ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

An Efficient Model for Publishing Microdata with Multiple Sensitive Attributes

Surapon Riyana[1], Kittikorn Sasujit[2], Nigran Homdoung[3] and Noppamas Riyana[4]

ABSTRACT

The purpose of this work is to propose an anonymization model. It is used to address privacy violation issues in datasets that have multiple sensitive attributes. To achieve privacy preservation constraints and maintain data utilities, the sensitive attributes of datasets are grouped to be nominal and continuous attributes. With the nominal sensitive attribute, the data utility and privacy are maintained by the confidence of data re-identification. With another data type, the continuous data, the data utility and privacy are maintained by the data bounding. The proposed model is evaluated by using extensive experiments. The experimental results indicate that the proposed model is more effective and efficient than the compared models. Moreover, the datasets satisfy the privacy preservation constraints of the proposed model, which can guarantee the confidence and bounding of data re-identification.

Article information:

Keywords: Privacy preservation models, Privacy violation Issues, Data anonymizations, Multiple sensitive attributes, Confidence of data re-identifications, Bounding of data re-identification

Article history:
Received: March 13, 2025
Revised: June 10, 2025
Accepted: July 29, 2025
Published: August 9, 2025 (Online)

DOI: 10.37936/ecti-cit.2025193.261156

1. INTRODUCTION

The issue of violating the user's privacy in datasets is serious when datasets are released for utilizing outside the scope of the data-collecting organizations. To address this issue, there are the privacy preservation models to be proposed such as k-Anonymity [1][2][3], l-Diversity [4][5][6][7], t-Closeness [8][9], Anatomy [10][11][12][13], (k, e)-Anonymous [14][15], (α, k)-Anonymity [16], LKC-Privacy [17][18], aggregate query frameworks [19], differential privacy [20][21][22], privacy preservation in video streaming [31], and privacy in image steganography-based copyright and image trading systems [32].

An example of privacy preservation is based on k-Anonymity [1]. We assume that Table 1 is a specified dataset that will be released for utilizing outside the scope of data collecting organizations. We suppose that the value of k is set to be 2, i.e., k = 2. For privacy preservation, the attributes of datasets are grouped to be quasi-identifier attributes (i.e., Age and Sex attributes) and a sensitive attribute (i.e., Disease attribute). Moreover, the unique quasi-identifier values are generalized by their less specific values to be at least 2 indistinguishable tuples. Therefore, a data version of Table 1 is satisfied by 2-Anonymity constraints to be shown in Table 2.

Aside from data generalization, the datasets can be satisfied by k-Anonymity constraints through data suppression. For example, let Table 1 be the specified dataset. We give the value of k to be 2. Therefore, a data version of Table 1 is satisfied by 2-Anonymity constraints that are based on using data suppression, it is shown in Table 3.

Table 1: *An example of raw datasets.*

♯	Age	Sex	Disease
t_1	52	Male	Flu
t_2	51	Male	Fever
t_3	52	Male	HIV
t_4	54	Male	Cancer
t_5	56	Female	Fever
t_6	55	Female	Fever
t_7	55	Female	HIV

Table 2: *A data version of Table 1 is satisfied by 2-Anonymity from using data generalization.*

♯	Age	Sex	Disease
t_1	50-51	Male	Flu
t_2	50-51	Male	Fever
t_3	52-54	Male	HIV
t_4	52-54	Male	Cancer
t_5	55-56	Female	Fever
t_6	55-56	Female	Fever
t_7	55-56	Female	HIV

[1,2,3,4]The authors are with the School of Renewable Energy, Maejo University, Sansai, Chiangmai, Thailand, Email: surapon_r@mju.ac.th, k.sasujit@yahoo.com, nigranghd@gmail.com and noppamas@mju.ac.th
[1]Corresponding author: surapon_r@mju.ac.th

With Tables 2 and 3, we can see that they lose some meaning of data utility. For example, only 55 as the age of users in Table 3 can be retained. Moreover, we observe that Table 2 has some counterfeit query conditions. An example of counterfeit query conditions in Table 2 is the query condition for a 53-year-old that is shown in QUERY 1.

QUERY 1: *SELECT Disease FROM Table 2 WHERE Age = 53;*

Furthermore, when we only focus on the tuples t_5, t_6, and t_7 of Table 2, we cannot ensure how many tuples are 55 years old. Therefore, we can conclude that although data suppression and data generalization can be used to address privacy violation issues in datasets, they still have data utility issues that must be improved. Moreover, in [4], the authors further demonstrate that k-Anonymity has a serious vulnerability that must be considered. That is, the privacy violation issues in datasets are when the sensitive value is non-variety. For example, with Table 4, we can see that although the adversary cannot ensure that a tuple is what the target user's profile tuple is in this table, he/she can infer that the disease of the target user in Table 4 is HIV.

Table 3: *A data version of Table 1 is satisfied by 2-Anonymity from using data suppression.*

♯	Age	Sex	Disease
t_1		Male	Flu
t_2		Male	Fever
t_3		Male	HIV
t_4			Cancer
t_5			Fever
t_6	55	Female	Fever
t_7	55	Female	HIV

Table 4: *A data version of Table 1 is satisfied by 2-Anonymity from using data suppression.*

♯	Age	Sex	Disease
...	...	...	...
t_3	52-54	Male	HIV
t_7	55-56	Female	HIV
...	...	...	...

Table 5: *A partitioned data version of Table 1, where $l = 2$.*

♯	**Age**	**Sex**	**Disease**	**PID**
t_1	52	Male	Flu	1
t_2	51	Male	Fever	
t_3	52	Male	HIV	2
t_4	54	Male	Cancer	
t_5	56	Female	Fever	3
t_6	55	Female	Fever	
t_7	55	Female	HIV	

To address privacy violation issues in datasets by considering the non-variety sensitive value, l-Diversity [4] is proposed. With this privacy preservation model, aside from distorting the unique quasi-identifier values, the number of distinct sensitive values is further considered in privacy preservation constraints. That is, every group of distorted quasi-identifier values must be related to at least l distinct sensitive values. For this reason, the datasets are satisfied by l-Diversity constraints, which can guarantee that all possible query conditions always have at least l distinct sensitive values that are satisfied. Thus, they seemingly have no concerns about privacy violation issues. Unfortunately, in [8], the authors show that l-Diversity still has a serious vulnerability that must be improved, i.e., the privacy violation issue from considering the distance of sensitive values. For example, the specified group of the distorted quasi-identifier values only relates to Gonorrhoea, Chlamydia, Syphilis, and HIV. In this situation, the adversary can infer that the data owner of these tuples has venereal diseases. To rid this vulnerability of l-Diversity, t-Closeness [8] is proposed. The privacy preservation idea of t-Closeness is that the sensitive values of each distorted quasi-identifier group must have a distance that is at least t. In addition, aside from data suppression and generalization, the datasets can be satisfied by k-Anonymity, l-Diversity, and t-Closeness constraints from using the added noise values [25][26]. However, these privacy preservation models have data utility issues that must be addressed. To address these issues, Anatomy [10][11][12][13] is proposed. With this privacy preservation model, datasets can satisfy privacy preservation constraints after their tuples are partitioned by l and anonymized to be the quasi-identifier and sensitive tables.

Table 6: *A data version of the quasi-identifier table of Table 1, where $l = 2$.*

♯	**Age**	**Sex**	**PID**
t_1	52	Male	1
t_2	51	Male	
t_3	52	Male	2
t_4	54	Male	
t_5	56	Female	3
t_6	55	Female	
t_7	55	Female	

Table 7: *A data version of the sensitive table of Table 1, where $l = 2$.*

♯	**Disease**	**PID**
t_1	Flu	1
t_2	Fever	
t_3	HIV	2
t_4	Cancer	
t_5	Fever	3
t_6	Fever	
t_7	HIV	

Table 8: *An example of datasets that have multiple sensitive attributes.*

♯	Age	Sex	Disease	Income
t_1	52	Male	Flu	12000
t_2	51	Male	Fever	14000
t_3	52	Male	HIV	23000
t_4	54	Male	Cancer	15000
t_5	56	Female	Fever	22000
t_6	55	Female	Fever	25000
t_7	55	Female	HIV	13000

An example of privacy preservation in datasets based on Anatomy. Let Table 1 be the specified dataset, and the value of l is 2. Thus, a partitioned data version of Table 1 is shown in Table 5, and a data version of the quasi-identifier and sensitive tables is shown in Tables 6 and 7, respectively. These anatomized tables can be utilized by joining their defined partition identifiers (PID), and they can guarantee that every possible query condition always obtains at least l distinctly satisfied sensitive values. Moreover, these tables do not have any counterfeit query conditions. Therefore, we can conclude that the privacy preservation is based on Anatomy constraints; it can maintain the data utility of datasets to be better than the privacy preservation that is based on k-Anonymity, l-Diversity, and t-Closeness constraints.

To the best of our knowledge about k-Anonymity, l-Diversity, t-Closeness, and Anatomy, it is insufficient to address privacy violation issues in datasets with multiple sensitive attributes [33], e.g., the dataset is shown in Table 8.

Table 9: *A 2-Diversity data version of Table 8 by maintaining the data utility of the Disease attribute as much as possible.*

♯	Age	Sex	Disease	Income
t_1	51-52	Male	Flu	12000
t_2	51-52	Male	Fever	14000
t_3	52-54	Male	HIV	23000
t_4	52-54	Male	Cancer	15000
t_5	55-56	Female	Fever	22000
t_6	55-56	Female	Fever	25000
t_7	55-56	Female	HIV	13000

Table 10: *A 2-Diversity data version of Table 8 by maintaining the data utility of the Income attribute as much as possible.*

♯	Age	Sex	Disease	Income
t_1	50-55	*	Flu	12000
t_7	50-55	*	HIV	13000
t_2	51-54	Male	Fever	14000
t_4	51-54	Male	Cancer	15000
t_3	52-56	*	HIV	23000
t_5	52-56	*	Fever	22000
t_6	52-56	*	Fever	25000

Table 11: *A 2-Diversity data version of Table 8 by maintaining the data utility of the Sex attribute as much as possible.*

♯	Age	Sex	Disease	Income
t_1	50-54	Male	Flu	12000
t_2	50-54	Male	Fever	14000
t_3	50-54	Male	HIV	23000
t_4	50-54	Male	Cancer	15000
t_5	55-56	Female	Fever	22000
t_6	55-56	Female	Fever	25000
t_7	55-56	Female	HIV	13000

Table 12: *A 2-Diversity data version of Table 8 by maintaining the data utility of the Age attribute as much as possible.*

♯	Age	Sex	Disease	Income
t_1	50-52	Male	Flu	12000
t_2	50-52	Male	Fever	14000
t_3	50-52	Male	HIV	23000
t_4	54-56	*	Cancer	15000
t_5	54-56	*	Fever	22000
t_6	55	Female	Fever	25000
t_7	55	Female	HIV	13000

To address privacy violation issues in datasets with multiple sensitive attributes, the privacy preservation model is proposed in [19], [29], [30], [31], and [33]. It is extended from k-Anonymity, l-Diversity, t-Closeness, or Anatomy. For this reason, it also has various data utility issues that must be improved. Moreover, we found that these privacy preservation models are highly complex with data transformation.

An example of data utility issues in datasets that have multiple sensitive attributes with l-Diversity. We give Table 8 to be the specified dataset. The value of l is 2. Thus, Tables 9, 10, 11, and 12 can be the 2-Diversity data version of Table 8. However, we can see that these tables have different levels of data meaning in terms of data utilization. That is, Table 9 has the data utility to be more than the others when the data analyst utilizes Disease. However, when the data analyst must utilize Income, he/she can see that Table 10 has the data utility to be more than others. While Table 11 has the data meaning of Sex to be more than the others. If the data analyst must utilize Age, he/she can observe that Table 12 has higher data utility than the others.

From these examples, we can conclude that a specified dataset and a specified privacy preservation constraint have various data versions that can be satisfied. Thus, only the high data utility version is desired. To achieve this aim, a new privacy preservation model is proposed in this work. It will be presented in Section 3.

2. CONTRIBUTIONS AND PAPER OUTLINES

In the previous section (Section 1), we discussed privacy violation issues that occur when datasets are

released. Moreover, we demonstrate the vulnerabilities of the existing privacy preservation models. Thus, a stronger privacy preservation model (i.e., the proposed model) is very needed; it will be proposed in Section 3. Then, the experimental evaluations can show the effectiveness and efficiency of the proposed privacy preservation model (Section 4). Finally, the conclusion of this work will be discussed in Section 5.

3. THE PROPOSED MODEL

In this section, a new privacy preservation model for datasets with multiple sensitive attributes is proposed. It aims to address privacy violation issues and maintain the data utility in datasets.

3.1 Dataset

Let $U = \{u_1, u_2, \ldots, u_n\}$ be the set of users. Let $QI = \{qi_1, qi_2, \ldots, qi_a\}$ be the set of quasi-identifier attributes. Let $S = \{s_1, s_2, \ldots, s_b\}$ be the set of sensitive attributes. Let DD_{qi_y} and DD_{s_x}, where $1 < y < a$ and $1 < x < b$, be the data domain of $qi_y \in QI$ and the data domain of $s_x \in S$ respectively. Let $D = \{d_1, d_2, \ldots, d_n\}$ be the raw dataset that collects n user profile tuples such that D is multiset. Every $d_i \in D$, where $1 < i < n$, represents the profile tuple of $u_i \in U$. That is, d_i is constructed by $QI^i \cup S^i$, where $QI^i \subseteq QI$ and $S^i \subseteq S$, such that QI^i and S^i are the set of u_i's quasi-identifier and sensitive attributes, respectively. Let $D[QI]$ and $D[S]$ be the data projection of QI and S of D respectively. Let $D[d_i]$ be the data projection of d_i of D. Also, let $d_i[QI]$ and $d_i[S]$ be the data projection of QI and S of $d_i \in D$, respectively.

3.2 Continuous sensitive value

Let DD_{s_x} be the data domain of $s_x \in S$ such that it is a set of continuous values, e.g., it is based on numeric and datetime. Let $SA_{s_x} \subseteq s_x$ be the set of specified sensitive values from s_x. Let $MAX(SA_{s_x})$ represent the maximum value (the upper bound value) of SA_{s_x}. Let $MIN(SA_{s_x})$ represent the minimum value (the lower bound value) of SA_{s_x}. Thus, the range of SA_{s_x} can be represented by $[MAX(SA_{s_x}), MIN(SA_{s_x})]$. Let $ERR(SA_{s_x})$ denote an error measure defined on SA_{s_x}, i.e., $E(SA_{s_x}) = MAX(SA_{s_x}) - MIN(SA_{s_x})$. More value of $E(SA_{s_x})$ means a larger range or more different values of SA_{s_x}. This property of SA_{s_x} can be used to address privacy violation issues in SA_{s_x}. That is, let $R_{s_x} \in I^+$ be the privacy preservation constraint. SA_{s_x} does not have any privacy violation issues when $E(SA_{s_x})$ is equal to or greater than $R_{s_x})$, i.e., $E(SA_{s_x}) \geq R_{s_x}$. In addition, the higher value of R_{s_x} leads to a high level of privacy preservation in SA_{s_x}.

The privacy and utility goals of continuous sensitive values are that when they are released to utilize outside the scope of data-collecting organizations, they can guarantee the range of data errors and data re-identifications to be at least $MAX(SA_{s_x}) - MIN(SA_{s_x})$.

3.3 Nominal sensitive value

Let DD_{s_x} be the data domain of $s_x \in S$ such that it is a set of nominal values, e.g., it is based on text and string. Let $SP_{DD_{s_x}} \subseteq DD_{s_x}$ be the set of protected sensitive values for s_x. Let $SA_{s_x} \subseteq s_x$ be the set of specified sensitive values from s_x. More cardinality of SA_{s_x} (i.e., $|SA_{s_x}|$) is more possible sensitive values in SA_{s_x}. Thus, the ability or error of data re-identification for SA_{s_x} can be denoted by $E(SA_{s_x})$, i.e., $E(SA_{s_x}) = |SA_{s_x}|$. With this property of SA_{s_x}, it can be used to address privacy violation issues in SA_{s_x}. That is, let $C_{s_x} \in I^+$ be the privacy preservation constraint. Let $SP_{SA_{s_x}} \subset SA_{s_x}$ be the protected sensitive values that are available in SA_{s_x}. Therefore, SA_{s_x} does not have any privacy violation issues when $\frac{|SP_{SA_{s_x}}|}{|SA_{s_x}|} \leq C_{s_x}$. In addition, the lower value of C_{s_x} leads to a high level of privacy preservation in SA_{s_x}.

The privacy goal of nominal sensitive values is that when they are released to utilize outside the scope of data-collecting organizations, they can guarantee the confidence of data re-identification to be at most C_{s_x}.

3.4 The optimal partition of continuous sensitive values

Let $f_{PCS}(D[QI] \cup D[s_x], R_{s_x} : D[QI] \cup D[s_x] \longrightarrow_{R_{s_x}} PAR_{s_x}$ be the function for partitioning $D[QI] \cup D[s_x]$, where $s_x \in S$, to be PAR_{s_x} That is, s_x is a continuous sensitive attribute of D, and PAR_{s_x} is in the form of $PAR_{s_x} = \{par^1_{s_x}, par^2_{s_x}, \ldots, par^z_{s_x}\}$ such that every partition $par^j_{s_x} \in PAR_{s_x}$, where $par^j_{s_x}$, where $1 < j < z$, must be satisfied by the limitations as follows.

- Each partition $par^j_{s_x}$, where $1 < j < z$, is constructed by the defined partition identifier j and the tuples $T \subseteq D[QI] \cup D[s_x]$.
- $E(par^j_{s_x}[s_x])$ is equal to or greater than R_{s_x}, i.e., $MAX(par^j_{s_x}[s_x]) - MIN(par^j_{s_x}[s_x]) \leq R_{s_x})$,
- $MAX(par^j_{s_x}[s_x]) \leq MIN(par^j_{s_x}[s_x])$
- $\bigcup_{j=1}^{z} par^j_{s_x} = D[QI] \cup D[s_x]$,
- $\bigcap_{j=1}^{z} par^j_{s_x} = \varnothing$, and
- $\sum_{j=1}^{z} E(par^j_{s_x}[s_x])$ is minimized.

3.5 The optimal partition of nominal sensitive values

Let $f_{PNS}(D[QI] \cup D[s_x], C_{s_x}) : D[QI] \cup D[s_x] \rightarrow_{C_{s_x}} PAR_{s_x}$ be the function for partitioning $D[QI] \cup D[s_x]$, where $s_x \in S$, to be PAR_{s_x} such that s_x is a nominal sensitive attribute of D, and PAR_{s_x} is also in the form of $PAR_{s_x} = \{par^1_{s_x}, par^2_{s_x}, \ldots, par^z_{s_x}\}$ such that every partition $par^j_{s_x} \in PAR_{s_x}$, where $par^j_{s_x}$, where $1 < j < z$, must be satisfied by the limitations as follows.

- Each partition $par^j_{s_x}$, where $1 < j < z$, is constructed by the defined partition identifier j and the tuples $T \subseteq D[QI] \cup D[s_x]$.
- $E(par^j_{s_x}[s_x])$ is lesser or equal to than C_{s_x}, i.e., $\frac{|SP_{par^j_{s_x}[s_x]}|}{|par^j_{s_x}[s_x]|} \leq C_{s_x}$,
- $\bigcup_{j=1}^{z} par^j_{s_x} = D[QI] \cup D[s_x]$,
- $\bigcap_{j=1}^{z} par^j_{s_x} = \varnothing$, and
- $\sum_{j=1}^{z} E(par^j_{s_x}[s_x])$ is minimized.

3.6 Data anonymization

Let PAR_{s_x} be the partitioned data version of each $s_x \in S$ of D, where $1 < x < b$. That is, if the data domain of s_x is continuous values, every partition $par^j_{s_x} \in PAR_{s_x}$ is satisfied by R_{s_x}. But if the data domain of s_x is nominal values, every partition $par^j_{s_x} \in PAR_{(}s_x)$ is satisfied by C_{s_x}. Let $f_{Ano}(PAR_{s_x}) : PAR_{s_x} \rightarrow PAR'_{s_x} \cup ATT\_NAME_{s_x}$ be the function for anonymizing PAR_{s_x} to be indistinguishable, where $ATT\_NAME_{s_x}$ is the attribute name. That is, $par^j_{s_x}[qi_1], par_{(}s_x)^j[qi_2], \ldots, par^j_{s_x}[qi_a]$, and $par^j_{s_x}[s_x]$ are anonymized by $qi'_1, qi'_2, \ldots, qi'_a$, and s'_x respectively such that $qi'_1, qi'_2, \ldots, qi'_a$, and s'_x are multisets that represent the values that are available in $par^j_{s_x}[qi_1], par^j_{s_x}[qi_2], \ldots, par^j_{s_x}[qi_a]$, and $par^j_{s_x}[s_x]$ respectively. Let f_{Released} $(f_{Ano}(PAR_{s_1}), f_{Ano}(PAR_{s_2}), \ldots, f_{Ano}(PAR_{s_b}):f_{Ano}(PAR_{s_1}), f_{Ano}$ $(PAR_{s_2}), \ldots, f_{Ano}(PAR_{s_b}) \rightarrow D'$ be the function for constructing the released data version D' of D, where $D' = f_{Ano}(PAR_{s_1}) \cup f_{Ano}(PAR_{s_2}) \cup \ldots \cup f_{Ano}(PAR_{s_b})$.

For example, let Table 8 be the specified dataset. Let the value of R_{Income} and $\text{C}_{\text{Disease}}$ be 3000 and 0.5, respectively. Given that HIV and Cancer are protected sensitive values. With these given instances, both sub-data versions of Table 8 must be considered to satisfy the given privacy preservation constraints, i.e., Tables 13 and 14. With Table 13, the data domain of the sensitive attribute, Income, is continuous values. For this reason, every partition must be satisfied by the data limitations that are defined in Section 3.4 with R_{Income}=3000. Thus, a partitioned data version of Table 13 is shown in Table 15. With Table 14, the data domain of the sensitive attribute, Disease, is nominal values. For this reason, every partition must be satisfied by the data limitations that are defined in Section 3.5 with $\text{C}_{\text{Disease}}$=0.5. Thus, a partitioned data version of Table 14 is shown in Table 16. For privacy preservation, the values of each attribute are available in each partition of Tables 15 and 16 to be indistinguishable from their set, which is based on a multiset. Therefore, a released data version of Table 8 is satisfied by the proposed privacy preservation constraints that are proposed in Section 3.6 with R_{Income}=3000 and $\text{C}_{\text{Disease}}$=0.5 to be shown in Table 17. With Table 17, we can see that the ability of data re-identification for every income of each partition is in the range between its lower and upper bounds. The confidence of data re-identification for every protected disease in every partition of this table is at most 0.5 (50%). Moreover, we can see that Table 17 does not have any counterfeit query conditions, and it does not have the issue about multiple released tables that must be considered when the released tables are utilized.

Table 13: *The data version of Table 8 with the Income attribute.*

Age	Sex	Income
52	Male	12000
51	Male	14000
52	Male	23000
54	Male	15000
56	Female	22000
55	Female	25000
55	Female	13000

Table 14: *The data version of Table 8 with the Disease attribute.*

Age	Sex	Disease
52	Male	Flu
51	Male	Fever
52	Male	HIV
54	Male	Cancer
56	Female	Fever
55	Female	Fever
55	Female	HIV

Table 15: *A data partition version of Table 13 is satisfied by $R_{Income} = 3000$.*

Age	Sex	Income
52	Male	12000
55	Female	13000
51	Male	14000
54	Male	15000
56	Female	22000
52	Male	23000
55	Female	25000

Table 16: *A data partition version of Table 13 is satisfied by $C_{Disease} = 0.5$.*

Age	Sex	Disease
52	Male	Flu
52	Male	HIV
51	Male	Fever
55	Female	HIV
54	Male	Cancer
56	Female	Fever
55	Female	Fever

Algorithm 1 The proposed anonymization algorithm

Require: D, $SP_{DD_{s_{q_1}}}, \ldots, SP_{DD_{s_{q_b}}}, C_{s_{q_1}}. \ldots, C_{s_{q_b}}, R_{s_{p_1}}. \ldots, R_{s_{p_b}}$;
Ensure: D';
  while $x = 1$ *to* b **do**
    $Temp \leftarrow D[QI] \cup D[s_{q_x}]$;
    while $|Temp| > 0$ **do**
      $par_{s_{q_x}} \leftarrow PARTITION(Temp)$, where $E(par_{s_{q_x}})$ is minimized;
      $SP_{par_{s_{q_x}}} \leftarrow par_{s_{q_x}}[s_{q_x}] \cap SP_{DD_{s_{q_x}}}$;
      if $(|SP_{par_{s_{q_x}}}|/|par_{s_{q_x}}[s_{q_x}]|) \leq C_{s_{q_x}}$ **then**
        $Temp \leftarrow Temp - par_{s_{q_x}}$;
        $D' \leftarrow D' \cup par_{s_{q_x}}$;
      end if
    end while
  end while
  while $x = 1$ *to* b **do**
    $Temp \leftarrow D[QI] \cup D[s_{p_x}]$;
    $Temp \leftarrow ASCENDING(Temp)$;
    while $|Temp| > 0$ **do**
      $par_{s_{p_x}} \leftarrow PARTITION(Temp)$, where $E(par_{s_{p_x}})$ is minimized;
      if $(MAX(par_{s_{p_x}}[s_x]) - MIN(par_{s_{p_x}}[s_x])) \geq R_{s_{p_x}}$ **then**
        $Temp \leftarrow Temp - par_{s_{q_x}}$;
        $D' \leftarrow D' \cup par_{s_{q_x}}$;
      end if
    end while
  end while
  return D';

Table 17: *A data version of Table 8 is satisfied by the proposed privacy preservation constraints, where R_{Income}=3000 and $C_{Disease}$= 0.5.*

Age	Sex	Sensitive values	Sensitive attributes
{52, 55, 51, 54}	{Male, Female}	{12000, 13000, 14000, 15000}	Income
{52, 55, 56}	{Male, Female}	{22000, 23000, 25000}	Income
{52}	Male	{Flu, HIV}	Disease
{51, 55}	{Male, Female}	{HIV, Cancer}	Disease
{55, 56}	{Female}	{Fever}	Disease

3.7 The proposed algorithm

In this section, we propose an algorithm that can be used for transforming D to become D' to satisfy the proposed privacy preservation constraints. With the proposed algorithm, its inputs are the specified dataset D, the set of protected sensitive values $SP_{DD_{s_{q_1}}}, \ldots, SP_{DD_{s_{q_b}}}$. The nominal privacy preservation constraints $C_{s_{q_1}}, \ldots, C_{s_{q_b}}$, and the continuous privacy preservation constraints $R_{s_{p_1}}, \ldots, R_{s_{p_b}}$. The output of the proposed algorithm is D' that is satisfied by $C_{s_{q_1}}, \ldots, C_{s_{q_b}}$, $R_{s_{p_1}}, \ldots, R_{s_{p_b}}$.

For privacy preservation, each nominal sensitive attribute is first transformed to satisfy $C_{s_{q_1}}, \ldots, C_{s_{q_b}}$ by the steps as follows.

At first, the sub-data version of D for each nominal sensitive attribute s_{q_x}, where $1 \leq x \leq b$, is constructed. It is collected by TEMP. Then, the tuples of TEMP are iterated and partitioned by $C_{s_{q_1}}$ such that each partition must have $E(par_{s_{q_1}})$ to be minimized. Moreover, the tuples of $par_{s_{q_1}}$ are removed from TEMP. Finally, $par_{s_{q_1}}$ is collected by D'.

Another processor of the proposed privacy preservation algorithm is the processor for transforming each continuous sensitive attribute s_{p_1}, where $1 \leq x \leq b$, to satisfy $R_{s_{p_1}}$ by the steps as follows.

In the first step, the sub-data version of D for each continuous sensitive attribute s_{p_x}, where $1 \leq x \leq b$, is constructed. It is also collected by TEMP. Then, the tuples of TEMP are re-sorted by the values of s_{p_x} in ascending order. Then, the tuples of TEMP are iterated and partitioned by $R_{s_{p_1}}$ such that each partition must have $E(par_{s_{q_1}})$ that is minimized. Moreover, the tuples of $par_{s_{q_1}}$ are removed from TEMP. Finally, $par_{s_{p_1}}$ is collected by D'. Finally, D' is satisfied by $C_{s_{q_1}}, \ldots, C_{s_{q_b}}$, $R_{s_{p_1}}, \ldots, R_{s_{p_b}}$ to be returned by the algorithm.

In addition, with the complexity of the proposed privacy preservation algorithm, it can be separated into three parts, i.e., the cost of partitioning the numeric sensitive values, partitioning the nominal sensitive values, and constructing the released data version (merging the constructed partitions of every sensitive attribute). Let $NuS \in S$ and $NoS \in S$, where

$NuS \cup NoS = S$ and $NuS \cap NoS = \phi$, be the numeric and nominal sensitive attributes of D, respectively. With the numeric sensitive attribute, the tuples of D are first sorted in ascending order with s_x. The cost of data ordering is $|D| \log |D|$ because the sorting processor of the proposed algorithm is based on merge sort [34]. After that, the sorted tuples are partitioned by R_{s_x}, i.e., the different ranges between the lower bound and the upper bound of each partition must be less than or equal to R_{s_x}. For this reason, all tuples in the temporary are considered in each iteration, and only one tuple is chosen to insert into an appropriate partition and removed from the temporary. Thus, the number of tuples in the iteration under consideration is always one row greater than the next iteration. Therefore, the cost of partitioning all numeric sensitive attributes is shown in Equation 1.

$$O(NuS) = |NuS| \cdot |D|^2 \cdot (|D| - 1) \cdot \log |D| \quad (1)$$

With nominal sensitive attributes, their cost is based on three factors, i.e., the number of protected sensitive attributes, the number of tuples in D, and the size of NoS. That is, the desired partition of every protected sensitive value in each attribute is constructed from its appropriate non-protected sensitive values. Also, only one tuple is inserted into a proper partition with each iteration after that it is removed from the temporary, so, the number of tuples in the iteration under consideration is also one row greater than the next iteration. Therefore, the cost of partitioning all nominal sensitive attributes is shown in Equation 2.

$$\mathrm{O(NoS)} = |NoS| \cdot (|D| - |SP_{DD}|) \cdot (|D| - |SP_{DD}| - 1) \quad (2)$$

Another cost of the proposed algorithm, the cost of constructing the released data version, is merging the constructed partitions of every sensitive attribute. It can be defined by Equation 3.

$$O(PAR(D)) = \sum_{x=1}^{|S|} |PAR_{s_x}|, \text{ where } s_x \in S \quad (3)$$

Therefore, the cost of constructing $\boldsymbol{D'}$ can be defined from $\boldsymbol{O(NuS)}$, $\boldsymbol{O(NoS)}$, and $\boldsymbol{O(PAR(D))}$, Equation 4.

$$O(D') = O(NuS) + O(NoS) + O(PAR(D)) \quad (4)$$

4. EXPERIMENT

In this section, the experiments for evaluating the efficiency and effectiveness of the proposed model are discussed by comparing with "l-Diversity [4]", "Anatomy [10]", "k-Anonymity and l-Diversity [32]", "$(l^{p1}, \ldots, l^{pz})$-Privacy [19]".

4.1 Experimental setup

The experiments are proposed to evaluate the efficiency and effectiveness of the proposed model. They are built on four Intel(R) Xeon(R) Gold 5318H CPUs @2.50GHz 2.49 GHz with 512 GB memory and 18.1TB M.2 NVMe HDD running on the Window Server 2022 Standard 64 bit. The proposed experiments are built and executed on Visual Studio 2022 with C♯ and MS SQL Server 2022. Moreover, they are further based on Adult dataset which is available at the UCI Machine Learning Repository. This dataset is constructed from 32561 tuples such that every tuple consists of 14 attributes that are Age, Workclass, Fnlwgt, Education, Education-num, Marital-status, Occupation, Relationship, Race, Sex, Capital-gain, Capital-loss, Hours-per-week, and Native-country. To conduct the experiments effectively, all user profile tuples include the missing values, "?"; they are removed. Thus, the experimental dataset only includes 1428 user profile tuples. Moreover, the data is available in the attributes that are Native-country, Workclass, Fnlwgt, Education-num, Marital-status, Capital-gain, Hours-per-week, it is also removed. Given Age, Sex, Race, Education, are the quasi-identifiers, and Occupation, Marital-status, and Capitalloss are the sensitive attributes. With the Occupation, we give Exec-managerial, Prof-specialty, Machine-op-inspect, and Armed-Forces to be the protected sensitive values. While Divorced, Separated, Married-spouse-absent are the protected sensitive values in Marital-status. The experimental dataset information is shown in Table 19.

4.2 Experimental result and discussion

4.2.1 Effectiveness

In this section, the experiments can evaluate the effectiveness of the proposed model to be discussed. With the proposed model and its compared models, the—y are based on data partitions or equivalence classes. Thus, the penalty cost of datasets can be defined by the discernibility metric (DM) that is shown in Equation 5. In addition, the privacy preservation model (e.g., the proposed model) is based on multiple data versions of the specified dataset. The penalty cost can be defined by the global discernibility metric (GDM) as shown in Equation 6. More scores of Equations 1 and 2 lead to more penalty costs in the dataset.

$$DM(PAR_{s_x}) = \sum_{j=1}^{|PAR_{s_x}|} |par_{s_x}^j|^2 \quad (5)$$

Table 18: *The information of the experimental dataset.*

Attribute name	Description
Quasi-identifier attributes	
Age	Continuous
Sex	Female, Male
Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US (Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
Sensitive attribute	
Occupation	Tech-support, