International Journal of Innovative Research and Scientific Studies, 8(1) 2025, pages: 1947-1963

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Privacy-preserving sensitive data publishing in case of outliners challenges

Supornpol Nukrongsin[1*], Chetneti Srisa-An[2]

[1,2]*College of Digital Innovation Technology, Rangsit University, Pathumthani 12000, Thailand.*

Corresponding author: Supornpol Nukrongsin (*Email: supornpol.n65@rsu.ac.th*)

Abstract

The purpose of keeping personal data is to prevent malicious individuals from hacking or violating privacy in order to extort or take advantage of the data owner or the person in charge of the data. Outliers present significant challenges in machine learning and data security due to their substantial deviation from the norm, making them attractive targets for potential security breaches. Outliers attract the attention of hackers due to their potential to reveal personally sensitive information. This paper addresses the outlier problem within data privacy concerns, focusing specifically on personal data that includes outliers. A novel framework, namely Handling Outlier Anomaly Privacy Violation (HOPV), designed for implementation on data controllers' web servers, is proposed in order to monitor and mitigate this issue. The HOPV framework incorporates an advanced outlier detection algorithm complemented by sophisticated data generalization and Laplace Mechanism Perturbation techniques. Empirical results elucidate the remarkable performance superiority of our software module when juxtaposed with existing products in the field.

Keywords: Data anonymization, Data privacy, Data security, Outliers, Privacy-preserving, Violation.

DOI: 10.53894/ijirss.v8i1.4839
Funding: This study received no specific financial support.
History: Received: 8 January 2025/**Revised:** 10 February 2025/**Accepted:** 13 February 2025/**Published:** 21 February 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: Both authors contributed equally to the conception and design of the study. Both authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

The challenge posed by outliers extends beyond machine learning and data mining, emerging as a recent concern in the realm of data privacy. This concern arises when specific data points deviate significantly from their respective groups, attracting the attention of hackers due to their potential to reveal personally sensitive information. Hackers exploit these outlier patterns through data mining, a technique extensively employed by both hackers and governments for identifying outliers [1]. Public attention to these concerns was heightened with the announcement of the Data-Mining Moratorium Act in 2003 [2]. This legislation prohibits all data mining programs related to Defense or Homeland Security until Congress reviews the Terrorism Information Awareness Program. Data privacy is a widely discussed topic in the data science research

field. A degree of data utility is inevitable when a data controller processes data privacy in their database before sending. Data Anonymization Techniques include data masking, generalization, data swapping, and data perturbation.

Generalization and perturbation have been the most famous techniques among their colleagues for years. This paper has chosen the best algorithm from both techniques. Both techniques have their benefits. For example, data perturbation works well on all numeric attributes, while data generalization techniques can cope with categorical attributes. Most cases need both techniques in order to protect personal data.

The data generalization techniques [3] were introduced by Sweeney [3]. It was found that a lot of personal data can be revealed with a piece of auxiliary information. This finding has been named a “re-identification attack” since then. The data generalization techniques add noise into some quasi-identifier fields such as age, gender, and zip code. This technique can prevent re-identification attacks, but it might cause a degree of data utility. It was also found that much personal data can be revealed with auxiliary information. Therefore, this finding has been named a “re-identification attack.” The data generalization technique adds noise to some quasi-identifier fields, such as age, gender, and zip code. This technique can prevent re-identification attacks but might cause a degree of data utility.

The author's work focuses directly on its sensitive data attributes by adding noise, while data generalization techniques tend to modify quasi-identifier attributes to hide their sensitive data.

1.1. Data Privacy Laws

Starting in 2009, the introduction of data privacy laws coincided with the announcement of the General Data Protection Regulation (GDPR) for all European countries. In response, the Thai government recognized the need to align with GDPR standards and subsequently introduced the Personal Data Protection Act (PDPA) laws.

The primary objective of data privacy laws is to safeguard all personal data, prohibiting its exposure to the public without explicit consent. Personal data is divided into two categories: Personally Identifiable Information (PII) and Sensitive Personal Data (SPD). SPD encompasses information that organizations are obligated to protect, such as salary details, tax information, medical records, and more. Importantly, Thailand's PDPA laws extend this protective scope to encompass a broader array of data, including political views and ethnicity.

Within the framework of the PDPA, two critical terminologies are introduced: Data Controllers (DC) and Data Processors (DP). Data Controllers are organizational units that bear full responsibility for decisions regarding the storage, usage, and publication of personal data. Currently, data processors execute operations in accordance with the directives issued by data controllers. In Thailand, these roles are often assumed by the same organizational unit.

Regrettably, data controllers may be vulnerable to data privacy violations, potentially facing substantial penalties, particularly in the aftermath of a data breach. The fundamental purpose of data privacy laws is to protect personal data from public exposure without explicit consent. Compliance with these laws is imperative for organizations operating within the regulatory framework.

According to Thailand's PDPA law, if personal data is leaked, the data controller of that organization must notify the data owner within 72 hours and must fix the data breach problem as soon as possible. Therefore, a data breach is one of the most serious problems in data privacy and data security in Thailand.

The PDPA introduced two novel terms: Data Controllers (DC) and Data Processors (DP). Data controllers are organizational units responsible for decisions regarding the storage, usage, and publication of personal data.

Conversely, data processors are organizational units responsible for processing data in accordance with the instructions of data controllers. It is noteworthy that in Thailand, these roles are often fulfilled by the same organizational unit.

However, under this definition, data controllers are at risk of becoming victims of data privacy violations. This vulnerability is particularly acute in the event of a data breach, where substantial penalties may be imposed on data controllers. In this context, the malefactors or adversaries are hackers employing various attack techniques to access personal data, ultimately leading to a data breach.

1.2. History of Data Breach

In 2018, the New York City Taxi and Limousine Commission released many taxicab data, including basic details like the location and time of pickups and drops. As a result, attackers can reidentify the driver's personal information, such as address and salary. AOL data breaches are the same type of incident.

The state of New York announced open data for taxi trip records in the New York area. The original idea was to cope with a traffic jam problem by providing information to open users. The anonymized data source is located at https://data.cityofnewyork.us. Unfortunately, attackers can reidentify the driver's personal information, such as address and salary.

Another incident that occurred around the same time was the AOL case. The release of anonymized AOL data raised concerns when the New York Times managed to identify an AOL searcher partially. User No. 4417749 [4] was revealed to the public with her real name and other sensitive personal information.

The National Institutes of Health (NIH) released a number of different statistics that included minor allele frequencies as well as various chi-squared statistics and known p-values for various tests, and these were things that they freely released before. Homer, et al. [5] showed that under certain technical conditions, it was possible to reidentify individuals who participated in a study. This attack was discovered by Homer, et al. [5] As a result, the NIH restricted free access to scientific information since then. It is possible to conduct these genomic studies or compute these statistics in an appropriately privacy-preserving manner.

There are many real case studies listed above, including AOL, New York City Taxi, and the National Institutes of Health (NIH). Data privacy is a widely discussed topic in the data science research field. The method for those case studies is to add noise to a dataset. This technique is well known as "data perturbation." Many techniques were invented for the data perturbation method. They all try to add noise as little as possible to preserve data utility.

Table 1.
History of data breach.

Case	Year, reference
AOL (Anonymized Internet Search Data)	2006,Barbaro and Zeller [4]
Netflix Prize contest data breach	2018,Lohr [6]
NIH Federal Credit Union sent out data breach letters to data owners	2003, Console and Associates [7]
Re-identification attacks on Massachusetts Governor William Weld's medical data	1997, Barth-Jones [8]
Yahoo's database was hacked, compromising 3 billion user accounts, including personally identifiable information (PII)	2013,Perlroth [9]
Census Bureau data from "reconstruction attacks"	2023Majeed and Lee [10]

Table 1 presents a history of personal data breach cases from 1997 to 2023.

1.3. The Outlier Case

In the Titanic tragedy, two individuals purchased ticket number 17755, priced at 512 pounds in 1912, while the regular ticket cost was only 7 pounds. Figure 1 illustrates that these two data points are outliers. Importantly, these outliers cannot be dismissed as human errors; instead, they represent genuine data. We refer to this scenario as "sensitive personal data containing outliers."

```
Age_x1_male = df['Age'][df['Sex']=='male']
Age_x2_female = df['Age'][df['Sex']=='female']
Fare_y1_male = df['Fare'][df['Sex']=='male']
Fare_y2_female = df['Fare'][df['Sex']=='female']
```

```
plt.figure(figsize=(15,6))
plt.scatter(Age_x1_male, Fare_y1_male, color='b')
plt.scatter(Age_x2_female, Fare_y2_female, color='r')
plt.show()
```

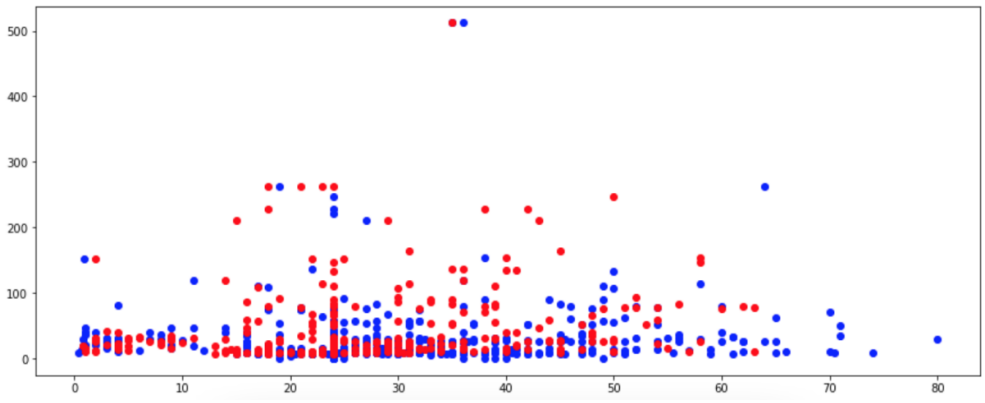

Figure 1.
Distribution of ticket prices.

Figure 1 illustrates the distribution of ticket prices. The Titanic case has been selected as a dataset for this paper due to the presence of outlier data within it. According to Thailand's PDPA laws, the personal data of deceased individuals may not be protected; however, this dataset serves as an illustrative example stemming from the analysis of outliers.

```
from scipy.stats import zscore
titanic["fare_zscore"] = zscore(titanic["fare"])
titanic["is_outlier"] = titanic["fare_zscore"].apply(
  lambda x: x <= -2.5 or x >= 2.5
)
titanic[titanic["is_outlier"]]
```

survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone	age_zscore	is_outlier	fare_zscore
1	1	male	36.0	0	1	512.3292	C	First	man	True	B	Cherbourg	yes	False	0.024111	True	5.681797
1	1	male	35.0	0	0	512.3292	C	First	man	True	B	Cherbourg	yes	True	-0.039875	True	5.681797

Figure 2.
Outlier detection by Z-score.

Figure 2 illustrates the use of the Z-score to detect outliers. While the Z-score is effective for identifying outliers, it does not perform clustering.

1.4. The IBM Watson Marketing Data Analysis Case

This experiment uses a dataset called the IBM Watson Marketing Data Analysis and Prediction dataset. This step is crucial as it involves detecting outliers using DBSCAN and the Z-score, as illustrated in Figure 6. Identifying outliers at this stage is essential. If no outliers are found, the program will exit.

```
# -1 value in outliners_flag column signifies that the data is outliner
dbscan = DBSCAN(eps = 5,min_samples = 4)
model = dbscan.fit(X)
labels = model.labels_
df["outliners_flag"] = labels
anomalies = df[df.outliners_flag == -1]
anomalies
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	outliners_flag
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	-1
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	-1
31	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	28.0	1	0	PC 17569	146.5208	B78	C	-1
54	55	0	1	Ostby, Mr. Engelhart Cornelius	male	65.0	0	1	113509	61.9792	B30	C	-1
92	93	0	1	Chaffee, Mr. Herbert Fuller	male	46.0	1	0	W.E.P. 5734	61.1750	E31	S	-1
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 3.
Outliner detection.

Figure 3 illustrates the classification of outliers using the outlier flag, as determined by DBSCAN, which assigns -1 values for outliers. The dataset contains a total of 74 outliers. However, it is important to note that this paper focuses on the detection of outliers rather than their quantity.

In this research paper, we have selected DBSCAN Deng [11] and Z-Score as our outlier detection tools due to their status as recent inventions and their availability in sklearn tools.

1.5. Differencing and Reidentification Attack

The differencing attack is one of the most serious threats to data privacy. A well-known case is that of the U.S. Census Bureau. The U.S. Census Bureau conducted a national Census of Population and Housing in 2010. Unfortunately, database differencing attacks occurred, leading to personal data breaches. Due to the large volume of data generated by the surveys, the associated risks are significantly heightened. A Laplace Mechanism Perturbation is applied to prevent these attacks in this research. For example, hackers query only the individual they want to target and then query entire datasets, excluding the records of the individual they wish to attack. This pattern occurred during the national Census of Population and Housing in 2010. Reidentification attacks happen when there is only one row in the group by function, as shown in Figure 1.

