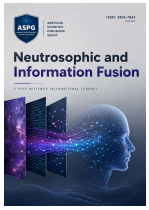
Neutrosophic and Information Fusion

♦ ISSN: 2836-7863 Vol: 04, No. 01, 2024 ♦ PP. 26–34

Article DOI: https://doi.org/10.54216/NIF.040105




FULL LENGTH ARTICLE

# Differentially Private Clustering with Dynamic Noise Adjustment (DPC-DNA) based Fusion Anonymity and Privacy Enhancement in Big Data

**Sergey Drominko**[1,*]

[1] Faculty of Information Technology and Robotics, Vitebsk State Technological University, Belarus

**Email:** Serdrominko1996@vsu.by

Received: December 17, 2023 Accepted: June 27, 2024 ⋆ Corresponding author

## ABSTRACT

Other few challenges faced during privacy preservation by anonymity e.g. difficulty in identifying the The main challenges in preserving anonymity for privacy are determining which attributes could undermine privacy and extracting useful information from massive databases without disclosing sensitive details. We developed a Novel Framework for Differentially Private Clustering with Dynamic Noise Adjustment (DPC-DNA) that addresses these issues. This novel approach can recognize sensitive and non-sensitive data aspects using Differentially Private Clustering with Dynamic Noise Adjustment (DPC-DNA). The accuracy of clusters formed by DPC-DNA was assessed using the silhouette score, which gauges how similar each item is to its own group versus others. DPC-DNA achieved a silhouette score of 0.62, signalling strong internal cluster composition. In contrast, traditional k-anonymity clustering yielded a lower score of 0.45, confirming that DPC-DNA significantly boosts accuracy. Our Novel Framework for Differentially Private Clustering with Dynamic Noise Adjustment (DPC-DNA) provides a robust solution for privacy-preserving data mining. By combining differential privacy with adaptive noise management, it safeguards sensitive material while sustaining high precision, integrity and usefulness of results.

**Keywords:** Fusion Anonymity and Privacy Enhancement ▪ Big Data ▪ Differentially Private Clustering ▪ Dynamic Noise Adjustment

## 1. INTRODUCTION

People are now aware of the disruptions that may occur in their private data, and as a result, they are unwilling to provide information that constitutes extremely sensitive information. The practice of protecting the confidentiality of sensitive information or personal data without compromising the usefulness of the data is referred to as privacy-preserving data mining or PPDM [1] for short. The most significant developments in the subject of data mining are those that pertain to privacy-preserving data mining, often known as PPDM. As a result, algorithms for privacy-preserving data mining (PPDM) were developed in order to extract precise information from a vast quantity of data while simultaneously taking into account the confidentiality of the data. The efficiency of the Privacy Preserving Data Mining (PPDM) method, on the other hand, was exposed to modification while the prevalent data mining approaches were being used.

### 1.1 The Significance

Because the data are spread in the majority of cases involving data mining, it is not very probable that the data that have been gathered will be observed in a particular area for the purpose of analysis in compliance with the privacy legislation or acts. Data mining [2] is a process that involves executing

© 2024 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

a secure and collaborative protocol on a specific distributed database in order to get the information that is sought and targeted. This is one of the most essential components of data mining that protects the privacy of users. The following is a list of the most important characteristics that the privacy-preserving algorithm must possess: it must be able to handle the various data mining techniques; it must be able to prevent the identification of information [3] that is sensitive; it must not have a high level of computational complexity built into it; and it must not easily permit the access and utilization of data that is not sensitive.

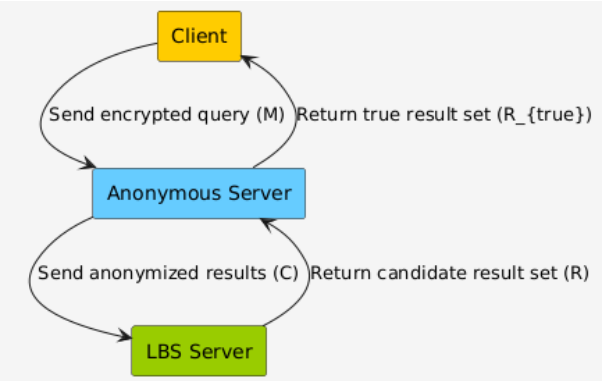



**Figure 1.** Baic Anonymous Server.

Up to this moment, a vast number of secret protocols have been proposed for the process of three association rule mining, clustering, decision tree classification, Bayesian Networks, and Neural Networks. These protocols have been proposed using a variety of different methods. Specifically, these protocols refer to methodologies that use data mining and machine learning. The most significant role of these algorithms is to protect the confidential information of the parties involved. This is, of course, in addition to the fact that they make it possible to get the knowledge that is both required and beneficial from the whole data sets. There was a possibility to avoid the creation of centralized warehouses in a number of different locations, which would have precluded the transfer of data from one site to another. This opportunity was offered by the preservation of data confidentiality. Because of this, the fundamental principles of association associated with data mining needed repeated database scanning. This scanning might be given in a distributed and centralized environment, where the required data could be described across a number of different websites. PPDM [4] is comprised of a number of fundamental components, one of which is the supply of universal data in conjunction with the notion of "differential privacy," which refers to the preservation of privacy in the prototype of statistics database.

### 1.2 Motivation

The objective is to extract precise data from massive databases without compromising the confidentiality of people's personal information. During the process of data mining, several ways were used to protect individuals' privacy. For protecting individuals' privacy during data mining, a number of different methods, including anonymization, cryptography [5], and randomization, may be used. Given the advancements that have been made in the analysis of the procedures, it is possible that the effectiveness of the privacy preservation will be improved. We are able to identify the extent of the loss of privacy that occurs during data mining thanks to the implementation of conditional privacy. The primary reason for this is to ensure that the data mining process does not violate any individuals' privacy and to enable security measures.

One of the most important aspects of this research is that a new anonymization approach is offered in order to protect the confidentiality of the sensitive data and information that is modified by the particular user. The structure of the article is as follows: Section 2 contains related work; Section 3 contains the methodology of the proposed work; Section 4 contains experimental analysis and the findings of the proposed work; Section 5 contains the conclusion and future work.

## 2. RELATED WORK

A technique of anonymization was studied by [6] using sub clustering, and it was said that this approach achieved maximum usefulness and privacy while costing very little time to execute [7]. In spite of the fact that PPDM algorithms were used for the purpose of the data streams, this method merely took into consideration a static database. During the course of this research, the problem of homogeneity attacks in SIT was handled by picking a number of different sub-clusters for relocation. After anonymization, the issue of similarity attack was overcome since the centroids of the subclusters were distinct from those that existed before anonymization. Skewness attack was used for the aim of detecting the remaining records since the concept of whole or partial records would not be helpful in detecting the remaining records. An increase in the number of subclusters was shown to provide a reduction for distortion, as demonstrated by the findings. The amount of documents that were protected from prying eyes, on the other hand, rapidly reduced. On the other hand, this specific strategy was only applicable to numerical elements, and it was particularly noteworthy for categorical items. It is summarized in the report that the method was computationally irreversible and that it avoided skewness attacks, reconstruction attacks, and similarity attacks, all of which are known to encourage data mining with the highest possible level of efficacy.

Reference [8] invested in privacy-preserving data mining methods and frameworks by building an algorithm to supply or conceal privacy to certain key data with the goal of preventing them from being accessible to unauthorized individuals [9]. The research looked at a variety of PPDM approaches that are currently in use. It was determined that the distinct values of the database should be modified for public access, which is referred to as data modification. This was done to retain a high level of privacy. The techniques of data alteration that are discussed in this study are as follows: (1) a disturbance that has a new value substituted for the distinctive value that was previously present; (2) blocking that resulted in a departure from the typical characteristic value; (3) swapping that included interchanging the value of each individual record; (4) sampling that only failed to collect data for a certain demographic; and (5) a kind of encryption that makes use of a variety of different cryptographic techniques. After doing the research, the researchers concluded that there is no single algorithm that can be used for privacy-preserving

data mining. This algorithm would be superior to other data mining algorithms in terms of cost, utility, performance, tolerance complexity, and other factors. In conclusion, different sorts of algorithms all have the potential to function more effectively than one another about a certain measure. A data warehousing technology known as Hive was used to ensure the protection of privacy in the setting of big data [10].

The work that was proposed was successful in achieving the release of data that protected individuals' privacy by using fusion anonymity and the NSB (nearest similarity-based clustering) algorithm, together with bottom-up generalization and flavors of clustering that eradicated the problems that were previously present. The whole process was carried out within the paradigm of Big Data [11], which resulted in an improvement in the dissemination of information and the availability of communication channels. In this investigation, the mortality rate was obtained using a single index value type. There is also the possibility of evaluating the sensitivity levels with different index values. It cleared the door for the straightforward implementation of the methods without requiring any form of modifications. Based on the findings of the whole research, it seems that the model that was used in this experiment was more effective than the other models that were prevalent at the time.

Cluster-based data reduction for the KNN classifier was developed in paper [12]. Classification is the process of assigning new data to a class that has already been established. This preprocessing has a cheap cost. It is possible to attain high reduction rates. The accuracy of classification is good. The proposed approach was carried out using two different ways. Reduction via homogeneous clustering and dynamic reduction through homogeneous clustering are the two methods that are being discussed here. Dynamic Reduction was used to cope with data that was not suitable for main memory by means of homogeneous clustering the data.

Reference [13] showed the grouping of numerical and categorical data based on density and distance. The approach of cluster center initialization was used in order to cluster both numerical and categorical data items. When dealing with numerical data, the cluster center initialization approach mixes centres that are comparable to the cluster center. Using the cluster center initialization approach for categorical data, it was possible to identify data items that had a high cluster membership. Because of the suggested solution, the issue of mixed data is resolved.

"Dimensionality Reduction in Hyperspectral Data Dependent upon Class Aware Tensor Neighborhood Graph" was the title of the study symbolized by [14]. Concerns with data redundancy and dimensionality were circumvented by the solution that was presented. In order to implement the suggested solution, the Classware Tensor Neighborhood Graph algorithm was used. The dimensionality reduction of hyperspectral data that has been proposed encompasses three distinct aspects: (1) tensor representation in hyperspectral data; (2) estimating the gap between the tensors; and (3) the Classware Tensor Neighborhood Graph. "Multi-mode radar signal sorting using spatial data" was the title of the study represented by [15]. The suggested technique included identification of the data field, the technique of first clustering, possibility of effacement of the canal, and the method of dividing. The multi-mode radar signal sorting by spatial data algorithm was successful in revealing the hidden dispersed information.

As stated in [16], these enhanced telecare medical information systems are characterized by high levels of fusion anonymity in addition to authentication. These efforts are mostly geared toward making it possible to get medical care from a distant location. Improvements were made to the healthcare procedure involving consumers and physicians. An adversary may be prevented from gaining access to servers and users by using remote authenticated protocol. Username and password are the two components that make up the two-factor authentication that is required for authentication. Users can get a session key, which allows for secure interaction between servers. It was suggested that a novel authentication mechanism be used to circumvent the security constraints. As part of the enhanced scheme, five distinct schemes have been accepted: the phase of system initialization, the phase of registration, the phase of login and verification, the phase of changing the password, and the phase of revoking the missing smart card. Finally, the revocation process was carried out if the smart card was misplaced. The password was updated during the password-changing period. The login and verification process included the validation of the user's ID and password. During the registration step [17], both the ID and the password are recorded. Mutual authentication, user anonymity, man-in-the-middle attack, impersonation attack, offline dictionary attack, replay attack, and privileged attack are some of the security performances that may be achieved via the usage of a unique method. Mutual authentication involves the validation of both the username and the password. A unique approach was used to achieve strong user anonymity. To the Telecare medical information system, authentication of the user was carried out. The storage overhead, communication expenses, and computation costs were investigated to determine how effective the unique technique was. To ensure safety, BAN logic and novel protocol are essential.

## 3. PROPOSED FRAMEWORK

Although explicit identifiers need to be eliminated, there is still the possibility of privacy invasion [19] occurring when quasi-identifiers are connected to data that is accessible to the public. This kind of assault is also known as a linking attack.

The date of birth, gender, race, and zip code are just examples of the kind of information that may be found in public documents such as voter lists. The method that is applied the most often in the present trends is known as k-anonymity. However, it still has the disadvantage of potentially losing information throughout the process of data transformation. Additionally, the k-anonymity model has two significant constraints that must be considered. When it comes to the external tables, it might be challenging to determine which characteristics are accessible and which are not available. Additionally, the adaptation of attack strategies to real-world settings is another item to consider. In most cases, there is a standard fusion anonymity model that is used for the queries that are provided by the user. This makes it simple to violate the privacy of the data. However, in this study, several methods of anonymization that are carried out for different inquiries that are made by the same user are carried out. This makes it difficult for the system to violate the confidentiality of the data.

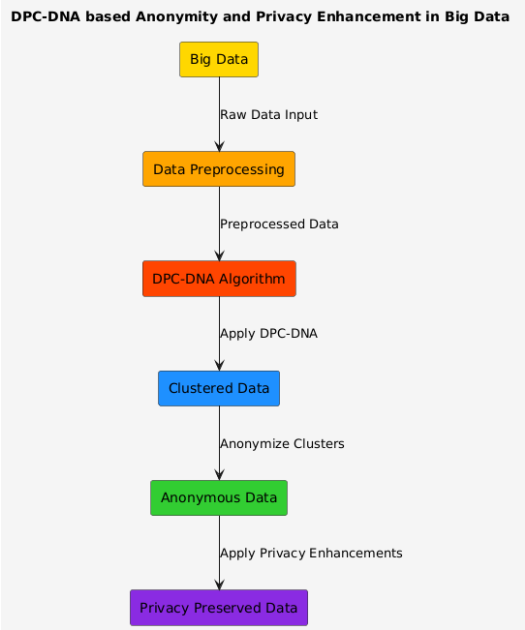



**Figure 2.** Block Diagram of Proposed Work.

Figure 2 provides an explanation of the workflow of a novel framework that is in the process of developing an efficient anonymity algorithm. This chapter also includes the presentation of methods such as the privacy score [20], attribute selection using principal component analysis, and novel-based anonymity algorithm. It is vital to separate the sensitive data characteristics from the non-sensitive data characteristics to reduce the amount of time that is needed via the calculating process. Specifically, this is because the preservation of the confidentiality of the whole data collection requires the most amount of time. Considering this, we have introduced the novel anonymity approach with the intention of achieving the highest possible privacy score while simultaneously minimizing the amount of time that is necessary for computation. In the method that we have developed, the use of the privacy score algorithm makes it feasible to preserve the secrecy of the data. Following that, the features of the data that were sensitive and those that were not sensitive were determined by the usage of attribute selection criteria that were obtained using principal component analysis. In conclusion, anonymous data was gathered using a one-of-a-kind algorithm that makes use of anonymity as its foundation throughout the process of data mining.

Figure 2 depicts the workflow of the DPC-DNA system, which is used for protecting the confidentiality of data that is generated throughout the data mining process. To begin, we will assume that the data that has been acquired from the dataset is the input. All the personal information pertaining to individual users has been saved and kept up to date using the datasets that were collected. Following the storage of personal information pertaining to a single user, the input data has been pre-processed via the use of the pre-processing method developed by data mining. In the context of data mining, the term "data pre-processing" refers to a procedure that converts the original data into a format that can be understood. To proceed with this procedure, the data that is being entered is pre-processed once again in preparation for the subsequent data mining process. The data that has been pre-processed is then made available for storage in the database system.

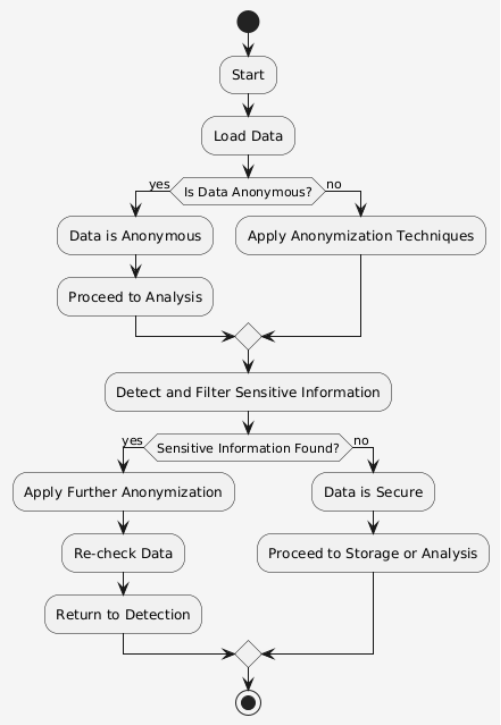



**Figure 3.** Flowchart of Anonymous Data Detection.

Using an attribute selection approach that is based on Principal Component Analysis, the attributes, such as sensitive and non-sensitive attributes, are defined after the input data has been pre-processed. It has been possible to differentiate between sensitive and non-sensitive data characteristics by using the attributes of a pre-processed data set that have been chosen. The determination of each aspect of anonymity is accomplished using a novel-based anonymity algorithm. The anonymous data has been obtained and assessed therefore. When it comes to pre-processing, the first step involves obtaining the input data from a dataset that includes personal information for many users in the context of data mining. During this stage of the pre-processing procedure, the noise reduction and special characters removal processes are both finished and estimated successfully. Characters, strings, and numerical values are the three categories of data that are included inside a dataset, respectively. This collection of unstructured data is transformed into a format that is organized using the pre-processing approach. This pre-processing method requires the transformation of various kinds of data into numerical values by making use of ASCII code. These numerical values are then used in the subsequent steps. Prior to being retrieved as pre-processed data, each value of data included in the dataset is treated until it reaches two digits. Following this, the data that have been pre-processed and are saved in the database are necessary in order to proceed with the subsequent operations. In the next step, the covariance matrix for each of the various scores of the data is assessed by applying the equation shown below:

$$CVM_{X,Y} = x\frac{Y_i}{N} \tag{1}$$

where $N$ represents the number of scores in each set of data; $X_i$ represents the $i$th raw score in the first set of scores; $Y_i$ represents the $i$th raw score in the first set of scores; and $CVM_{X,Y}$ represents covariance for respective scores in two

sets of data. After that, the Eigen values and vectors in the pre-processing technique are calculated and determined. With the help of the Eigen vector, the score attributes of original data are obtained by the following equation:

$$SC = [\text{ori } i_{dt}, E_{\text{Vec}}] \tag{2}$$

where ori $i_{dt}$ represents Original Data and $E_{\text{Vec}}$ represents Eigen Vector.

A privacy score, also known as privacy preserving in data mining, is defined as the process of achieving valuable results that render input data private. There are several types of techniques that are used to preserve privacy, including heuristic, cryptography, and reconstruction-based techniques. Within this privacy score, there are four types of attributes, including explicit identifiers, quasi-identifiers, sensitive attributes, and non-sensitive attributes. Explicit identifiers are defined as attributes that contain information that helps and identifies the owner of the record explicitly, such as the owner's name and other details.

**k-Anonymity Condition:**

$$k\text{-Anonymity} \Rightarrow \forall t \in T, \left|\left\{t' \in T : QI(t) = QI(t')\right\}\right| \geq k \tag{3}$$

where $T$ is the set of tuples and $QI(t)$ represents the quasi-identifiers of tuple $t$.

**Generalization for k-Anonymity:**

$$G(t) = \{\text{generalized values of } t \text{ for } k\text{-Anonymity}\} \tag{4}$$

**Information Loss (IL):**

$$IL = \frac{1}{|T|} \sum_{t \in T} \frac{|QI(t)|}{|QI_{\max}|} \tag{5}$$

where $QI_{\max}$ is the maximum number of quasi-identifiers.

**Data Perturbation:**

$$\widetilde{X} = X + \varepsilon \tag{6}$$

where $X$ is the original data and $\varepsilon$ is the noise added to perturb the data.

**Privacy Loss (PL):**

$$PL = 1 - \frac{\sum_{i=1}^{n} |X_i - \widetilde{X}_i|}{n} \tag{7}$$

where $X_i$ is the original value and $\widetilde{X}_i$ is the perturbed value.

An attribute that has the power to identify the owner of records that are merged with publicly accessible data is referred to as a quasi-identifier. A characteristic that contains precise information on a particular sensitive individual, such as their pay, sickness, and other features of sensitive data, is referred to as a sensitive attribute. If untrustworthy users of data were to be discovered in a data mining system, it is argued that a non-sensitive attribute is an attribute that does not cause any problems for users.

**Variance for Data Perturbation:**

$$\text{Var}(\varepsilon) = \sigma^2 \tag{8}$$

where $\sigma^2$ is the variance of the added noise.

It is necessary to eliminate the explicit identifiers; yet, there is a risk of privacy invasion when the quasi-identifiers are coupled with the data that is accessible to the public. These are the kinds of assaults that are referred to as connecting attacks for short. The characteristics that are accessible in public data, such as a person's name, gender, date of birth, and race, are among the most prominent instances of linking attacks. This may be exploited for the purpose of inferring the identity of the relevant person with a higher likelihood. It is possible to prevent the prediction of an individual's identity by using the k-anonymity model, which helps to limit the risk of potential invasions of privacy.

**Degree of Uncertainty (DU):**

$$DU = H(X) - H(X \mid Y) \tag{9}$$

where $H(X)$ is the entropy of the original data and $H(X \mid Y)$ is the conditional entropy given the perturbed data $Y$.

The client, the anonymous server, and the LBS server are all components of the architecture of the central system that protects users' privacy. In the event that the anonymous server is trusted, then the communication that takes place between the client and the trusted anonymous server is encrypted from the specific system. It is not possible to ensure the security of communication between a trusted anonymous server and LBS providers, and attackers have already begun to carry out attacks such as eavesdropping and interception, respectively.

**Quantification of Privacy (QP):**

$$QP = \frac{1}{|T|} \sum_{t \in T} \log_2 \left( \frac{1}{P(t \mid QI)} \right) \tag{10}$$

where $P(t \mid QI)$ is the probability of identifying $t$ given the quasi-identifiers $QI$.

The user in preserved data sends a query request $M$ and gives the privacy requirements, which include the minimum number of users $k$ in the anonymous group and the minimum area value of the anonymous region, and sends the encrypted query information $M$ to the anonymous server later. The anonymous server of preserved data obtains anonymous results set $C$ by k-anonymity based on clustering and then sends $C$ to the LBS server.

**Anonymization Mapping:**

$$A(t) = t' \tag{11}$$

where $t$ is the original tuple and $t'$ is the anonymized tuple.

**Privacy Score (PS):**

$$PS = \frac{1}{1 + e^{-\lambda(DU)}} \tag{12}$$

where $\lambda$ is a constant and $DU$ is the degree of uncertainty.

$R$ is the candidate result set that is sent from the LBS server to the anonymous server. Finally, the anonymous server filters the candidate result set $R$ based on the user's actual location information in order to get the genuine result set and the inquiries of a specific user have been answered. The most significant phase in the process of maintaining private data

is the anonymous processing of saved data, which involves generalizing the precise geographical information of the user into an anonymous area. In the event that the local density of user objects in the anonymous group is larger, then the query service faults will also be higher.

Distributing the users has been done in order to fulfill the goal of achieving private data. Through the use of the clustering density approach, this dispersed data is condensed to a substantially greater extent and then separated into many anonymous groups. The selection of individual data from a specific set of anonymous data that has been partitioned is another key aim that has to be accomplished. The item in the anonymous group that has the highest k density is more closely related to other objects, and picking it allows one to take into consideration several different criteria that the user may have.

A determination and evaluation of the correctness of privacy-preserving data for anonymous data has been carried out. The concept of anonymity is described as a characteristic or model that allows for the re-identification of data that has been disclosed versus data that may be released. Privacy is an extremely important factor in the processing of a wide variety of information via data mining. The process of measuring and indicating how the initial value of an attribute from a collection of characteristics has been computed and properly forecasted is referred to as measurement. In proportion to the increase in the degree of privacy, the amount of data loss that is created also increases. To obtain the required amount of privacy and data loss, it is important to strike a balance between the two categories.

As a method of evaluating and determining the original private data, quantification has been used to assess data privacy as a degree of uncertainty throughout the evaluation and determination process. When the degrees of uncertainty are reached to a great degree, the level of security afforded to the data pertaining to privacy also increases. The data perturbation technique in privacy preservation is also assessed, and its effectiveness is determined by the variance between the original data values and the perturbed data values.

## 4. RESULTS AND DISCUSSION

Although a large number of data mocking libraries are available for practically every platform and language, the Mockaroo dataset makes it possible to download a large quantity of test data that has been prepared without discrimination in a simple and fast manner. These datasets that have been downloaded are based on the needed specificity, and they can be inserted into the test environment using either the CSV or SQL formats, without the need for any further programming.

Figure 4 has a representation of a typical Mockaroo dataset. This dataset makes use of the field name, type, and choices. To facilitate the testing, modeling, and demonstration processes, Mockaroo makes it possible to create datasets that are both realistic and synthetic. It does this by using schematic data because of the prevalent data that the user has previously established. This provides the foundation for the creation of the dataset. The Mockaroo data structure has two different kinds of components: the Field type and the Field name. Both components are present. There is a correlation between the

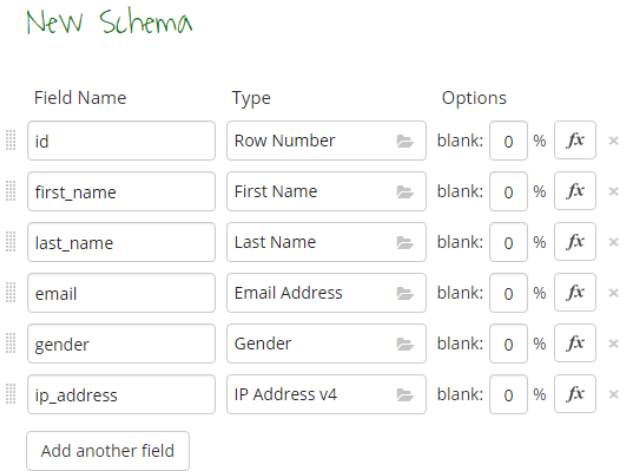



**Figure 4.** A Typical Mockaroo dataset (Courtesy: VMOKSHA).

field type and the field name, which may be either a numerical or a categorical value. The field name represents an attribute name that is present in the dataset that is supplied.

When Mockaroo builds the dataset, it does so based on a data set that first generates the field names and then randomly assigns the category field type to the related characteristic range. When it comes to creating an original request to the API, it is really challenging. After then, the user and the API explain the issue with the first request to improve the quality of the application flow, timeliness, and design later on. It is possible for users to have access to URL response and error condition with the assistance of this dataset, which will make it simpler for them to send various applications. It gives users the ability to submit data into a certain dataset during a predetermined amount of time. It is not possible to implement a significant volume of datasets in a short period. If the test dataset contains data that is representative of the actual world, then the tester will have access to correct data. Genuine data may include apostrophes or Unicode characters that are derived from languages other than English. The Mockaroo Generate API dataset is used for the purpose of generating single endpoint data requests, which includes the management of request parameters and failures. In order to generate realistic test data in formats such as CSV, JSON, SQL, and Excel, it enables users to generate one thousand rows. Every object needs to be responsible for determining its name, type, description, filename, project, and default attribute. In this case, the array size is returned by JSON if the string argument is greater than 1. To enable the user to construct data values based on the field name, a comma-separated value is provided. Excel and SQL both deliver the result as a document in the sql-xml format.

The modeling of the qualities and the assessment of marketing are two areas in which the Mockaroo dataset is used. Because digital marketing is becoming more prevalent in today's world, each of these two factors has a direct impact on the marginalization of the marketing industry. As a result, the last-click analysis is employed in the marketing sector for assessing the numerous channel performances that are included in the dataset. In conclusion, the application of cooperative game theory is undertaken with the purpose of improving the distribution of marketing analytics. For the purpose of gaining a better grasp of the Mockaroo datasets, Figure 4 is

used above.

In most cases, Eigen values may be discovered in matrix applications in the form of each set of values that includes a non-zero solution (that is, a magnitude of a certain value). The comparative analysis is very much essential since it gives the comparative study of the Eigen value, which is the most significant and initial factor that in turn influences parameters such as sensitivity values, SQRT and sensitivity levels, which are involved in this novel method proposed for the determination of the anonymity required.

**Table 1.** Comparison of Eigen Values of the proposed and existing systems

| Attributes | Existing System | Proposed System |
|---|---|---|
| AT1 | 7.26 | 12.16 |
| AT2 | 3.16 | 6.06 |
| AT3 | 1.75 | 5.41 |
| AT4 | 1.64 | 4.65 |
| AT5 | 1.37 | 4.50 |
| AT6 | 1.06 | 4.29 |
| AT7 | 1.00 | 3.59 |

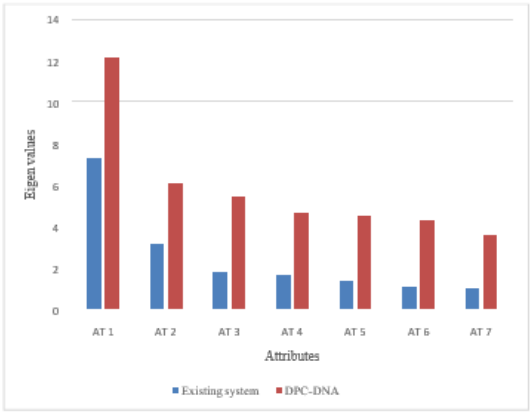



**Figure 5.** Comparison of Eigen Values of the proposed and existing system.

Figure 5 shows the graph that depicts the comparison between the attributes in the existing system in the phenomena of efficient anonymity and the Novel Framework for Efficient Anonymity Algorithm (NFAA).

### 4.1 Computational Cost

Computational cost or time is a measure of time taken in completing or executing a particular task or works or part of a task or work in the completely computational process considered.

Computational cost is to be compared with four other existing works and justification will be made concerning the better performance of this proposed work. For facilitating the comparison, the data from other papers are taken and given in Table 2. From this table, the computational cost data are used and correlated with the proposed work by plotting the graph between the system and time to find the improvement of the computational cost in the proposed work.

**Table 2.** Comparison of Computational time of the proposed and existing systems

| System | Time |
|---|---|
| [10] | 3.57 ms |
| [11] | 2.38 ms |
| [12] | 3.54 ms |

Figure 6 depicts the graph which shows the computational cost analysis for the DPC-DNA system, and it is compared with the various existing systems by taking the system and time in the x-axis and y-axis, respectively, for plotting the graph involved [17]. The comparison will be made with four existing methods given in the table one by one for performance understanding in terms of computational time.

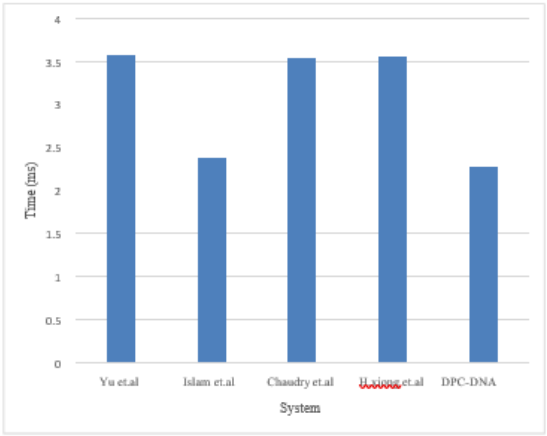



**Figure 6.** Comparison of Computational cost of the proposed and existing system.

When the existing Yu et al. system is considered, the DPC-DNA system has 36.13% less computational time when compared with the result obtained by this work. When the existing Islam et al. system is considered, the DPC-DNA system has 4.20% less computational time when compared with the result obtained by this work. When the existing H. Xiong et al. system is considered, the DPC-DNA system has 35.77% less computational time when compared with the result obtained by this work. In addition, we can infer that the proposed Novel Framework for Efficient Anonymity Algorithm (DPC-DNA) system yields the far better computational time of 2.28 ms, and the efficient computational time yielded by all the systems including the proposed work is 2.28 ms by the DPC-DNA system.

On comparing, we came to know that our DPC-DNA system has the lowest computational time incurred in the processes towards anonymity and H. Xiong et al. system has the highest computational time that had been incurred in the processes towards anonymity. When a computational cost is lower for any systems/processes running, it provides a better result in general, which is also not an exception in this case. Hence, ultimately the proposed Novel Framework for Efficient Anonymity Algorithm (DPC-DNA) system is giving better yields.

### 4.2 Privacy Level

Privacy level is nothing but the nature of one of the data sources to be isolated from the other data sources in any system considered. Firstly, Table 3 is tabulated for the purpose of correlating and analyzing the DPC-DNA system with the existing ones by plotting the graph.

**Table 3.** Comparison of Privacy level of the proposed and existing systems

| Attribute | BPS | LKCA | KDP | Proposed System |
|---|---|---|---|---|
| AT1 | 3.83 | 4.60 | 5.11 | 2.81 |
| AT2 | 5.66 | 6.97 | 8.98 | 2.36 |
| AT3 | 7.01 | 7.95 | 8.99 | 2.30 |
| AT4 | 8.70 | 9.01 | 9.80 | 2.21 |
| AT5 | 10.11 | 11.30 | 11.90 | 2.19 |
| AT6 | 11.30 | 12.50 | 13.25 | 2.17 |
| AT7 | 9.30 | 10.15 | 11.30 | 2.07 |

The graph is plotted by taking the attributes on the x-axis and privacy level on the y-axis, in which seven attributes are taken and analyzed with the proposed system of DPC-DNA and existing techniques of BPS, LKCA and KDP. The comparison will be made with four existing methods given in the table one by one for performance understanding in terms of computational time.

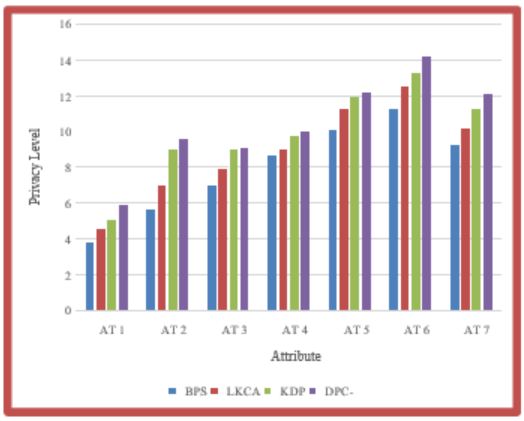



**Figure 7.** Comparison of Privacy level of the proposed and existing systems.

In comparison to the current BPS methods, which have a privacy level of 3.83, the existing LKCA techniques, which have a privacy level of 4.6, and the existing KDP techniques, which have a privacy level of 5.11, the suggested DPC-DNA system has a privacy level of 2.81 when AT1 is considered.

## 5. CONCLUSION AND FUTURE SCOPE

Using this algorithm, the confidentiality of the personal information was maintained. Within the framework of the data mining approach, the protection of personal privacy was given a vital role. To acquire the dataset that was used with the qualities such as alphabets and integers, the NBA method was used. This algorithm provided the sensitivity that was listed in the table. A hash code conversion that was completely anonymous was used to carry out the process of converting the alphabetical and numerical characteristics data record. A comparison was made between the information loss that was acquired by the previous system and that of DPC-DNA. The privacy level of the DPC-DNA technology takes into consideration a variety of distinct characteristics. Keeping the sensitive information about each individual user secure is the primary purpose of this thesis, and we can do this via the use of our innovative anonymization method to choose the sensitive characteristic from the information pertaining to each unique user. Additionally, the performance of the DPC-DNA system was confirmed using the experimental analysis. The Mockaroo dataset was used to conduct an analysis of the performance of the suggested system. The output that this program ended up producing was data that was anonymous. Through performance study, we have shown that the suggested technique has the greatest Eigen values, the lowest amount of information loss, and the lowest amount of computational cost. The suggested approach of privacy models was shown to be effective. According to the results of the comparative study, the suggested framework provides superior performance in comparison to the system that is already in place. A comparison was made between the proposed framework and the present system, and the conclusion that was reached was that the suggested framework offers better performances.

## REFERENCES

[1] B. Suma and G. Shobha, "Fractional salp swarm algorithm: An association rule based privacy-preserving strategy for data sanitization," *Journal of Information Security and Applications*, vol. 68, p. 103224, 2022.

[2] Y. Chen, Q. Zhao, P. Duan, B. Zhang, Z. Hong, and B. Wang, "Verifiable privacy-preserving association rule mining using distributed decryption mechanism on the cloud," *Expert Systems with Applications*, vol. 201, p. 117086, 2022.

[3] A. Kiran and N. Shirisha, "K-anonymization approach for privacy preservation using data perturbation techniques in data mining," *Materials Today: Proceedings*, vol. 64, pp. 578–584, 2022.

[4] V. H. Durelli, R. S. Durelli, S. S. Borges, A. T. Endo, M. M. Eler, D. R. Dias, and M. P. Guimaraes, "Machine learning applied to software testing: A systematic mapping study," *IEEE Transactions on Reliability*, vol. 68, no. 3, pp. 1189–1212, 2019.

[5] C. Murphy, G. E. Kaiser, and M. Arias, "An approach to software testing of machine learning applications," 2007.

[6] L. C. Briand, "Novel applications of machine learning in software testing," in *2008 The Eighth International Conference on Quality Software*. IEEE, 2008, pp. 3–10.

[7] N. Baskiotis, M. Sebag, M. C. Gaudel, and S. D. Gouraud, "A machine learning approach for statistical software testing," in *IJCAI*, 2007, pp. 2274–2279.

[8] M. Noorian, E. Bagheri, and W. Du, "Machine learning-based software testing: Towards a classification framework," in *SEKE*, 2011, pp. 225–229.

[9] A. R. Lenz, A. Pozo, and S. R. Vergilio, "Linking software testing results with a machine learning approach," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 5–6, pp. 1631–1640, 2013.

[10] H. B. Braiek and F. Khomh, "On testing machine learning programs," *Journal of Systems and Software*, vol. 164, p. 110542, 2020.

[11] D. Marijan, A. Gotlieb, and M. K. Ahuja, "Challenges of testing machine learning based systems," in *2019 IEEE International Conference on Artificial Intelligence Testing (AITest)*. IEEE, 2019, pp. 101–102.

[12] J. Kahles, J. Törrönen, T. Huuhtanen, and A. Jung, "Automating root cause analysis via machine learning in agile software testing environments," in *2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST)*. IEEE, 2019, pp. 379–390.

[13] R. Lachmann, "Machine learning-driven test case prioritization approaches for black-box software testing," in *The European Test and Telemetry Conference*, Nuremberg, Germany, 2018.

[14] S. Nakajima, "Quality assurance of machine learning software," in *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2018, pp. 601–604.

[15] R. Gove and J. Faytong, "Machine learning and event-based software testing: Classifiers for identifying infeasible gui event sequences," in *Advances in Computers*. Elsevier, 2012, vol. 86, pp. 109–135.

[16] D. Zhang, "Machine learning in value-based software test data generation," in *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*. IEEE, 2006, pp. 732–736.

[17] D. G. e Silva, M. Jino, and B. T. De Abreu, "Machine learning methods and asymmetric cost function to estimate execution effort of software testing," in *2010 Third International Conference on Software Testing, Verification and Validation*. IEEE, 2010, pp. 275–284.

[18] S. Huang, E. H. Liu, Z. W. Hui, S. Q. Tang, and S. J. Zhang, "Challenges of testing machine learning applications," *International Journal of Performability Engineering*, vol. 14, no. 6, p. 1275, 2018.

[19] S. Masuda, K. Ono, T. Yasue, and N. Hosokawa, "A survey of software quality for machine learning applications," in *2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2018, pp. 279–284.

[20] J. J. Li, A. Ulrich, X. Bai, and A. Bertolino, "Advances in test automation for software with special focus on artificial intelligence and machine learning," *Software Quality Journal*, vol. 28, no. 1, pp. 245–248, 2020.