International Journal of Intelligent Information Systems
2025, Vol. 14, No. 3, pp. 44-59
https://doi.org/10.11648/j.ijiis.20251403.11

SciencePG
Science Publishing Group

Research Article

An Effective Clustering Based Privacy Preserving Model Against Feature Attacks

Muhammad Zulqurnain[1, *], Muazzam Ali Khan Khattak[1], Adeel Anjum[2], Tehsin Kanwal[3]

[1]Departmentof Computer Science, Quaid-I-Azam University, Islamabad, Pakistan

[2]Institute of Information Technology Quaid-I-Azam University, Islamabad, Pakistan

[3]Departmentof Computer Science Comsats University Islamabad, Pakistan

Abstract

The rise in healthcare-related illnesses has generated a substantial amount of patient data, making the safeguarding of patient data imperative. Existing privacy protection methods face challenges, including longer execution times, compromised data quality, and increased information loss as data dimensions expand. Effective attribute selection is vital to enhance preservation methods. Our research introduces a privacy-preserving clustering approach that addresses these concerns through two stages: feature selection and anonymization. The first stage selects relevant features using symmetrical uncertainty (SU) and eliminates duplicates with Kendall’s Tau Correlation Coefficient. The Utility Preserved Anonymization (UPA) algorithm is employed in the second phase to achieve privacy preservation. Additionally, our approach reduces data dimensionality to simplify cluster creation for anonymization. Experimental analysis on real-time data demonstrates the strategy’s effectiveness, with outstanding sensitivity (97.85%) and accuracy (95%), efficiently eliminating unnecessary features and simplifying clustering complexity.

Keywords

Privacy-preserving, Clustering, Feature attack, Anonymization, Clinical Data, Feature Selection, Sensitive Attributes

1. Introduction

Recent IT advances have simplified personal data storage, particularly in healthcare where clinical data is retained for further analysis ([6]). This data often contains private information, and individuals are more willing to share it if confidentiality is ensured ([23]). Data mining has emerged to glean insights from health data while preserving privacy, finding applications in various domains like social networks, online services, commerce, and healthcare.

Privacy laws mandate the confidentiality of medical records, yet specific privacy threats persist in handling sensitive medical data ([13]). Consequently, many countries have implemented regulations to safeguard this information.

In our increasingly interconnected world, data protection has become a global compliance priority. Nations are enacting privacy legislation to safeguard personal information from public exposure. Various anonymization techniques,

[*]Corresponding author: lasharibscs@gmail.com (Muhammad Zulqurnain)

Received: 21 March 2025; **Accepted:** 19 April 2025; **Published:** 19 June 2025

including k-anonymity, l-diversity, (α, k)-anonymity, and t-closeness, have emerged ([15]).

K-anonymized datasets can be vulnerable to privacy breaches, like similarity, homogeneity, and background knowledge attacks, particularly when critical attributes have low variability [10]. Re- searchers have explored clustering methods, such as k-means clustering, to address these issues by identifying useful traits for anonymization [3]. How- ever, safeguarding sensitive data on cloud platforms requires more than k-means clustering, necessitating privacy-preserving outsourcing of the process [14].

Privacy concerns have surged, covering personal information protection to prevent adverse physical, psychological, and financial consequences. For example, revealing a patient's cancer diagnosis to their insurer or employer can profoundly affect their life. Ensuring personal information protection is critical throughout data collection and publication [4].

Privacy regulations mandate the removal of personally identifying information (PII) from patient data, such as names and Social Security numbers ([28]). However, privacy is not maintained solely by eliminating PII, as quasi-identifiers like age, gender, blood type, and religion, when combined with external data like a voter list, can still identify individuals. A privacy-preserving technique safeguards clinical data privacy using clustering- based anonymization on a dataset with 699 features. Symmetrical uncertainty-based feature selection is employed to protect sensitive demographic details, minimizing loss of information.

Privacy protection traditionally relies on anonymization methods like generalization and suppression to secure sensitive data. However, connecting quasi-attributes with publicly available information can still reveal identities, underscoring the goal of preserving individual privacy in the anonymity model.

To enhance privacy and data quality, redundant features are removed using Kendall's Tau rank correlation coefficient before clustering. The K- anonymization-based clustering method accelerates healthcare data anonymization with high accuracy. The following list explains the main contributions of the suggested system:

1. By combining generalisation and clustering algorithms, one may design an efficient anonymization solution enhancing data quality and reducing information loss.
2. The dimensionality of attributes is streamlined using Kendall's Tau Based Feature Selection (KTFS) in order to lessen complexity during the clustering phase.
3. Through the creation of clusters, healthcare data may be anonymized while maintaining patient privacy, establishing a balance between less information loss and enhanced data quality.

A. Research Motivation

Healthcare institutions have vast patient data repositories due to increased IT adoption. Sharing this data for research and improved patient care faces ethical and legal hurdles. K-anonymity is a common privacy safeguard, but it has drawbacks like information loss and classification errors due to reduced data granularity.

Our proposed framework integrates data anonymization, feature selection, and clustering techniques, offering a holistic solution. This approach enables organizations to leverage their data while protecting privacy. This study equips organizations with a robust methodology for responsible data utilization in a data-centric world. The remainder of this paper is organized as follows: related work discussed in section II, proposed model of privacy explained in section III, formal modelling and analysis section in section IV, results and discussion observed in section V, and conclusion of this paper is in section VI.

2. Related Work

Improvements in patient care quality, epidemiological research, and overall healthcare management have been sparked by recent breakthroughs in the healthcare industry, especially in the context of exchanging sensitive clinical information. But the increasing worry about privacy violations in patient diagnostic data emphasizes how important it is to protect data privacy.

A. Sensitive semi-identifier data anonymization

Privacy preservation models like Diversity and Proximity ([22]) are essential for anonymization. We've introduced l-diversity and t-closeness strategies for safeguarding sensitive quasi-identifiers. However, distinguishing the sensitive property from quasi-identifiers remains challenging. For example, in clinical data, both illness (sensitive) and attributes like age, residence, and work (quasi-identifiers) contain sensitive information. Our research proposes a method to anonymize features within each attribute.

Our approach comprises two algorithms: one for anonymization and one for reconstruction. It enhances microdata tables using the (p+, *alpha*)- sensitive k-anonymity characteristic from the Enhanced P Sensitive K-Anonymity Model ([27]). This approach reduces the likelihood of similarity attacks and distortion compared to the p-sensitive k-anonymity method. Additionally, the l-diversity method addresses k-anonymity's limitations.

B. Anonymization of Data Based on Features

Gachanga et al. ([5]) proposed a feature-based anonymization method for high-dimensional data, reducing dimensionality by selecting relevant features through information acquisition and ranking. Their approach combines feature selection, data slicing, and distortion minimization, enhancing data value and was evaluated using classifiers on anonymized datasets.

Chunhui Piao et al. ([20]) introduced CATDS, a Cluster-Based Anonymous Table Data-exchange Privacy Protection Method, designed for secure government data exchange with personal information. It involves data preparation and k-medoids clustering- based division of data tables. Experiments comparing it to the Incognito algorithm showed its effectiveness in reducing data quality preservation and information loss.

C. Model that protects privacy using clustering techniques

In ([19]), a probabilistic preservation method based on clustering is introduced to enhance data security in large datasets. It prioritizes privacy and involves identifying sensitive data within clusters and making alterations to protect it. Clustering is central to preserving individual data privacy with minimal disruption.

In ([11]), a novel (a, k)-anonymity model-based approach is presented for privacy-preserving data collection in healthcare services. It employs a user- to-client-server technique to assess the risk model and create anonymous records on the client side. Bottom-up clustering is used to build clusters that meet the (a1, k1) anonymity privacy level.

D. Techniques for Generalising and Suppressing health Data

We assessed the impact of anonymization methods on machine learning models [24], finding that in- creased anonymity requirements reduce classification accuracy, but the impact varies by dataset and technique. A novel generalization and suppression- based approach ([9]) mitigates information loss and improves data quality, addressing computational costs and scalability concerns.

Rodriguez et al. demonstrated the effect of k- anonymous micro-aggregation on small datasets ([21]).