```
df.groupby(['State','Gender','Education','Policy','Response','
```

	State	Gender	Education	Policy	Response	EmploymentStatus	count
0	Arizona	F	Bachelor	Corporate L1	No	Employed	5
1	Arizona	F	Bachelor	Corporate L1	No	Unemployed	2
2	Arizona	F	Bachelor	Corporate L1	Yes	Unemployed	1
3	Arizona	F	Bachelor	Corporate L2	No	Employed	8
4	Arizona	F	Bachelor	Corporate L2	No	Unemployed	5
5	Arizona	F	Bachelor	Corporate L3	No	Disabled	1
6	Arizona	F	Bachelor	Corporate L3	No	Employed	14
7	Arizona	F	Bachelor	Corporate L3	No	Medical Leave	2
8	Arizona	F	Bachelor	Corporate L3	No	Unemployed	6
9	Arizona	F	Bachelor	Corporate L3	Yes	Disabled	1
10	Arizona	F	Bachelor	Corporate L3	Yes	Employed	2
11	Arizona	F	Bachelor	Corporate L3	Yes	Retired	1
12	Arizona	F	Bachelor	Corporate L3	Yes	Unemployed	2
13	Arizona	F	Bachelor	Personal L1	No	Disabled	2
14	Arizona	F	Bachelor	Personal L1	No	Employed	25

Figure 4.
Risk records.

Figure 4 shows all records at risk from differencing attacks and reidentification attacks. For example, Figure 1 illustrates one risk row in line 11 where the retired female customer lived in Arizona, holding Corporate L3, and her educational background is a bachelor's degree. With additional knowledge from other databases, if there is only one person in a database, then this row can identify the owner of this record. This technique is called a "reidentification attack." According to Sweeney [3], "State," "Gender," "Education," "Policy," "Response," and "Employment Status" attributes become "quasi attributes." This problem can be solved using a generalization process, as shown in Figure 8.

```
df['EmploymentStatus'] = df['EmploymentStatus'].replace('Disabled','Unemployed')
df['EmploymentStatus'] = df['EmploymentStatus'].replace('Retired','Unemployed')
df['EmploymentStatus'] = df['EmploymentStatus'].replace('Medical Leave','Unemployed')
```

```
df.groupby(['State','Gender','Education','Policy','Response','EmploymentStatus']).siz
```

	State	Gender	Education	Policy	Response	EmploymentStatus	count
0	Arizona	F	Bachelor	Corporate L1	No	Employed	5
1	Arizona	F	Bachelor	Corporate L1	No	Unemployed	2
2	Arizona	F	Bachelor	Corporate L1	Yes	Unemployed	1
3	Arizona	F	Bachelor	Corporate L2	No	Employed	8
4	Arizona	F	Bachelor	Corporate L2	No	Unemployed	5
5	Arizona	F	Bachelor	Corporate L3	No	Employed	14
6	Arizona	F	Bachelor	Corporate L3	No	Unemployed	9
7	Arizona	F	Bachelor	Corporate L3	Yes	Employed	2
8	Arizona	F	Bachelor	Corporate L3	Yes	Unemployed	4
9	Arizona	F	Bachelor	Personal L1	No	Employed	25
10	Arizona	F	Bachelor	Personal L1	No	Unemployed	9
11	Arizona	F	Bachelor	Personal L1	Yes	Employed	3

Figure 5.
Generalized attributes.

Figure 5 shows that three categories (Disabled, Retired, Medical Leave) are regrouped into an unemployed status. This technique is called "generalization." By using this technique, some fields are regrouped; however, their data utility is degraded, as shown in section 5.

1.6. Laplace Mechanism Perturbation Technique

The Laplace mechanism perturbation technique was first introduced by Dwork and Roth [12].

Equation 1: The Laplace Distribution with scale b and a location parameter μ is the distribution with a probability density function:

$$f(x|\mu, b) = \frac{1}{2b} exp^{(-\frac{|x-\mu|}{b})} \quad (1)$$

Equation 1 shows the Laplace distribution, a symmetric form of the exponential distribution. The Laplace Mechanism perturbation technique uses its properties to protect data privacy because its distribution affects datasets very little, as shown in Figure 3.

```
def laplaceMechanism(x, epsilon):
    x +=  np.random.laplace(0, 1.0/epsilon, 1)[0]
    return x
```

