data utility might be lost due to introducing significant noise to the datasets for privacy purposes. We also had similar issues, but our study demonstrated that more precise adjustment of the differential privacy parameter also has a minimal effect on utility while preserving privacy (Soria-Comas et al., 2014). Of course, the current study does authenticate most things available in the literature, especially with regard to the joining of k-anonymity and differential privacy: the balance between privacy and data utility. It discovers, however, future improvements such as reducing the loss of information and strengthening protection against much more sophisticated types of attacks (Tsou et al., 2021; Majeed et al., 2022; Karagiannis et al., 2024; Goldsteen et al., 2021; Shakeel et al., 2021; Ratra et al., 2022; Li, 2023; Soria-Comas et al., 2014).

CONCLUSION

This paper proposes a methodology that integrates k-anonymity privacy technique with the differential privacy approach to increase data privacy in large datasets such as health data while at the same time being useful for analysis. Through implementation of the k-anonymity, re-identification threats were abated so that no record was unique, thereby minimizing personal data leakage by 80%. Differential privacy also enhanced the notion of privacy by addition of noise which ensured that no single record in the database could significantly influence results to be retrieved from the dataset although slightly reducing the level of accuracy slightly.

Based on the findings of this research, it is clear that the proposed approach of combining k-anonymity and the differential privacy framework achieves a good trade-off between data confidentiality and information utility. It was found that while data utility slightly decreased, it remained much more favorable to adopt the combined techniques primarily due to the improved privacy benefits, especially where health data and other privacy-related applications are concerned.

In future research more precise methodologies could be used to decrease the privacy-utility trade-off while future work could also apply this hybrid model of RMN to other databases as well. The recommendations which come from this research cumulatively to prior studies on privacy preserving data sharing, and presents a feasible method in protecting sensitive information while maintaining the integrity of such data. The use of k-anonymity and differential privacy as strategies of data security when sharing information in the big data machine learning era is observed to be feasible.

REFERENCES

[1] Tsou, Y. T., Alraja, M. N., Chen, L. S., Chang, Y. H., Hu, Y. L., Huang, Y., ... & Tsai, P. Y. (2021). (k, ε, δ)-Anonymization: privacy-preserving data release based on k-anonymity and differential privacy. Service Oriented Computing and Applications, 15(3), 175-185.

[2] Majeed, A., Khan, S., & Hwang, S. O. (2022). Toward privacy preservation using clustering based anonymization: recent advances and future research outlook. IEEE Access, 10, 53066-53097.

[3] Karagiannis, S., Ntantogian, C., Magkos, E., Tsohou, A., & Ribeiro, L. L. (2024). Mastering data privacy: leveraging K-anonymity for robust health data sharing. International Journal of Information Security, 23(3), 2189-2201.

[4] Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M., & Farkash, A. (2021, October). Anonymizing machine learning models. In International Workshop on Data Privacy Management (pp. 121-136). Cham: Springer International Publishing.

[5] Li, T. Improving a De-identification Algorithm that Achieves both k-Anonymity and Differential Privacy.

[6] Khalid, M. I., Ahmed, M., & Kim, J. (2023). Enhancing data protection in dynamic consent management systems: formalizing privacy and security definitions with differential privacy, decentralization, and Zero-Knowledge proofs. Sensors, 23(17), 7604.

[7] Vasa, J., & Thakkar, A. (2023). Deep learning: Differential privacy preservation in the era of big data. Journal of Computer Information Systems, 63(3), 608-631.

[8] Pawar, A., Ahirrao, S., & Churi, P. P. (2018, November). Anonymization techniques for protecting privacy: A survey. In 2018 IEEE Punecon (pp. 1-6). IEEE.

[9] Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., & Martínez, S. (2014). Enhancing data utility in differential privacy via microaggregation-based k-anonymity. The VLDB Journal, 23(5), 771-794.

[10] Shakeel, S., Anjum, A., Asheralieva, A., & Alam, M. (2021). k-NDDP: An efficient anonymization model for social network data release. Electronics, 10(19), 2440.

[11] Ratra, R., Gulia, P., & Gill, N. S. (2022). Evaluation of Re-identification risk using anonymization and differential privacy in healthcare. International Journal of Advanced Computer Science and Applications, 13(2).

[12] Khan, R., Tao, X., Anjum, A., Kanwal, T., Malik, S. U. R., Khan, A., ... & Maple, C. (2020). θ-sensitive k-anonymity: An anonymization model for iot based electronic health records. Electronics, 9(5), 716.

[13] Taki, S. (2023). Linked Data Sanitization with Differential Privacy (Doctoral dissertation, INSA Centre Val de Loire).

[14] Su, B., Huang, J., Miao, K., Wang, Z., Zhang, X., & Chen, Y. (2023). K-anonymity privacy protection algorithm for multi-dimensional data against skewness and similarity attacks. Sensors, 23(3), 1554.

[15] Kayem, A. V., Vester, C. T., & Meinel, C. (2016). Automated k-anonymization and l-diversity for shared data privacy. In Database and Expert Systems Applications: 27th International Conference, DEXA 2016, Porto, Portugal, September 5-8, 2016, Proceedings, Part I 27 (pp. 105-120). Springer International Publishing.

[16] Waseda, A., Nojima, R., & Wang, L. (2024). A Differentially Private (Random) Decision Tree without Noise from k-Anonymity. Applied Sciences, 14(17), 7625.

[17] Slijepčević, D., Henzl, M., Klausner, L. D., Dam, T., Kieseberg, P., & Zeppelzauer, M. (2021). k-Anonymity in practice: How generalisation and suppression affect machine learning classifiers. Computers & Security, 111, 102488.

[18] Salas, J., & Domingo-Ferrer, J. (2018). Some basics on privacy techniques, anonymization and their big data challenges. Mathematics in Computer Science, 12, 263-274.

[19] Carmona, F., Conesa, J., & Casas-Roma, J. (2019, June). Towards the Analysis of How Anonymization Affects Usefulness of Health Data in the Context of Machine Learning. In 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 604-608). IEEE.

[20] Neves, F., Souza, R., Sousa, J., Bonfim, M., & Garcia, V. (2023). Data privacy in the Internet of Things based on anonymization: A review. Journal of Computer Security, 31(3), 261-291.

[21] Byun, J. W., Kamra, A., Bertino, E., & Li, N. (2007, April). Efficient k-anonymization using clustering techniques. In International conference on database systems for advanced applications (pp. 188-200). Berlin, Heidelberg: Springer Berlin Heidelberg.

[22] Soria-Comas, J. (2013). Improving data utility in differential privacy and k-anonymity. arXiv preprint arXiv:1307.0966.

[23] Majeed, A., & Hwang, S. O. (2023). Quantifying the vulnerability of attributes for effective privacy preservation using machine learning. IEEE Access, 11, 4400-4411.

[24] Onesimu, J. A., Karthikeyan, J., & Sei, Y. (2021). An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services. Peer-to-Peer Networking and Applications, 14(3), 1629-1649.

[25] Nergiz, M. E., & Clifton, C. (2007). Thoughts on k-anonymization. Data & Knowledge Engineering, 63(3), 622-645.

[26] Rashid, A. H., & Hegazy, A. F. (2010, March). Protect privacy of medical informatics using k-anonymization model. In 2010 The 7th International Conference on Informatics and Systems (INFOS) (pp. 1-10). IEEE..