(MD): This is known as a table linkage attack, where the goal is to determine if an individual's record is in the table. A robust privacy-preserving system should prevent membership identification. The k-anonymity model achieves this by applying generalization methods to protect the records table. Attack using Attribute Disclosure: When an unauthorized third party tries to learn someone's personal information, it is known as an attribute disclosure or attribute linkage assault. The intrusive party has some background knowledge about the target and wants to obtain sensitive data. To repel such assaults, one might use strategies like generalisation and suppression. Similarity Attack: An identical attack, also called a similarity or matching attack, occurs when an adversary tries to deduce sensitive information about an individual by comparing their data with external datasets, exploiting similarities, patterns, or correlations.

E. Privacy Attacks

Privacy attacks refer to malicious or unauthorized actions taken to compromise an individual's or an organization's private information. These attacks exploit vulnerabilities in systems, processes, or human behavior to gain access to sensitive data. Identity Disclosure (ID) attack: These attacks are also known as "record linkage attacks." Our method employs k-anonymity principles to mitigate identity disclosure risks. With k-anonymity, the likelihood of exposure is only 1/k, even when an adversary has access to both sensitive attributes and quasi- identifiers. Attack using Membership Disclosure

Table 1. *Execution times and privacy assaults are compared with newly suggested approaches and state-of-the-art privacy preservation techniques.*

study	time evaluation	s-attack	md	id	ad	**([12])**
	10^4 ms	X	√	√	X	([29])
	10^3 ms	X	√	√	√	([25])
	10^3 ms	√	√	√	√	([16])
	10^3 ms	X	√	√	√	([25])
	10^2 ms	√	√	√	√	([26])
	10^1 ms	√	√	√	√	
Proposed	10^1 ms	√	√	√	√	

3. Proposed Model

In this section, we present our patient clinical data anonymization scheme, emphasizing the importance of anonymizing healthcare records before sending them to data servers or cloud platforms. Our approach employs clustering-based anonymization and an enhanced k-anonymity model to minimize information loss.

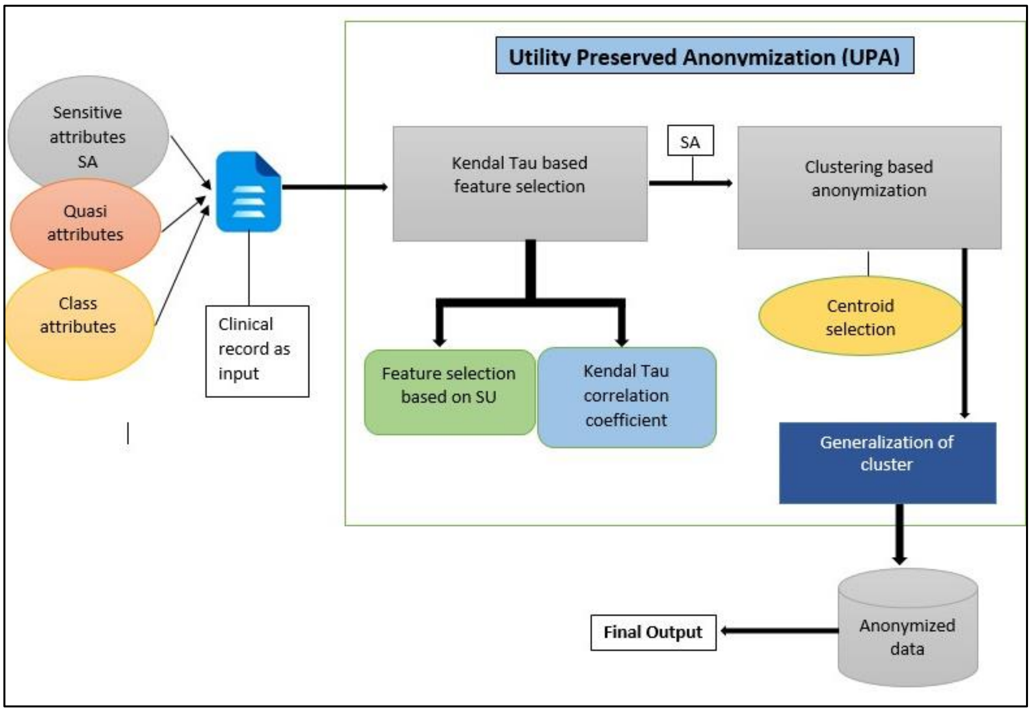

Utility Preserved Anonymization (UPA)
Sensitive attributes SA
Quasi attributes
Class attributes
Clinical record as input
Kendal Tau based feature selection
SA
Clustering based anonymization
Feature selection based on SU
Kendal Tau correlation coefficient
Centroid selection
Generalization of cluster
Anonymized data
Final Output

Figure 1. *Schematic Representation of Proposed system.*

Our healthcare data collection method focuses on privacy protection through anonymization, as depicted in Figure 1. The process begins with dataset collection, specifically clinical records related to healthcare. Feature selection is the first step, utilizing the Symmetrical Uncertainty (SU) approach to simplify the selection of essential attributes. Following feature selection, the next step involves removing redundant attributes within the dataset. The selected attributes are then used in the Utility Preserved Anonymization (UPA) algorithm, which employs generalization and clustering to achieve anonymization. The clustering process starts with the calculation of centroid points, grouping instances into clusters based on their proximity to these centroids. This results in generalized anonymized data, and the following section provides a detailed step-by-step explanation of this anonymization approach.

A. Description of Proposed System

The generation of anonymized data from the collected clinical records encompasses several stages. The purpose of the anonymization process is to safeguard sensitive information, such as personal identities, from potential privacy breaches. The clustering- based k-anonymization approach has been developed for this suggested system to provide privacy protection while reducing data loss. The sequential process of the suggested privacy-preserving anonymization approach is elucidated as follows.

B. Collection of datasets

Healthcare datasets, typically stored in secure databases, comprise data from various medical facilities, including both clinical and non-clinical patient information. These datasets contain both numerical and categorical data, categorized into three groups: explicit identifiers (e.g., names, license numbers, social security numbers), quasi-identifying attributes (e.g., age, sex, zip codes), and sensitive features (e.g., living situations, salaries, occupations). While explicit identifiers must be removed for data analysis before publication, sensitive characteristics may be retained in their original state. Anonymization techniques are most applied to quasi- identifying attributes.

C. Selection of Features depending on SU

Symmetrical uncertainty (SU) measures the effectiveness of feature classification by evaluating the relationship between features and the target concept, with higher SU values indicating greater relevance. A dataset-specific machine learning method aids in selecting quasi-attributes for anonymization from a pool of existing attributes.

Feature subset selection involves identifying and eliminating redundant features, simplifying attribute dimensionality to enhance machine learning algorithm efficiency and classification accuracy. After feature filtering, a relevance index, or scoring, is generated to assess the chosen feature’s connection to classification. This index serves as a heuristic criterion in filter techniques, highlighting the effectiveness of feature selection compared to other wrapper approaches for classification.

To achieve anonymization, features at the lower end of the relevance index are selected as quasi- attributes. Utilizing a feature selection approach expedites the clustering process for anonymization, improving data quality. SU plays a role in calculating the relevance index for each feature ([8]).

Symmetrical uncertainty (SU) quantifies the relevance between two random variables, yielding values from 0 to 1.

When SU equals one, variables X and Y are dependent; when it's zero, they're independent.

In classification, SU assesses the relationship between the class and features, calculated as follows for random variables X and Y.

Equation 1 shows the unbiased estimator of population variance σ^2

$$SU\ (x, y) = 2\frac{\mathrm{I(x|y)}}{H(x) + H(y)} \quad (1)$$

Gaining knowledge in pairs between the variables Y and X is denoted in the equation above by the symbol I(x,y). H(Y) is representation of variable Y's entropy. H(X) is a symbol for the variable X's information entropy. I(X—Y) evaluates X's understanding of Y if Y, X represent the classification name and characteristic, respectively. After, it is possible to calculate the mutual information between X and Y as illustrated below.

$$Ix, y = H(x) + H(y) - H(x, y) \quad (2)$$

$$H(x) = \sum_{x\in X} p(x) \log p(x) \quad (3)$$

$$H(x) = \sum_{x\in Y} p(y) \log p(y) \quad (4)$$

$$H(x) = \sum_{x\in X} \sum_{x\in Y} p(x, y) \log p(x, y) \quad (5)$$

While $x \in X$ and $y \in Y$ denotes potential values of both X, Y. The distributions of x, y the joint distribution are indicated, respectively, by the notations p(x), p(y), and p(x, y).