```
df['Noised Income'] = df['Income'].apply(laplaceMechanism,
df1=df
col=['Gender','EmploymentStatus','Income','Noised Income']
df1=df1[col]
df1=df1.dropna()
df1.head(10)
```

	Gender	EmploymentStatus	Income	Noised Income
0	F	Employed	56272.795197	56272.610909
1	F	Unemployed	0.562929	0.280897
2	F	Employed	48765.968356	48765.940848
3	M	Unemployed	0.373316	3.012610
4	M	Employed	43835.448111	43837.172410
5	F	Employed	62900.545246	62899.522933

Figure 6.
Laplace mechanism perturbation.

Figure 6 shows how the added noise affects the original income so little, but it causes attackers to hesitate to attack. The Laplace mechanism applies the Laplace distribution by computing 1/f as noise added to the dataset. The value of noise added from using the Laplace distribution property is to perturb each row. The amount of noise applied to each row is called the sensitivity of f. The more noise there is, the less data utility (more loss).

Equation 2 (ε-Laplace-based data perturbation): The Laplace Mechanism is such that a function f(x) returns a number, then F(x) satisfies ε-Laplace-based data perturbation.

$$F(x) = f(x) + Lap(\frac{s}{\varepsilon}) \quad (2)$$

Where S is the sensitivity, and Lap(S) denotes sampling from the Laplace distribution.

Our platform can prevent data breaches unintentionally by using unpublished datasets and intentional hacking techniques such as re-identification and differencing techniques. In these studies, a researcher intends to compare Laplace-based data perturbation and data generalization techniques to determine their strong benefits. Firstly, we aim to compare the effect of data utility on machine learning applications. Secondly, they should be applied to their strong characteristics. Lastly, we will compose a new dataset.

1.7. Receiver Operating Characteristic (ROC) Curve

Many research papers by Feingold, et al. [2]; Kara and Eyupoglu [13] and Narayanan and Shmatikov [14] suggest that the generalization process inevitably causes the deterioration of data utility. We will present them using ROC and AUC for ease of presentation.

The ROC Curve is a line used to measure the performance of a binary classification model. ROC measures how accurately it can predict the issues of interest. AUC stands for Area Under the ROC Curve. ROC and AUC are used in this study to represent other metrics to show the decline in data utility over the processing generalization of the K-Anomaly algorithm.

In this research, ROC and AUC are used to measure the efficiency of the classification model because they encompass many metrics. Using ROC and AUC is better than other metrics because not only is it a visualization model, but it is also a number that can be compared more clearly.

The paper is structured into six sections. Section 1 introduces the problem statement and provides an overview. Section 2 presents the literature review. Section 3 explains the data preparation process, while Section 4 details the methodology. The experiment and its results are discussed in Section 5, and finally, Section 6 concludes the study.

2. Literature Review

Dinur and Nissim [15] proposed a new theory called a polynomial reconstruction algorithm in 2004, which is a fundamental Laplace-based data perturbation algorithm.

Dwork and Roth [12] introduced the concept of Laplace-based data perturbation to solve a differencing attack. The main idea is to add small noise to preserve data privacy and utility. Laplace-based data perturbation is one of the perturbation techniques.

Kara and Eyupoglu [13] proposed their algorithm for privacy-preserving data publishing using graph and tabular data techniques.

Kara, et al. [16] developed an e-health dataset using various data generalization technique-based algorithms. They also stated that using the data generalization techniques concept to create an e-health dataset is inevitable.

Narayanan and Shmatikov [14] revealed that removing names alone does not ensure the privacy of the data, so they released an anonymized aggregate heat map of location data from across the world. It was found that they had released sensitive data about the military.

Chen, et al. [17] demonstrated that removing names alone does not ensure the privacy of the data, so they released an anonymized aggregate heat map of location data from across the world. It was found that they had released sensitive data about the military.

Narayanan and Shmatikov [14] demonstrated a differencing attack that targets aggregate statistics like summary statistics, histograms, and charts. The differencing attack works by singling out an individual from multiple aggregate statistics.

Narayan and Haeberlen [18] propose a technique called DJoin that enables differentially private join queries over distributed databases. DJoin utilizes this framework to protect the privacy of individual records during the joint operation.

Li and Miklau [19] present the design and evaluation of the adaptive mechanism. It discusses the theoretical foundations and practical considerations of balancing accuracy and privacy. The authors provide experimental results on real-world datasets to demonstrate the effectiveness of their approach in achieving accurate query answers under Laplace-based data perturbation.

Wilson and Rosen [20] stated that database administrators must balance data utility and privacy because businesses need to analyze data while complying with laws.

The Information Commissioner's Office [21] introduced privacy-enhancing technologies (PETs) as a framework for the software industry to develop tools to enhance privacy and data protection within the data industry in 2002. An example of PETs is Cloud Link TEE, which has been used by Alibaba Tech since 2019.

Gionis and Tassa [22] presented the concept of minimal loss of information from k-anonymization caused by the process of generalization.

In 1996, Ester, et al. [23] introduced the density-based spatial clustering of applications with noise (DBSCAN) algorithm, which is designed for unsupervised density-based clustering. Recognized as one of the premier outlier detection methods, DBSCAN stands out for its ability to detect outliers and, importantly, disregard them. It is readily available in the sklearn library, which enhances its accessibility and usability. The distinctive advantage of DBSCAN, compared to algorithms like K-means, is its ability to identify and subsequently disregard outliers. In this research, DBSCAN is employed for outlier detection due to its capabilities. The algorithm assigns labels to clusters, and in instances where a label cannot be assigned, it designates the data point as an outlier with a value of -1.

Another noteworthy algorithm implemented in our experiments is the isolated forest (IF), introduced by Liu, et al. [24].

3. Data Preparation

The dataset for an experiment is chosen and prepared in this section. The IBM Watson Marketing Data Analysis & Prediction dataset is selected because it consists of numeric data for data perturbation and categorical data for data generalization techniques. The dataset URL is https://www.kaggle.com/code/pankajjsh06/ibm-watson-marketing-data-analysis-prediction.

The data of the IBM Watson Marketing Data Analysis & Prediction dataset consists of a total of 24 field columns and has only two data types: numeric data type and categorical data, as shown in the data dictionary in Table 2.

Table 2.
Data dictionary.

Attribute	Type
Customer	Categorical data
State	Categorical data
Customer lifetime value	Numerical data
Response	Categorical data
Coverage	Categorical data
Education	Categorical data
Effective to date	Categorical data
Employment status	Categorical data
Gender	Categorical data
Income	Numerical data
Location code	Categorical data
Marital status	Categorical data
Monthly premium auto	Numerical data
Months since the last claim	Numerical data
Months since policy inception	Numerical data

Attribute	Type
Number of open complaints	Numerical data
Number of policies	Numerical data
Policy type	Categorical data
Policy	Categorical data
Renew offer type	Categorical data
Sales channel	Categorical data
Total claim amount	Numerical data
Vehicle class	Categorical data
Vehicle size	Categorical data

This paper demonstrates the effect of applying noise in machine learning analytics since this dataset contains a lot of sensitive personal data, such as responses, income, and total claim amounts.

3.1. Step 1: Classify Data Type

Laplace-based data perturbation works well only with numeric data; therefore, response fields need to be converted into numbers using the Label Encoder function.

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score

df=pd.read_csv('WA_Fn-UseC_-Marketing-Customer-Value-Analysis.csv')
#Convert Response field into number
le = LabelEncoder()
df['Response']=le.fit_transform(df['Response'])
df['Gender']=le.fit_transform(df['Gender'])
df['Education']=le.fit_transform(df['Education'])

df.shape
```

```
(9134, 24)
```

Figure 7.
Dataset transformation.

Figure 7 illustrates how to convert data into numeric values for the Laplace-based data perturbation mechanism; therefore, noise can be added to those fields.

However, some fields such as "Employment Status," "State," "Gender," "Education," "Policy," and "Response" are quasi-identifiers that are not numeric data, while some numeric data, such as Income and Total Claim Amount, are sensitive data.

3.2. Step 2: Exploratory Data Analysis

IBM Watson Marketing Data Analysis and Prediction dataset consists of 24 attributes, as shown in Figure 4. Response attributes are a class in this experiment. A classification model is suitable for analyzing this dataset.

```
ax = sns.countplot(x='Response',data = df)
plt.ylabel('Total number of Response')
annot_plot(ax, 0.08,1)
plt.show()
```

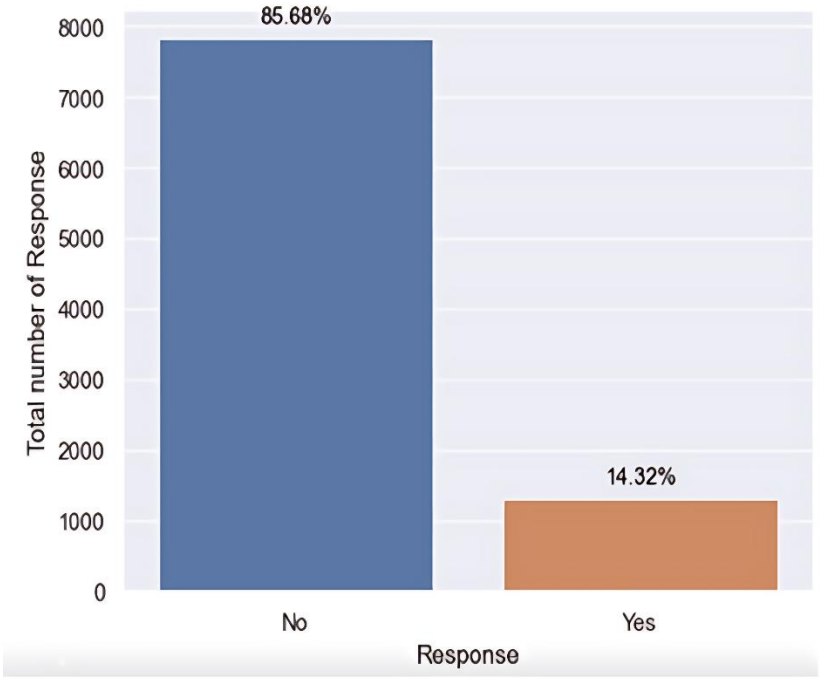

Figure 8.
Percentage of response distribution.

Figure 8 shows the percentage of response distribution. To protect data privacy, both algorithms add noise to the dataset. Finally, we can compare the effect of data utility from both algorithms, as shown in Section 5.

3.3. Step3: Compute Feature Importance

A random forest algorithm is a well-known algorithm for calculating these values to compute feature importance.

```
fs.plot(kind='barh')
<Axes: >
```

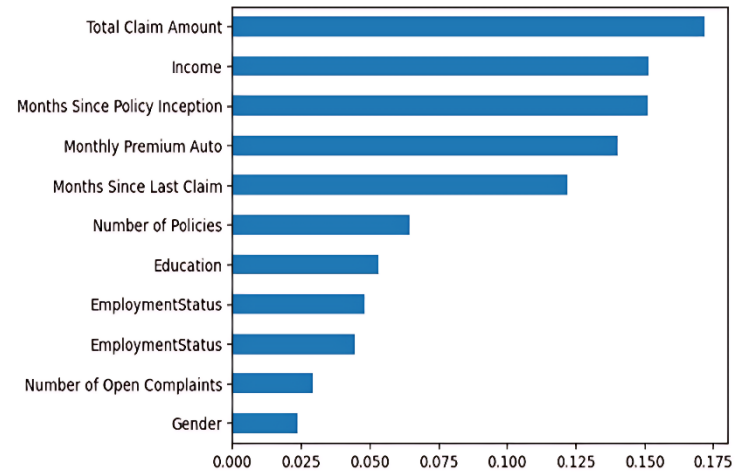

Figure 9.
Feature importance factors.

Figure 9 shows all features that strongly correlate with the "Response" class. The Total Claim Amount is one of the most important features in Figure 7; therefore, it is included in the sharable dataset. Additionally, the numeric value of the Total Claim Amount is suitable for both algorithms. Since this dataset consists of individual customer records, we separate and generalize the data, even though there are no direct identifiers such as personal ID, name, or last name.

```
df=pd.read_csv('K-ANONYMITY.csv')
df.head(10)
```

	Customer	State	Response	Education	EmploymentStatus	Gender	Income	Marital Status	Policy	L1_Total Claim Amount	L2_Total Claim Amount
0	FQ79358	California	No	Bachelor	Employed	M	99981	Married	Special L2	[0,250)	[0,500]
1	SG96898	Oregon	No	College	Employed	F	99961	Married	Personal L1	[250,500)	[0,500]
2	KF88603	Oregon	No	Master	Employed	M	99960	Married	Corporate L3	[250,500)	[0,500]
3	QO70290	Oregon	No	Bachelor	Employed	M	99934	Married	Special L3	[250,500)	[0,500]
4	JN26745	California	No	Master	Employed	F	99875	Divorced	Personal L3	[0,250)	[0,500]
5	JO64487	California	No	High School or Below	Employed	M	99874	Married	Corporate L2	[250,500)	[0,500]
6	KH53118	Arizona	Yes	High School or Below	Employed	M	99845	Married	Corporate L3	[500,750)	[500,1000]
7	BQ88033	California	Yes	High School or Below	Employed	M	99845	Married	Personal L2	[500,750)	[500,1000]
8	QJ77448	California	Yes	High School or Below	Employed	M	99845	Married	Personal L3	[500,750)	[500,1000]
9	FK17363	Oregon	Yes	High School or Below	Employed	M	99845	Married	Corporate L2	[500,750)	[500,1000]

Figure 10.
Sharable dataset.

Figure 10 shows a shareable dataset extracted from the IBM Watson Marketing Data Analysis and Prediction dataset. A new dataset is anonymized using data generalization techniques and saved as "Data Generalization Techniques.csv," as shown in Figure 7.

The Pearson correlation is used to calculate the relationship between variables. A heatmap diagram is used to visually show how all variables are related. The Pearson correlation indicates a linear relationship between the x and y variables.

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2(y_i-\bar{y})^2}} \quad (3)$$

Where: r = Correlation coefficient.
x_i= Values of x-variable in as a sample.
$\bar{x}$= Mean of the values of the x-variable.
y_i= Values of x-variable in as a sample.
$\bar{y}$ = Mean of the values of the y-variable.

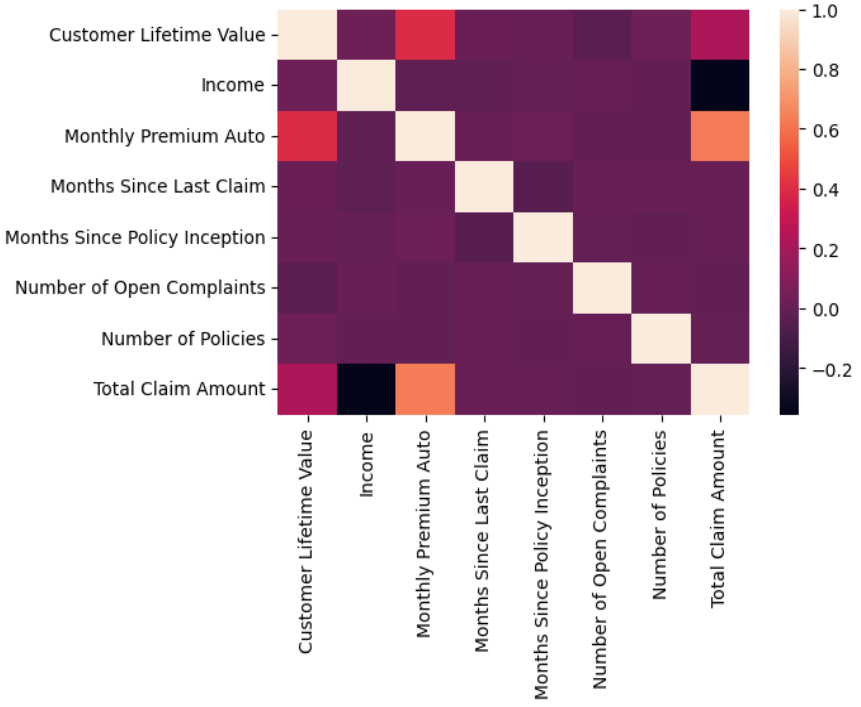

Figure 11.
Heatmap diagram of data sets.

Figure 11 is a heatmap that visually illustrates the relationships among all variables.

3.4. Step 4: Partition Each Quasi Attribute

Generalization is a process that makes some attributes less specific by broadening their scope, as shown in Figure 8.

```
df['EmploymentStatus'] = df['EmploymentStatus'].replace('Disabled','Unemployed')
df['EmploymentStatus'] = df['EmploymentStatus'].replace('Retired','Unemployed')
df['EmploymentStatus'] = df['EmploymentStatus'].replace('Medical Leave','Unemployed')
```

```
df['Education'] = df['Education'].replace('College','Prior University')
df['Education'] = df['Education'].replace('High School or Below','Prior University')
df['Education'] = df['Education'].replace('Bachelor','University level')
df['Education'] = df['Education'].replace('Master','University level')
df['Education'] = df['Education'].replace('Doctor','University level')
```

Figure 12.
Generalization on quasi-identifiers.

Figure 12 illustrates how to generalize both quasi-identifiers by making them less specific in order to conceal them within a larger group. For example, Disabled, Retired, and Medical Leave are regrouped into the unemployed category to obscure a record, as shown in Figure 6.

This research includes two quasi-attributes: Employment Status and Total Claim Amount.

```
df['EmploymentStatus'].unique()
```

```
array(['Employed', 'Unemployed', 'Medical Leave', 'Disabled', 'Retired'],
      dtype=object)
```

```
df[['EmploymentStatus','Total Claim Amount']]
```

	EmploymentStatus	Total Claim Amount
0	Employed	384.811147
1	Unemployed	1131.464935
2	Employed	566.472247
3	Unemployed	529.881344
4	Employed	138.130879

Figure 13.
All quasi-attributes.

Total claim amounts and income are sensitive data that need to be protected. They are real numeric values that are suitable for the data generalization mechanism.

EmploymentStatus	Gender	Income	Marital Status	Policy	L1_Total Claim Amount	L2_Total Claim Amount
Employed	M	99981	Married	Special L2	[0,250]	[0,500]
Employed	F	99961	Married	Personal L1	[250,500]	[0,500]
Employed	M	99960	Married	Corporate L3	[250,500]	[0,500]

Figure 14.
L1_Total Claim Amount and L2_Total Claim Amount.

Figure 11 shows how total claim amounts are generalized to disguise attackers. By using ranges of intervals, attackers cannot directly identify the records. Since income is not a quasi-attribute but real sensitive data, it does not need to be partitioned. Since data generalization techniques modify attributes differently, a researcher implements two experiments, as shown in Section 5.

4. Methodology

In the case of an outlier case, our algorithm HOPV implements both data generalization techniques and data perturbation because data perturbation works well only on numeric attributes. Data generalization techniques work well on text attributes that have a quasi-identifier characteristic.

A classification model is chosen to make a comparison. Response attributes are a class in this experiment. A classification model is suitable for analyzing this dataset.

Pseudo Code

```
1. If Exist (Outliners) then
2. Begin
3. Find all possible quasi-identifiers using group-by function (k==1)
4. If Exist (k==1) then
5.     Loop
6.     If numeric attributes then
           Perform data perturbation
7.     Else
           If Categorical attributes then
                 Perform generalization process
8.     Endif
9.     End loop
10.        Endif
```

Figure 15.
Pseudo code.

Figure 15 shows the pseudocode of our algorithm. Firstly, identifying outliers at this stage is crucial. If no outliers are found, the program will exit. Line 3 begins to check if quasi-identifiers exist. A group-by function, as shown in Figure 1, is used to perform this task. If one or two records satisfy the risk group conditions, then the quasi-attribute is generated to be less specific.

The research paper intends to show the effects of machine learning by comparing data perturbation and data generalization techniques; therefore, this research consists of two methodologies: Methodology for data generalization and methodology for perturbation techniques.

4.1. Methodology for Data Perturbation

Since our dataset is a well-known dataset that has been shared publicly for years, global Laplace-based data perturbation is chosen to protect each record by applying the Laplace mechanism. Since noise is applied to each record, a high value is chosen.

Data perturbation is a well-known algorithm that works effectively on all numeric attributes. The method involves adding noise to each piece of data. In this research, Laplace noise is applied to protect the real values of personal data.

```
df1=df
df1['Income'].head(5)
```

```
0    56274
1        0
2    48767
3        0
4    43836
Name: Income, dtype: int64
```

```
import numpy as np

def laplaceMechanism(x, epsilon):
    x +=  np.random.laplace(0, 1.0/epsilon, 1)[0]
    return x
```

```
df1['Income'] = df1['Income'].apply(laplaceMechanism, args=(2,))
df1['Months Since Policy Inception'] = df1['Months Since Policy Inception'] \
.apply(laplaceMechanism, args=(2,))
df1['Income'].head(5)
```

```
0    56272.667755
1       -0.627596
2    48766.465559
3       -0.277330
4    43837.750457
Name: Income, dtype: float64
```

Figure 16.
Apply noise to income attributes.

Figure 16 shows that a small noise is applied to a sensitive field called "Income" individually.

4.2. Methodology for Data Generalization Techniques

Figure 6 in Section I shows that some attributes are regrouped into new attributes.

```
: df['EmploymentStatus'].unique()

: array(['Employed', 'Unemployed', 'Medical Leave', 'Disabled', 'Retired'],
       dtype=object)
```

Figure 17.
Distinct values of employment status attributes.

Figure 17 shows three distinct values of Employment Status attributes. To generalize this attribute, we can regroup it into two values: Employed and Unemployed, as shown in Figure 6.

5. Experimental

This experiment aims to outline our procedures in the following steps.

5.1. Step 1: Effect on Data Utility Between Two Algorithms as Follows

Table 3 displays the percentage of accuracy scores from four algorithms applied to the dataset before and after.

Table 3.
Effect on data utility.

	Before (%)	L1_Total claim amount	Loss (%)	L2_Total claim amount	Loss (%)
KNeighbors	99.2473	97.7692	1.4781	97.6187	1.6286
Logistic regression	85.7397	85.6576	0.0821	85.7071	0.0326
Random Forest	96.4898	86.2863	10.2035	84.5145	11.9753
Laplace-based data perturbation	85.6712	85.6712			0

Note: L1_Total Claim Amount is four intervals (bin) of quasi-attributes. L2_Total Claim Amount is two intervals (bin) of quasi-attributes.

```
X = df1.iloc[:,df.columns != 'Response']
y = df1.Response

from sklearn.linear_model import LogisticRegression
import seaborn as sns
model = LogisticRegression()

X_train, X_test, y_train, y_test = train_test_split(X, y,\
test_size=0.20, random_state=5, stratify=y)

model.fit(X_train, y_train)
y_pred=model.predict(X_test)

#without noise Acc= 85.68495962775421

train_acc = model.score(X_train, y_train)
print("The Accuracy for Training Set is {}".format(train_acc*100))

The Accuracy for Training Set is 85.67127412070617

# without noise Acc= 85.65955117679256
test_acc = accuracy_score(y_test, y_pred)
print("The Accuracy for Test Set is {}".format(test_acc*100))

The Accuracy for Test Set is 85.6048166392994
```

Figure 18.
Effect of applying privacy mechanisms.

Figure 18 shows that Laplace-based data perturbation mechanisms affect logistic regression applications very little compared to data generalization techniques. The accuracy of the training set with noise is 85.6712% compared to that without noise, which is 85.6849%.

```
def laplaceMechanism(x, epsilon):
    x +=  np.random.laplace(0, 1.0/epsilon, 1)[0]
    return x
```

```
df1=df
df1.head(5)
```

	Response	Education	Gender	Income	Number of Open Complaints	Months Since Policy Inception	Number of Policies
0	0	0	0	56274	0	5	1
1	0	0	0	0	0	42	8
2	0	0	0	48767	0	38	2
3	0	0	1	0	0	65	7
4	0	0	1	43836	0	44	1

```
df1['Income'] = df1['Income'].apply(laplaceMechanism, args=(2,))
df1.head(5)
```

	Response	Education	Gender	Income	Number of Open Complaints	Months Since Policy Inception	Number of Policies
0	0	0	0	56274.272193	0	5	1
1	0	0	0	-0.676978	0	42	8
2	0	0	0	48766.349751	0	38	2
3	0	0	1	-0.018874	0	65	7
4	0	0	1	43834.089497	0	44	1

Figure 19.
Noise added to Income attributes.

Figure 19 illustrates how to introduce noise into the Income attribute as sensitive data in this experiment.

```
df['Education'].value_counts()
```

	Education
Bachelor	2748
College	2681
High School or Below	2622
Master	741
Doctor	342

```
epsilon = 1

f = lambda x: x + np.random.la
s = df['Education'].value_coun
s.to_frame().head(5)
```

	Education
Bachelor	2747.580831
College	2681.532647
High School or Below	2624.493422
Master	740.414911
Doctor	341.282864

Figure 20.
Histogram attacks.

Figure 20 shows that the noise added to datasets does not make a significant difference but can clearly disguise attackers.

5.2. Step 2: Apply Noise

Since the proof in step one, the Laplace mechanism noise is applied throughout the table to all numeric values because data perturbation does not affect data utility.

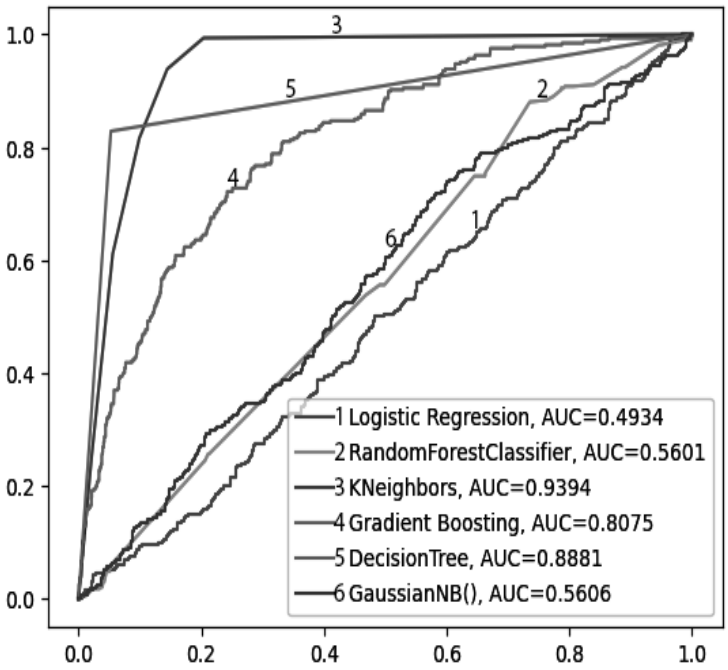

Figure 21.
ROC and AUC on an original dataset.

Figure 21 shows the Receiver Operating Characteristic (ROC) and Area Under the ROC Curve (AUC) diagram of the six classification algorithms of the original dataset without any division.

5.3. Step 3: Apply Generalization on Quasi Attributes

There are some quasi-identifiers in the preparation section. In this experiment, two quasi-identifiers qualify enough to be generated: Employment Status and Education, as shown in Figure 8. All other quasi-identifiers are unnecessary to generalize because they are all single-level attributes, such as gender and policy. The more they are generated, the more data utility is lost.

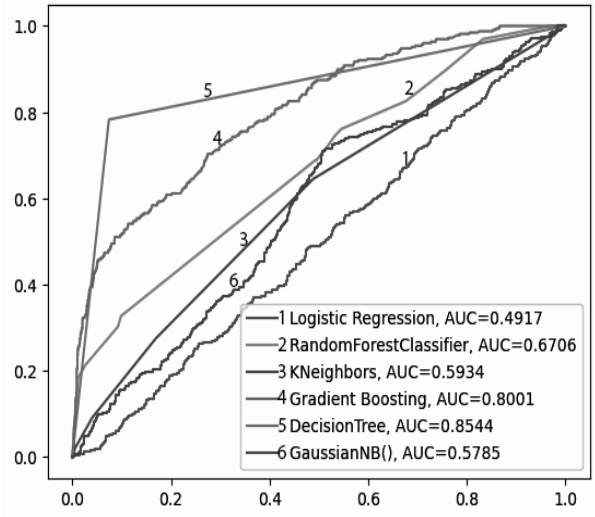

Figure 22.
ROC and AUC on Data Generalization Techniques dataset (modified).

Figure 22 illustrates the ROC diagram of the six classification algorithms from the Data Generalization Techniques dataset without any division.

6. Discussion

Table 2 shows that data generalization does not affect data loss due to the machine learning process; therefore, it can protect data privacy better than Data Generalization Techniques. Data generalization works best with all numeric values; therefore, all numeric sensitive data should be utilized.

This research aims to demonstrate a step-by-step implementation of how data generalization outperforms Data Generalization Techniques in our experiment.

Data perturbation does not affect the machine learning process, while Data Generalization Techniques suffer from it. Therefore, data generalization is suitable for privacy-preserving mechanisms. In other words, the noise added does not affect the meaning of data utility. This paper concludes that 1) data generalization can protect against differencing attacks, 2) it is also neutral to reidentification attacks, and 3) it does not affect the machine learning process.

6.1. Performance Comparison

Outliers pose a significant challenge in databases, leading to inaccurate data models and potential privacy violations due to their distinctive characteristics. These outliers can become targets for data breaches, as hackers may exploit their unique features to compromise system security. While truncating outliers is a common approach, this paper justifies the decision not to truncate the 74 outliers identified by DBSCAN and Z-score. The instances exemplify their substantial percentage and their representation of real and valuable cases in Figure 1.

Data controllers are faced with addressing outlier issues, which may involve either truncation or the application of privacy protection algorithms such as Data Generalization Techniques. In this study, DBSCAN and Z-score are utilized for both outlier detection and clustering; Data Generalization Techniques are applied to quasi-identifiers, and Laplace-based data perturbation is used for numeric sensitive data. Figure 19 highlights the enhancement in performance evaluation achieved when outlier issues are effectively managed.

The proposed HOPV framework provides a comprehensive solution for handling outlier problems. It involves the detection and generalization of outliers, separating the dataset into Personal Information Identifiers (PII) and de-identification segments. The de-identification segment is anonymized using the HOPV framework to ensure compliance with privacy laws.

In this research, we chose to use ROC and AUC metrics because such diagrams are accepted representations of evaluating the efficiency of the classification model.

It is clearly seen that the performance deteriorated in the classification model. This is evident in every algorithm, more or less. For example, KNN is reduced from AUC=0.9394 in Figure 18 to only AUC=0.5934 in Figure 19. We can conclude that it is evident that the k-anonymity process will definitely cause data utility losses.

From this fact, we should use the Laplace-based data perturbation algorithm rather than Data Generalization Techniques, but in the case where the field is not a number, perturbation cannot be used.

7. Conclusion

Due to the necessity of complying with Thailand's PDPA laws, a common approach is to make data anonymous, which is called “data anonymization.” In this article, two methods of data anonymization were chosen: Generalization and Data Perturbation Techniques.

With the property of a graph, the authors will present them using ROC and AUC for ease of presentation. The authors also showed a comparison of the data loss caused by the generalization process of Data Generalization Techniques in Figures 21 and 22. It can be seen that this is an unavoidable process because quasi-identifiers are usually not numerical data. The experiment in this research highlights the necessity of generalization and perturbation techniques in ensuring the anonymization of certain personal data.

The experimental results show that data perturbation has no effect on differencing attacks and is also neutral to reidentification attacks. Table 1 illustrates that Laplace-based data perturbation mechanisms have no impact on reidentification, as further demonstrated in Figure 10. The effect of data generalization is data loss because some quasi-attributes are modified to protect data privacy, as shown in Table 2.

The HOPV algorithm effectively preserves data privacy while maintaining utility by leveraging the Laplace Mechanism. In addition, HOPV also uses the concept of Data Anonymization to help deal with text attributes that the Laplace Mechanism cannot use. In summary, the HOPV algorithm can work with all types of information while ensuring data privacy protection and preserving data utility.

References

[1] S. Pachgade and S. Dhande, "Outlier detection over data set using cluster-based and distance-based approach," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 6, pp. 12-16, 2012.

[2] M. Feingold, M. Corzine, M. Wyden, and M. Nelson, "Datamining moratorium act of 2003," *U.S. Senate Bill (proposed)*, 2003.

[3] L. Sweeney, "Data generalization techniques: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, pp. 557-570, 2002. https://doi.org/10.1142/s0218488502001648

[4] M. Barbaro and T. Zeller, "A face is exposed for AOL searcher No. 4417749," Retrieved: https://www.nytimes.com/2006/08/09/technology/09aol.html, 2006.

[5] N. Homer *et al.*, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genetics*, vol. 4, no. 8, p. e1000167, 2008. https://doi.org/10.1371/journal.pgen.1000167

[6] S. Lohr, "Netflix cancels contest after concerns are raised about privacy," Retrieved: https://www.nytimes.com/2010/03/13/technology/13netflix.html, 2010.

[7] P. C. Console and Associates, "NIH federal credit Union notifies 14,706 members of data breach," Retrieved: https://beyondmachines.net/event_details/nih-federal-credit-union-reports-data-breach-to-14706-members-a-a-m-c-b, 2023.

[8] D. Barth-Jones, "The're-identification'of governor William Weld's medical information: A critical re-examination of health data identification risks and privacy protections, then and now," *Then and Now (July 2012)*, 2012. https://doi.org/10.2139/ssrn.2076397

[9] N. Perlroth, "All 3 billion Yahoo accounts were affected by 2013 attack," Retrieved: https://www.nytimes.com/2017/10/03/technology/yahoo-hack-3-billion-users.html, 2017.

[10] A. Majeed and S. Lee, "Anonymization techniques for privacy preserving data publishing: A comprehensive survey," *IEEE Access*, vol. 9, pp. 8512-8545, 2020. https://doi.org/10.1109/access.2020.3045700

[11] D. Deng, "DBSCAN clustering algorithm based on density," 7th ed. Hefei, China: International Forum on Electrical Engineering and Automation (IFEEA), 2020, pp. 949-953.

[12] C. Dwork and A. Roth, "The algorithmic foundations of Laplace-based data perturbation," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211-407, 2013.

[13] B. C. Kara and C. Eyupoglu, "Anonymization methods for privacy-preserving data publishing," presented at the In The International Conference on Artificial Intelligence and Applied Mathematics in Engineering (pp. 145-159). Cham: Springer International Publishing, 2021.

[14] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets." Oakland, CA, USA: In IEEE Symposium on Security and Privacy, 2008, pp. 111-125.

[15] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *In Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, San Diego, CA, USA, 2003, pp. 202-210*, 2003.

[16] B. C. Kara, C. Eyüpoğlu, S. Uysal, and S. Bayraklı, "Collection of an e-health dataset and anonymization with privacy-preserving data publishing algorithms," *Electrica*, vol. 23, no. 3, pp. 658-665, 2023. https://doi.org/10.5152/electrica.2023.23042

[17] R. Chen, B. C. M. Fung, B. C. Desai, and N. M. Sossou, "Differentially private transit data publication," presented at the In the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 213-221, 2012.

[18] A. Narayan and A. Haeberlen, "DJoin: Differentially private join queries over distributed databases," presented at the In the 10th USENIX Conference on Operating Systems Design and Implementation, Hollywood, CA, USA, 2012, pp. 149-162, 2012.

[19] C. Li and G. Miklau, "An adaptive mechanism for accurate query answering under Laplace-based data perturbation," *Proceedings of the VLDB Endowment*, vol. 5, no. 6, pp. 514-525, 2012.

[20] R. L. Wilson and P. A. Rosen, "Does protecting databases using perturbation techniques impact knowledge discovery?," *Advanced Topics in Database Research*, vol. 4, pp. 96-107, 2005.

[21] The Information Commissioner's Office, "ICO publishes guidance on privacy enhancing technologies," Retrieved: https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/09/ico-publishes-guidance-on-privacy-enhancing-technologies/, 2022.

[22] A. Gionis and T. Tassa, "K-Anonymization with minimal loss of information," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 2, pp. 206-219, 2008. https://doi.org/10.1109/icde.2008.4497483

[23] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *In Proceedings of Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, USA, 1996, pp. 226-231*, 2021.

[24] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," presented at the IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413-422, 2008.