D. Elimination of redundant features through Kendall's Tau correlation coefficient

Kendall's Tau correlation coefficient, often simply referred to as Kendall's Tau or Kendall's rank correlation coefficient, is a statistical measure used to quantify the degree of agreement or association between two ranked variables. It assesses how similar the ordering of values in one variable is to the ordering of values in another variable. This makes it suitable for cases where the real numerical the variables' values don't much important, but their relative rankings are. Here is the equation of Kendall's Tau correlation coefficient.

Kendall's Tau:

Calculate concordant couples (*C*). Calculate discordant couples (*D*). Calculate tied couples (*T*), if applicable. Plug the values into the formula:

$$\tau = \sqrt{\frac{\mathrm{C-D}}{\mathrm{(C+D+T)\times(C+D+T-1)}}}$$

Interpret the Result:

The resulting value of Kendall's Tau (τ) remains

-1 and +1. τ = +1 indicates perfect positive agreement. τ = −1 indicates perfect negative agreement. τ = 0 indicates no agreement.

E. KTFS Algorithm

The following section outlines the stages that make up the KTFS algorithm. The quasi-identifier property and the parameter *delta* are included in the input. The output of running this algorithm is a collection of chosen features.

1. Feature set initialization, where *FS* is an empty set.
2. Utilizing SU, determine each feature attribute's relevance index. These attributes are part of *FS*.
3. The qualities are ordered in decreasing order based on the relevance index.
4. Then, using the Kendall's Tau coefficient, the redundant characteristics are eliminated from *FS*.
5. Using the ranking function *delta*, Select the desired characteristics in *FS*.

Algorithm 1: Pseudo code for KTFS Algorithm Input: *QIS* = {*A*1, *A*2, . . . , *Am*}, δ

```
Output: FS
1. $RI = SU\ (QIS)$ // Using SU, find related values for everyone.
2. $X1 = DO(RI)$ // sort relevance index into descending order
3. $X2 = KTCC(X1)$ // Using Kendall's Tau Correlation Coefficient, redundant features are eliminated
4. If ($X2 < \delta$) //Utilize ranking to determine the FS attributes. $< \delta$
5. FS = $X2$
```

F. enhanced k-anonymization's clustering technique

Clustering groups similar data together, a technique extensively employed in k-anonymization to protect sensitive data. Various clustering approaches have been introduced by researchers for anonymizing sensitive information. Clustering-based k-anonymization aims to enhance the quality of disclosed data by grouping similar instances and reducing information loss ([17]). This approach uses minimal generalizations within similar groups, ultimately creating anonymized data clusters with minimal information loss while limiting cluster size to fewer than k.

G. proposed design for safeguarding privacy

The primary objective of this system is to distribute all sensitive data within the SA domain into their respective equivalence classes while ensuring robust privacy protection. The system works with a dataset D, sensitive values denoted by s for sensitive attributes SA, and a user-defined anonymization level k as input parameters. Group creation adheres to the specified s and k criteria, and the method incorporates two stopping criteria.

Criteria 1: must (K ≤ s)

The number of equivalence classes formed is determined by the least frequent sensitive values, with each class encompassing potential SA-related values. This automatically fulfills the anonymization requirement, as the created groups match the value of s. Each equivalence class is associated with a specific number of sensitive SA values to meet the anonymization parameter, but if the count falls short of the possible values, adjustments are made.

Criteria 2: (K>s).

In this case, more equivalence classes were formed than initially possible. To meet the anonymization criteria, the cluster group created using equivalence is divided to include the minimum number of instances required.

The table below illustrates the process of counting all cluster groups formed with the given dataset while maintaining the anonymization criteria.

According to the second criteria, the split point is reached If there are too many cluster groups created the potential sensitive values. The following method may be used to divide a cluster.

$$\text{split} = \text{round}(k/s) \tag{6}$$

Each group consists of a list of the delicate situations that happen the least often values related to SA, calculated as (k/s). According to the initial criterion, if the count of formed cluster groups is fewer than the potential sensitive values, additional instances are introduced to fulfill this requirement. The inclusion of instances can be achieved using the following formula.

$$AI = \text{Split} - \text{Card}(ds) \tag{7}$$

The equation above introduces the term ”AI,” which signifies the extra instances required to be matched the possible sensitive value by being included into the cluster group. The clustering procedure for producing anonymized data is carried out under the guidance of the above described criteria.

The foremost step in this clustering-based anonymization approach involves selecting centroids. A detailed explanation of the centroid selection process follows.

H. selection of the clustering centroid

In the clustering method, selecting centroids is crucial and can be done through various methods, including user-defined criteria, random selection,

$$NOE = \frac{\text{Card(ds)}}{k/s} \tag{8}$$

or probabilistic estimation. The choice of centroid significantly impacts cluster quality and processing

In the above equation 8 NOE denotes number of equivalence class, k the anonymization attribute, and s denotes range of feasible SA values. The time.

In our proposed technique, the initial centroid is randomly selected but with an oversampling factor.

Centroids are continuously computed based on mean values and adjusted as needed until a cluster contains all data points. When there are more centroids than k, the last N/K centroids are retained within C, organized in ascending order.

The steps of this centroid selection algorithm are as follows: Algorithm 2: Pseudocode for that, the instances are randomly chosen to provide the centroids, as was covered in greater depth in preceding part. To generate the clusters, the chosen centroids are separated from the examples by determined ([18]). This distance calculation use Euclidean distance equation. The Equation for Euclidean Distance shown below.

Clustering by Choosing the Centroid

$$d(i,c) = \sqrt{(i_1 - c_1)^2 + (i_2 - c_2)^2 + \ldots + (i_n - c_n)^2}$$

```
Input: $D$ = Dataset with attributes $FS$, $C = 0$
Output: $c$ centroids
1. Initialize $D$ = Dataset with attributes $FS$, $C$ = NULL // $FS$ that is the output of Algorithm 1.
2. $C_1$ = Select an instance at random from Dataset
3. $C = C_1$ // $C_1$ assign to $C$
4. $n\psi = \phi(C_1)$ // $\phi(C_1)$ calculates some probability value
5. while $I \le \log(\psi)$ do
  a) for all instances $j$ in $D$
  b) for all $I$ centroids in $C$
  c) Calculate $D2_{j,k}$ // using Euclidean distance formula
  d) $C = C \cup$ {Instances with maximum probability}
6. end for
7. end for
8. end while
9. Sort $C$ in ascending order
10. Select final centroids for $N/K$ centroids in $C$
// $N$ is the length of the dataset and $K$ is the cluster size.
```

I. Cluster formation

Using the centroids chosen as outlined above, the instances are organized into clusters. The process begins by inputting the dataset, which includes selected features and anonymization parameters.

$$\text{d(i, c)} = \sqrt{(i1 - c1)2 + (i2 - c2)2 + \ldots + (in - cn)2} \tag{9}$$

In equation 9,” Instances” represents dataset items, and ”c” signifies the selected centroid. Instances are assigned to the nearest centroid based on calculated distances, and centroids are updated iteratively by calculating the mean of allocated instances until each instance is assigned to a cluster.

If the number of instances in a cluster falls below the anonymization value, it is augmented by merging an instance. Conversely, if the instance count exceeds the anonymization value, the cluster is split into a minimum of k instances. Ultimately, clusters are generalized to produce anonymous data, following the procedures outlined below.

Input: Dataset *D* with *FS* and *k* parameters

result: anonymous data set *ADS*

1: As cluster centroids, choose *K* randomly chosen instances.

2: For all of the dataset’s remaining occurrences

For each cluster centroid

Determine the separation between the cluster centroid and the instance.

The instance will be assigned to the nearest centroid. 3: Determine the centroid for each cluster as the average of all occurrences.

```
4: $Iter++$
5: loop steps 2 to 4 |previous centroid − current centroid| > $M$ and $Iter <$ MaxCnt
6: Within each cluster $|C| < k$
7: combine with the cluster is arranged such that $NCP\ (GcupG')$ is reduced;
8: within each cluster $|C| \geq 2k$
clusters into two groups, with possible $k$ instances
9: Make the clusters more inclusive.
```

With the help of this recommended technique, anonymizing healthcare data may be effectively completed with the least amount of information loss. Simultaneously, the approach maintains data quality to address the challenge of over-generalization. Ultimately, the proposed technique effectively preserves the privacy of sensitive information. To ascertain the method's effectiveness, an experimental analysis is conducted, and the results from this analysis are deliberated upon in the subsequent section.

4. Formal Modeling and Analysis Using High-Level Petri Nets (HLPNs)

A. Formal modeling and analysis using High-Level Petri Nets (HLPNs) for proposed approach

In this section, we model and analyses the pro- posed approach for feature attack-based privacy disclosure and provide its invalidation. We have used high-level Petri nets (HLPN) for the modelling and analysis of the proposed approach. In first transition, data set pass their feature attributes values to SU to calculate the relevance index for every feature attribute based on Symmetrical Uncertainty (SU) as given in (10).

$$R(SU) = \forall i2 \in x2,\ \forall i4 \in x4,\ \forall i5 \in x5|$$

$$i2[3] \in i2[4] \rightarrow i3[3] := i2[2] \wedge i3[3] := i12\ [4]$$

Table 2. Mapping of data types on places.

Types	Descriptions
PID	An integer type for describing Patient user ID
QI	Quasi-identifier values for Patient dataset
GQI	Generalized quasi-identifier values for data collector
SAx	Sensitive Attribute of instance x
SAy	Sensitive Attribute of instance y
δ	An integer type for Delta
RFA	Ranked Attribute value of feature At-tributes
C	An integer type for describing class ID
Cr	Random value of centroid
Cs	Sorted value of centroid
DFSA	Disclosed Feature Attributes value

Table 3. Types used in HLPN for proposed approach.

Types	Description
φ(QDS)	P (PID × QI × SAx × SAy × C)
$\varphi(\delta)$	$P\ (\delta)$
$\varphi(RIA)$	$P\ (PID \times QI \times RFA \times C)$
$\varphi(SD)$	$P\ (PID \times QI \times SRFAn \times SRFAn1 \times SRFAn2 \times C)$
$\varphi(FSA)$	$P\ (PID \times RFSAs \times C)$
$\varphi(Cr)$	$P\ (C_r)$
$\varphi(Cn)$	$P\ (C_n)$
$\varphi(SC)$	$P\ (C_s)$
$\varphi(FSD)$	$P\ (PID \times GQI \times RFSAs \times C)$
$\varphi(BK)$	$P\ (QI)$
φ(FS − Dis)	P (PID × DFSA)

Relevance Indexes feature attributes are arranged into descending order and assign to place SD in transitions R(DO). In (12) the redundant attributes are removed using Kendall's Tau Correlation Co- efficient (KTCC) ranking model. It Utilize Ranked Attribute value of feature Attributes is determined by the condition of FS attributes ¡δ.

$$R(DO) = \forall i4 \in x4,\ i5 \in x5|\ \wedge\ x3 := x3 \cup \{i3[3]\} \qquad (10)$$

$$i5[3] := \mathrm{Dsrt}(i3[3])\ x5 := x5 \cup \{i5[3]\} \qquad (11)$$

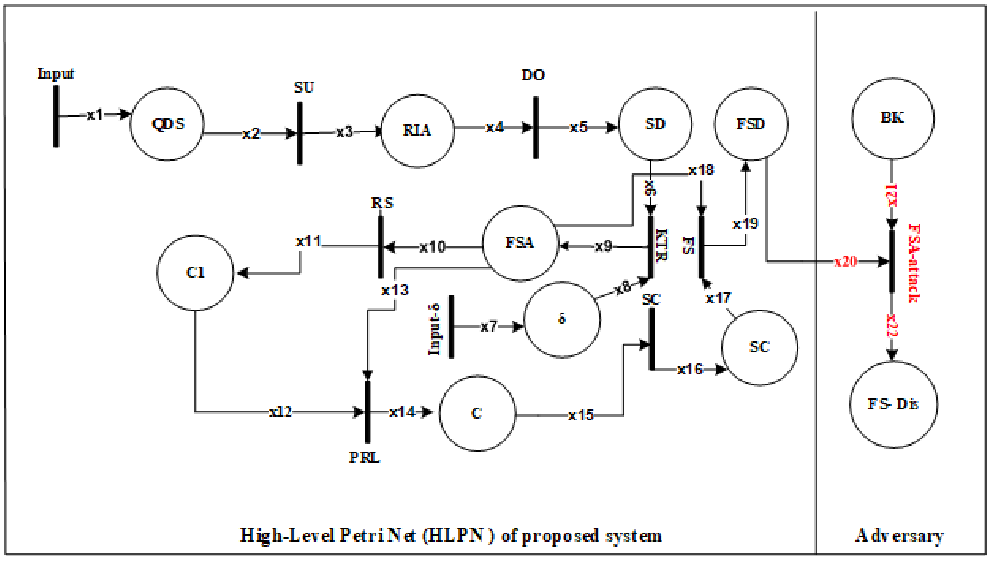


Figure 2. High level Petri Net for Proposed Approach.

$$R(KTCC) =\forall i8 \in x8, \forall i6 \in x6, \forall i9 \in x9 \mid i9[1]:=i6[1] \wedge i9[2]:=\sqrt{\frac{Concrd(i6[4],i4[5])-Discrd(i6[4],i6[5])}{(i6[3]-i6[4])-(i6[3]-i6[5])}}< i8[1] \quad x9' := x9 \cup \{i9[1],i9[2]\} \tag{12}$$

in place C. These values are then sorted in transition R (SC) . In (16) final centroids are selected and resulted data is saved at place FSD.

$$R(RS)=\forall i1 \in x10, \forall i1 \in x11 \mid i11[1]:=Rndm(i10[2]) \wedge x11:= x11 \cup \{i11[1]\} \tag{13}$$

In transition R (RS), an instance from place FSA is randomly selected. In (14) some probability value is calculated for all instances in FSA and randomly selected centroids. Using Euclidean distance formula in function Eucld() and

$$R(PRL) = \forall i12 \in x12, \forall i13 \in x13, \forall i14 \in x14 \mid i14[1]:=MaxprbEucldprb(i12[1],i13[2] \wedge x14' := x14 \cup \{i14[1]\} \tag{14}$$

choosing centroid with maximum probability value resultant value I s saved performance. These processes are discussed in ([1]). The system's efficiency is established when both minimal information loss and sustained data quality are achieved. For conducting the analysis, a dataset is procured. The next section goes

$$R(SC) = \forall i15 \in x15, \forall i16 \in x16 \mid (i16[1] := i15[1]) \wedge (i16[2] := i15[2]) \wedge x16' := x16 \cup \{i16[1],i16[2]\} \tag{15}$$

$$R(FS) = \forall i17 \in x17,\ \forall i18 \in x18,\ = i18[2] \cup i19[3] = i18[3]\ \forall i19 \in x19 \mid i20[1] := i19[1] \wedge i20[2] = i19[2] \wedge i20[3] := i19[3] \wedge i20[4] = i19[4]$$

$$i19[4] := i17[1] \cup i19[4] \cup i19[1] = i18[1] \cup i19[2] \wedge x19:= x19 \cup \{i19[1], i19[2], i19[3], i19[4] \tag{16}$$

Table 4. Origional Dataset.

preg	plas	pres	skin	insu	mass	pedi	age	class
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0

two crucial processes that determine the system's

$$R(FSA\text{-}attacks) = \forall i20 \in x20,\ \forall i21 \in x21,\ \forall i22 \in x22 | \ (i21[1] \cup i20[2]) \nvDash i22[1] \vee (i21[1] \cup i20[3]) \nvDash i22[2] \quad (17)$$

The last transition FSA-attacks is the main attack transitions that we have shown on proposed approach. In (17), it is shown that the privacy attacks FSA are effectively mitigated as we can preserve privacy using generalized unique values against the above-mentioned privacy attack. The minimum and maximum value of a particular attribute in group results in hard identification of a record and privacy can be preserved. We are dealing with multiple attacks like similarity attacks, Membership attack, and attribute disclosure attack.

5. Experimental Results

The proposed system is implemented using the Python platform through Anaconda Navigator and Jupiter Notebook. The system's execution takes place on a machine equipped with a 2.66 GHz Intel IV processor and 4 GB of RAM. This machine operates on the Microsoft Windows 10 Professional Edition.

This suggested system's experimental study in- tends to evaluate how well it produces anonymized data from uncooked clinical data. Clustering and the generalization approach used for anonymization are into further detail on the dataset that was utilized for the study.

A. Diabetes dataset for Pima Indians

To perform the study, this dataset was taken from the UCI Machine Learning Repository. 768 occurrences and 8 characteristics are included in the dataset. The outcomes for patients are expressed using the values 0 and 1.

While 0 indicates absence of diabetes in a patient, value 1 indicates the presence of diabetes. The age, Pedi, mass, skin, pres, plas, and pregnancy features of the dataset are included. Class 1 has about 268 cases, while class 0 has more than 500 occurrences.

B. Wisconsin cancer data set

The dataset used in this study is freely available at the following link: Breast Cancer Wisconsin Dataset. It comprises 699 instances and 11 attributes, categorized into benign and malignant types. There are 458 benign and 241 malignant instances in the dataset. Notable attributes include uniform cell shape, naked nuclei, normal nucleoli, neutral chromatin, and normal cell size. Approximately 16 cases were removed due to missing values, resulting in a foundation for subsequent analyses.

Table 5. *Ranking of Attributes.*

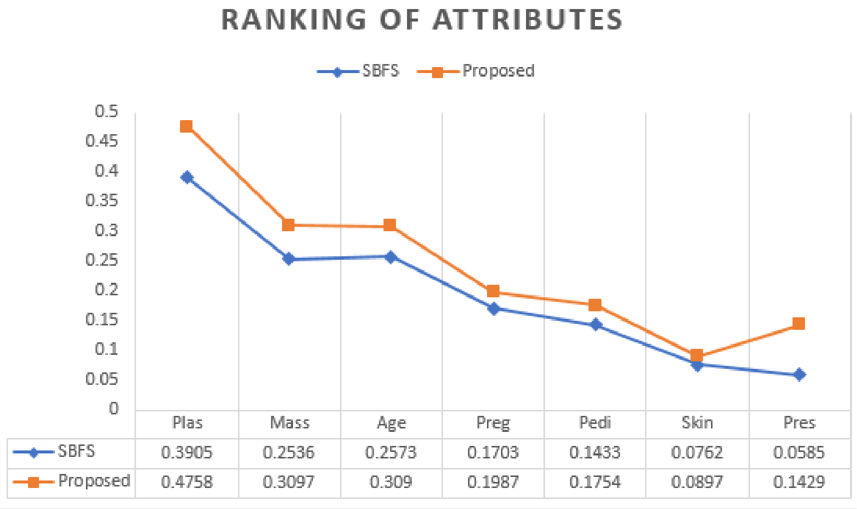
Attributes	Kendal Tau	sbfs	pca	lda
plas	0.3905	0.4758	0.3991	0.3999
mass	0.2536	0.3097	0.2854	0.2868
age	0.2573	0.3090	0.2375	0.2485
preg	0.1703	0.1987	0.2156	0.2182
pedi	0.1433	0.1754	0.1690	0.1768
insu	0.1192	0.0665	0.1345	0.1365
skin	0.0762	0.0897	0.0850	0.0873
pres	0.0585	0.1429	0.0556	0.0587

Table 6. *Anonymized table.*

Preg	Plas	Pres	Skin	Insu	Mass	Pedi	Age	Class
3-8	148	72-76	26-35	0-0	33.6	0.276-0.627	23	1
3-8	109	72-76	26-35	0-0	36.0	0.276-0.627	21-26	0
3-8	84	72-76	26-35	0-0	38.3	0.276-0.627	23	0
3-8	97	72-76	26-35	0-0	35.6	0.276-0.627	21-26	1
3-8	84	72-76	26-35	0-0	37.2	0.276-0.627	21-26	0
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0

These observations serve as input for the Utility- Preserving Anonymization (UPA) process, which consists of two primary stages: feature selection and K-anonymization-based clustering. Initially, the dataset is processed through the KTFS algorithm, utilizing the Sequential Unsupervised (SU) approach for essential feature selection. Subsequently, the Kendall's Tau correlation coefficient is applied to rank attributes, aiding in the removal of redundant characteristics. The provided information presents attribute rankings generated by the suggested algorithm, as well as comparisons with Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in Table 5. Figure 3 shows the ranking of attributes using the proposed method and SBFS method, the proposed method significance shows better ranking than the previous method used.

The KTFS algorithm plays a vital role in identifying attributes with significant impact on classification.



	Plas	Mass	Age	Preg	Pedi	Skin	Pres
SBFS	0.3905	0.2536	0.2573	0.1703	0.1433	0.0762	0.0585
Proposed	0.4758	0.3097	0.309	0.1987	0.1754	0.0897	0.1429

Figure 3. Ranking of attributes.

Rather than anonymizing all attributes, only those minimally affecting classification undergo anonymization, reducing the extent of generalization and alleviating excessive data loss. The attribute with the lowest rank score, considering a threshold of 0.2, becomes a candidate for anonymization. In this case, attributes like preg, pedi, pres, skin, and insu are anonymized.

However, age remains a potential privacy vulnerability, as knowing a patient's age can lead to identification. To address this, unique age values are identified and replaced with a range, such as ”21-26,” minimizing the risk of easy identification and preserving privacy.

The selected features identified by the KTFS algorithm are then utilized in the subsequent K- anonymization-based clustering process. In this phase, these features are treated as instances. The clustering procedure begins by randomly selecting a centroid and assigning instances to the nearest centroid based on distance calculations. This iterative process results in the formation of clusters, which continue to expand until all instances are allocated to a cluster. The resulting clusters are generalized to provide data that is anonymous. The supplied table shows the results after anonymization. VI.

The provided table summarizes the results of our proposed system, highlighting attributes such as pregnancy, pregnancies, skin, insu, and pedi with the lowest Kendall's Tau correlation coefficients. These attributes undergo selective generalization, minimizing information loss while maintaining data quality. To assess our approach's effectiveness, we employ specific performance metrics, which we will thoroughly explain in the following section.

C. indicators of performance

The system's performance is assessed using performance metrics, often known as statistical indicators. These parameters are selected in accordance with the specific analysis procedure. For instance, measures from a confusion matrix including recall, accuracy, specificity, and sensitivity are used in machine learning.

D. measure for discernibility

Utility loss is measured using discernibility metric. Another way to put it is the dimension equal to 50 or greater than 50 then discernibility metric provides a better result than the previous approach.

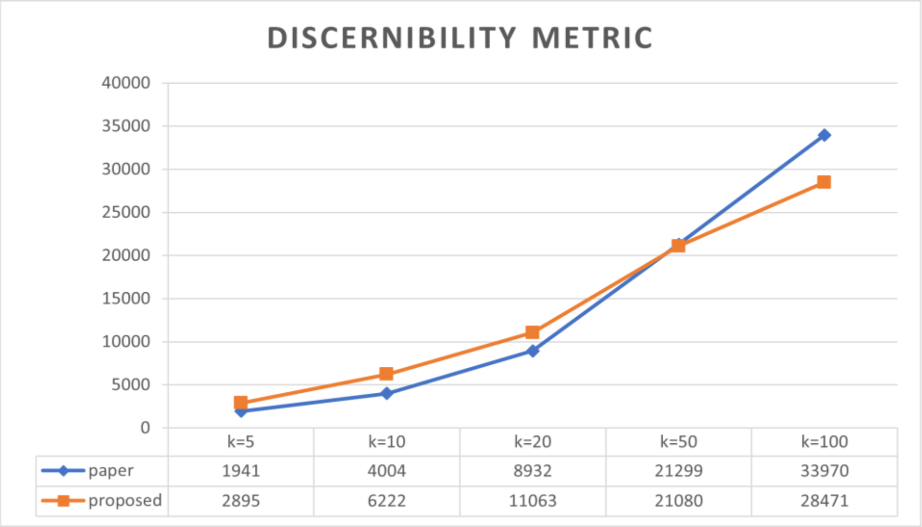


Figure 4. Discernibility Metric.

E. KL-divergence metric

The Kull back Liebler divergence metric is a measurement of the difference between the original distribution and the distribution obtained after anonymization. Below is shown a mathematical illustration of the Kull back Liebler divergence.

$$\text{DM}= \sum_{C\in D} |C|2 \qquad (18)$$

The ”C” equivalence class is constructed using the procedure in the above equation, and ”D” stands for the dataset with privacy maintained. For less utility loss, the goal is to diminish the value of DM. More utility loss is indicated by a larger value of DM obtained.

In Figure 4 using discernibility equation 18, we calculate the discernibility metric against the multiple values of k, it is seemed that when the value of k remains less than 50 then the proposed method does not give a better result, whenever the value of k is

$$D_{\text{KL}}\ (\text{P}\|\text{Q}) = \sum_x P(x)log\left(\frac{Q(x)}{P(x)}\right)\text{P(x)log Q(x) x P(x)} \qquad (19)$$

Here, p(x) and q(x) represent two distributions whose computation of the divergence is required. The KL- divergence metric should yield a smaller value when there is minimal change in the distribution values. In Figure 5, after performing the experiments it is shown that the proposed approach gives better results against the multiple values of k’s compared to the existing approaches. When using KL divergence, it’s important to be cautious of issues like dealing with zeros in the distributions and selecting the appropriate base for the logarithm (common choices are natural logarithm and logarithm base 2).

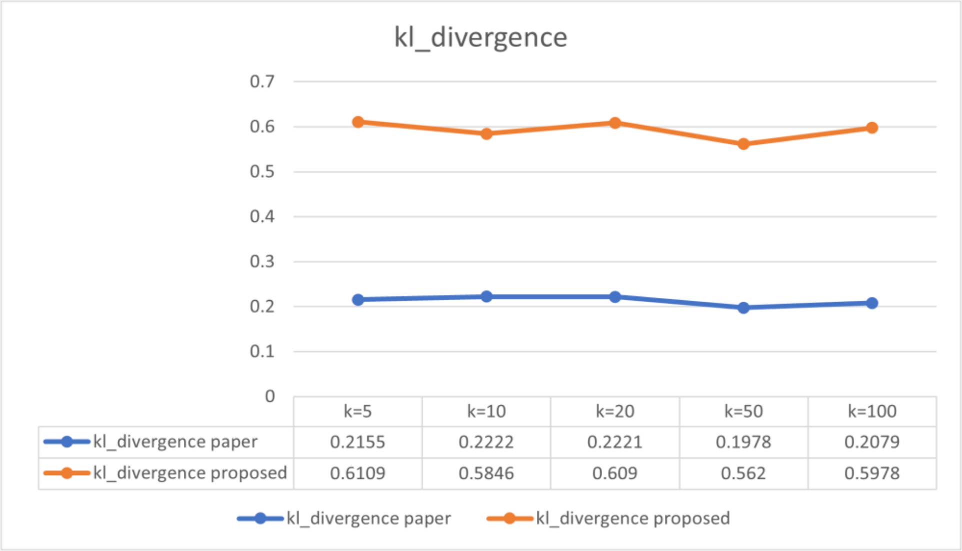


Figure 5. KL divergence metric.

F. Size of the average equivalence class

An equivalence relation on splits a set into disjoint subsets called equivalence classes. Each element in a given equivalence class is related to all other elements in the same class and unrelated to elements in other classes([2]).

The assessment of utility loss also considers the average size of equivalence classes. A lower average equivalence class size is indicative of reduced utility

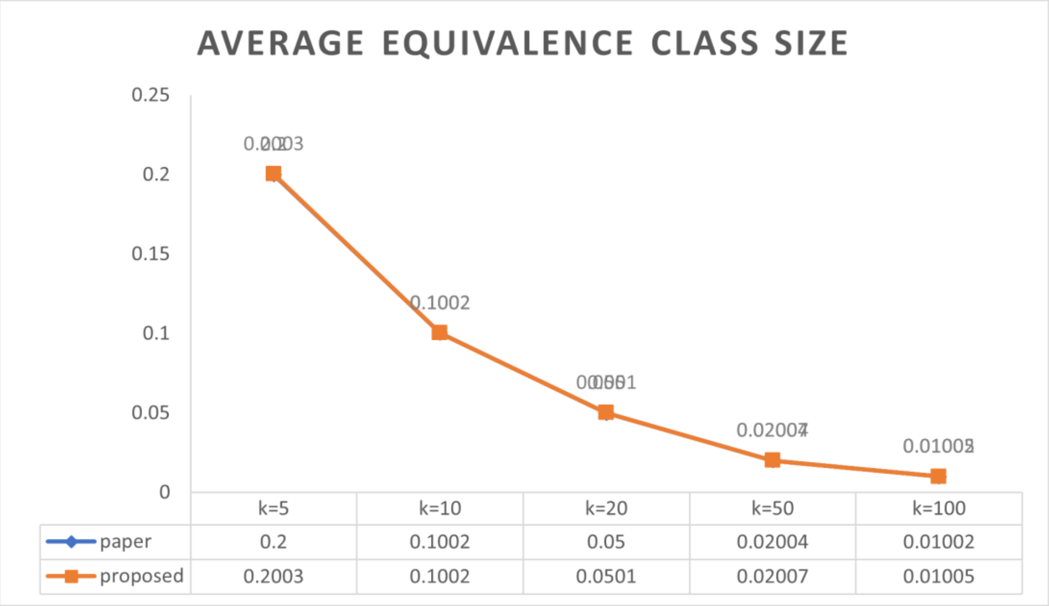


Figure 6*. Average equivalence class size.*

G. Silhouette Score

The silhouette score is a metric used to evaluate the quality of clusters in a clustering algorithm, such as k-means clustering. It measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette score ranges from -1 to +1, where:

A high silhouette score indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters.

A score close to 0 indicates that the object is on loss.

$$A(AVG) = \frac{\frac{D}{NOE}}{\mathrm{k}} \tag{20}$$

or very close to the decision boundary between two neighboring clusters.

In the equation provided above, ”NOE” represents the count of equivalence classes, ”D” stands for the dataset, and ”k” represents the anonymization parameter.

In Figure 6, The proposed method’s output is contrasted with that of the current method. there is a little bit of difference between the proposed and existing approach. An equivalence relation on splits a set into disjoint subsets called equivalence classes. Each element in a given equivalence class is related to all other elements in the same class and unrelated to elements in other classes.

A negative score indicates that the object might have been assigned to the wrong cluster.

$$\text{Silhouette Score} = \frac{1}{N}\sum_{i=1}^{N} s(i) \tag{21}$$

According to the Figure 7 If the silhouette score rises alongside the increasing value of ’k,’ it could appear paradoxical. The silhouette score gauges the extent of cluster separation, with higher values indicating more clearly defined and distinct clusters. The comparative research is conducted to verify the efficacy of the suggested method for privacy protection in comparison to an previous strategy, namely WFS(Wrapper Feature Selection with Mondrian). ([7]). Evaluation makes use of two datasets and classification methods including Naive Bayes and C4.5 classifier. It gives an explanation of the frequently used comparison method. The results of the comparison analysis are presented in the provided table.

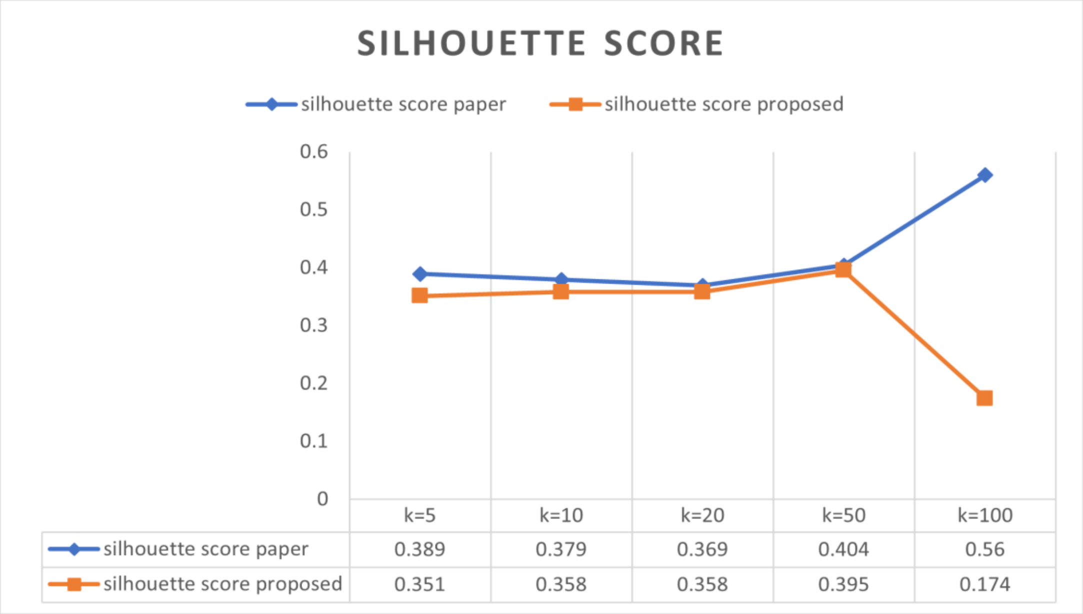


Figure 7. Silhouette Score.

VII. The primary concern of K-anonymity is the risk of record linkage attacks, where Quasi-Identifiers (QI) are linked to external databases, compromising privacy. Using feature selection enhances privacy, as indicated by our findings, which highlight the positive impact of clustering-based generalization and anonymization on performance.

Our proposed model customizes datasets for analysis, streamlining the data publication process. Anonymizing real-time datasets with extensive QI sets and diverse features can be challenging. However, by anonymizing only a subset of QI characteristics, we reduce the need for extensive generalization, lowering processing costs. This offers a significant opportunity for improving the anonymization process, especially with the growth of high-dimensional datasets. We conducted a comparative analysis (Table 7) between the existing UPA approach and our proposed method, visually represented in Figures 8 and 9. Our goal was to assess the proposed system's classification accuracy across different degrees of anonymization denoted by "k." The results consistently show that our method outperforms the current approach at every level of anonymization (Table 7), confirming its effectiveness in safeguarding sensitive data.

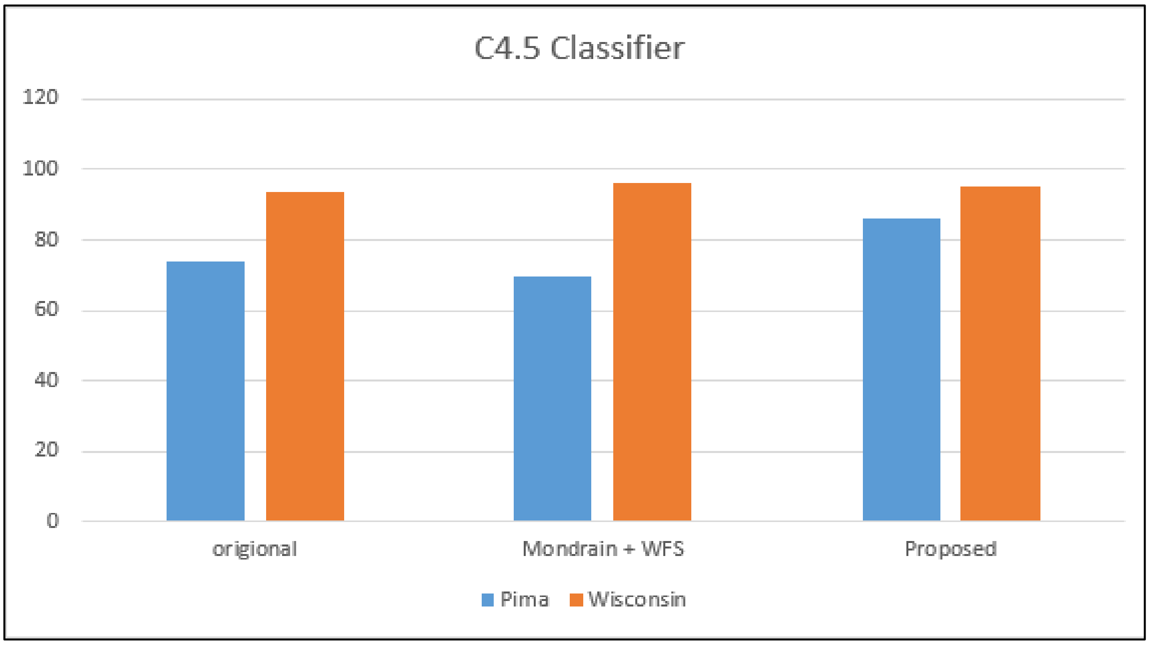


Figure 8. C4.5 classifier.

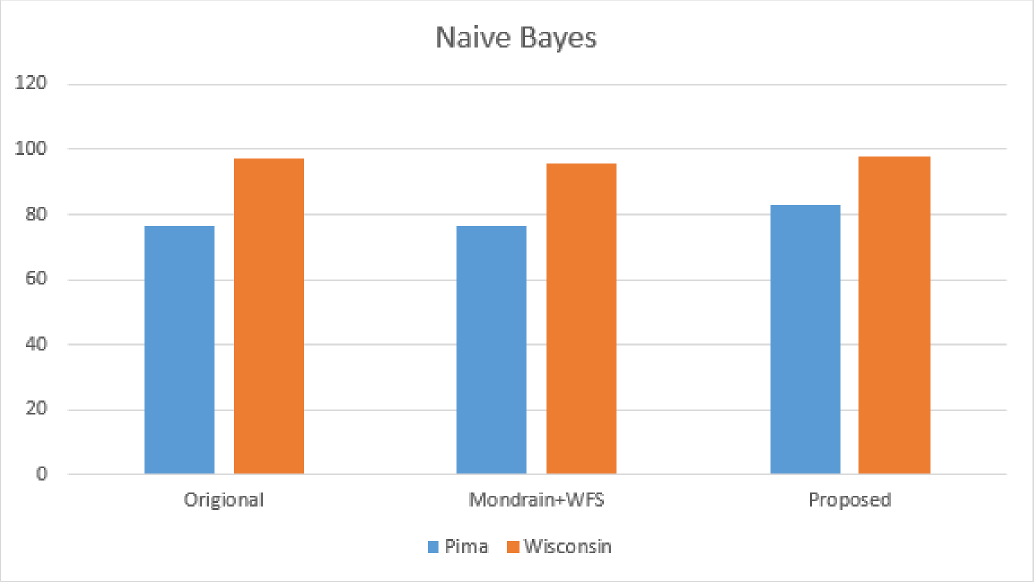


Figure 9. Naive Bayes classifier.

Regarding information loss, it measures the re- duction in valuable data during processing. In data anonymization, it specifically refers to the loss of original data values or characteristics while preserving utility. Table 8 This entails assessing the performance and efficacy of the current methodologies via the evaluation of quality metrics and datasets. When compared to the earlier strategies mentioned in the linked study, the current offered strategy produces better results. The suggested approach, which highlights its improved performance, also exhibits lower information loss when compared to other strategies. In Figure 10, we illustrate the Age attribute's unique values and their significance across various cluster sizes represented by k. Larger k values correspond to more clusters, resulting in a greater number of unique values, as seen with k=5. Conversely, when k=100, fewer clusters lead to fewer unique values. The Age attribute is considered vulnerable because knowledge of a patient's age can potentially reveal other patient information, increasing the risk of privacy breaches. To mitigate this risk, we propose identifying unique age values within specific groups and generalizing them to a range between the minimum and maximum values, excluding duplicates. This approach makes it harder to pinpoint individual identities within a group of identical age values.

Table 7. Comparison of Classifications.

Dataset	Classification Algorithm	Original	Mondrian+WFS	Proposed
Pima diabetes	C4.5 classifier	73.83	69.4	86.13
Wisconsin cancer	C4.5 classifier	93.41	96.0	95.0
Pima diabetes	Na¨ıve Bayes	76.32	76.56	83.16
Wisconsin cancer	Na¨ıve Bayes	97.36	95.90	97.85

Table 8. Information Loss Comparison.

Method	Information Loss
Onesimu et al ([19])	1.9
Li et al ([11])	1.85
Rodr´ıguez et al ([21])	1.83
Srijayanthi, Sethukarasi ([26])	1.5

Method	Information Loss
Proposed	0.51

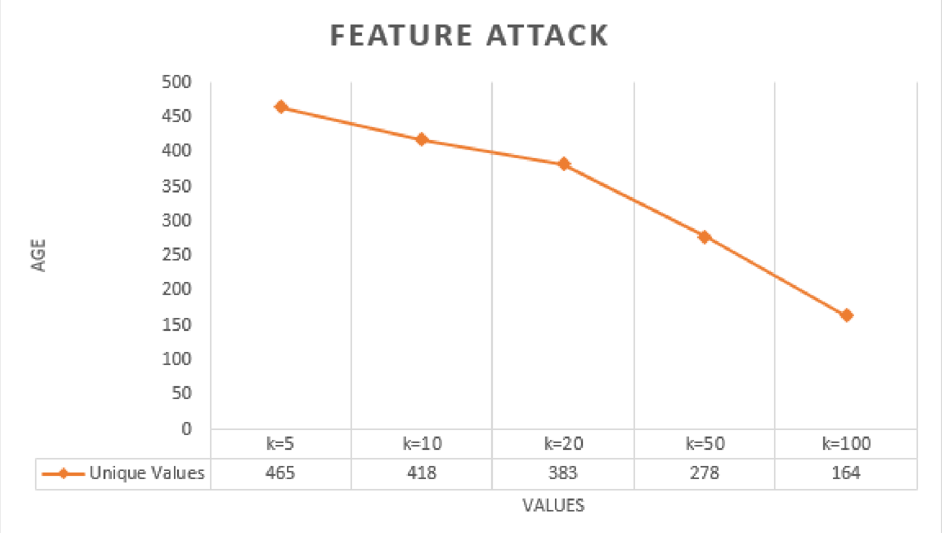


Figure 10. feature attack.

The significance values associated with different k values indicate vulnerable records that pose privacy risks. Future work will focus on reducing these vulnerabilities.

6. Conclusion

This study introduces a novel k-anonymity approach that reduces the number of traits through feature selection in clustering, maintaining privacy. We demonstrate its superiority over existing methods in terms of information loss, effectiveness, and scalability using real-world datasets. Our technique enhances classification, accuracy and privacy by selectively anonymizing quasi-attributes. Feature selection aids in preserving utility by preventing overgeneralization. We address challenges in safeguarding large datasets, emphasizing scalable privacy preservation in cloud systems. Our method prioritizes data privacy, minimizes data loss, and addresses attribute identification challenges through entity linking and dependency parsing. Equivalence class sizes are assigned using statistical tables or knowledge bases. Our future work will refine these challenges and pave the way for further inquiries. We will explore improving clustering and feature selection techniques for diverse data types and complexities, assess real-time processing applications, and investigate scalability with large datasets.

Abbreviations

SU	Symmetrical Uncertainty
UPA	Utility Preserved Anonymization
KTFS	Kendall’s Tau Based Feature Selection
KTCC	Kendall’s Tau Correlation Coefficient
KL	Kull back Liebler

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Afsoon Abbasi and Behnaz Mohammadi. A clustering-based anonymization approach for privacy-preserving in the healthcare cloud. *Concurrency and Computation: Practice and Experience*, 34(1): e6487, 2022.

[2] Lanny Fields, Sharon A Hobbie-Reeve, Bar- bara J Adams, and Kenneth F Reeve. Effects of training directionality and class size on equivalence class formation by adults. *The Psychological Record*, 49(4): 703-723, 1999.

[3] Mohamed R Fouad, Khaled Elbassioni, and Elisa Bertino. A supermodularity-based differential privacy preserving algorithm for data anonymization. *IEEE Transactions on Knowledge and Data Engineering*, 26(7): 1591-1601, 2014.

[4] Benjamin CM Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)*, 42(4): 1-53, 2010.

[5] Esther Gachanga, Michael Kimwele, and Lawrence Nderu. Feature based data anonymization for high dimensional data. 2019.

[6] Andreas Holzinger and Igor Jurisica. Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. *Inter- active knowledge discovery and data mining in biomedical informatics: state-of-the-art and future challenges*, pages 1-18, 2014.

[7] Yasser Jafer, Stan Matwin, and Marina Sokolova. Task oriented privacy preserving data publishing using feature selection. In *Advances in Artificial Intelligence: 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montre´al, QC, Canada, May 6-9, 2014. Proceedings 27*, pages 143-154. Springer, 2014.

[8] Wenjun Ke, Chunxue Wu, Yan Wu, and Neal N Xiong. A new filter feature selection based on criteria fusion for gene microarray data. *IEEE Access*, 6: 61065-61076, 2018.

[9] Florian Kohlmayer, Fabian Prasser, and Klaus A Kuhn. The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. *Journal of biomedical informatics*, 58: 37-48, 2015.

[10] Hyukki Lee, Soohyung Kim, Jong Wook Kim, and Yon Dohn Chung. Utility-preserving anonymization for health data publishing. *BMC medical informatics and decision making*, 17(1): 1-12, 2017.

[11] Hongtao Li, Feng Guo, Wenyin Zhang, Jie Wang, and Jinsheng Xing. (a, k)-anonymous scheme for privacy-preserving data collection in iot-based healthcare services systems. *Journal of Medical Systems*, 42: 1-9, 2018.

[12] Tong Li, Chongzhi Gao, Liaoliang Jiang, Witold Pedrycz, and Jian Shen. Publicly verifiable privacy-preserving aggregation and its application in iot. *Journal of Network and Computer Applications*, 126: 39-44, 2019.

[13] Hoon Wei Lim, Geong Sen Poh, Jia Xu, and Varsha Chittawar. Privacy-preserving integration and sharing of datasets. *IEEE Transactions on Information Forensics and Security*, 15: 564- 577, 2019.

[14] Wen-Yang Lin, Duen-Chuan Yang, and Jie- Teng Wang. Privacy preserving data anonymization of spontaneous ade reporting system dataset. In *Proceedings of the ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics*, pages 2-2, 2015.

[15] Brijesh B Mehta and Udai Pratap Rao. Im- proved l-diversity: scalable anonymization approach for privacy preserving big data publishing. *Journal of King Saud University-Computer and Information Sciences*, 34(4): 1423-1430, 2022.

[16] Shinya Miyakawa, Nobuyuki Saji, and Takuya Mori. Location l-diversity against multifarious inference attacks. In *2012 IEEE/IPSJ 12th International Symposium on Applications and the Internet*, pages 1-10. IEEE, 2012.

[17] J Jesu Vedha Nayahi and V Kavitha. An efficient clustering for anonymizing data and protecting sensitive labels. *International Jour- nal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 23(05): 685-714, 2015.

[18] J Jesu Vedha Nayahi and V Kavitha. Privacy and utility preserving data clustering for data anonymization and distribution on hadoop. *Future Generation Computer Systems*, 74: 393- 408, 2017.

[19] J Andrew Onesimu, J Karthikeyan, and Yuichi Sei. An efficient clustering-based anonymization scheme for privacy-preserving data collection in iot based healthcare services. *Peer-to- Peer Networking and Applications*, 14: 1629- 1649, 2021.

[20] Chunhui Piao, Liping Liu, Yajuan Shi, Xue- hong Jiang, and Ning Song. Clustering-based privacy preserving anonymity approach for table data sharing. *International Journal of System Assurance Engineering and Manage- ment*, 11: 768-773, 2020.

[21] Ana Rodriguez-Hoyos, Jose Estrada-Jimenez, David Rebollo-Monedero, Javier Parra-Arnau, and Jordi Forne´. Does *k*-anonymous microaggregation affect machine-learned macrotrends? *IEEE access*, 6: 28258-28277, 2018.

[22] Yuichi Sei, Hiroshi Okumura, Takao Take- nouchi, and Akihiko Ohsuga. Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness. *IEEE transactions on dependable and secure computing*, 16(4): 580-593, 2017.

[23] Daniel Sierra-Sosa, Begonya Garcia-Zapirain, Cristian Castillo, Ibon Oleagordia, Roberto Nuno-Solinis, Maider Urtaran-Laresgoiti, and Adel Elmaghraby. Scalable healthcare assessment for diabetic patients using deep learning on multiple gpus. *IEEE transactions on indus- trial informatics*, 15(10): 5682-5689, 2019.

[24] Djordje Slijepcˇevic´, Maximilian Henzl, Lukas Daniel Klausner, Tobias Dam, Peter Kieseberg, and Matthias Zeppelzauer. k- anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Computers & Security*, 111: 102488, 2021.

[25] Jordi Soria-Comas, Josep Domingo-Ferrer, David Sanchez, and Sergio Martinez. t- closeness through microaggregation: Strict privacy with enhanced utility preservation. *IEEE Transactions on Knowledge and Data Engineering*, 27(11): 3098-3110, 2015.

[26] S Srijayanthi and T Sethukarasi. Design of privacy preserving model based on clustering involved anonymization along with feature selection. *Computers & Security*, 126: 103027, 2023.

[27] Xiaoxun Sun, Hua Wang, Jiuyong Li, and Traian Marius Truta. Enhanced p-sensitive k-anonymity models for privacy preserving data publishing. *Transactions on Data Privacy*, 1(2): 53-66, 2008.

[28] Kok-Seng Wong, Nguyen Anh Tu, Dinh- Mao Bui, Shih Yin Ooi, and Myung Ho Kim. Privacy-preserving collaborative data anonymization with sensitive quasi-identifiers. In *2019 12th CMI Conference on Cybersecurity and Privacy (CMI)*, pages 1-6. IEEE, 2019.

[29] Yinghui Zhang, Robert H Deng, Gang Han, and Dong Zheng. Secure smart health with privacy-aware aggregate authentication and access control in internet of things. *Journal of Network and Computer Applications*, 123: 89- 100, 2018.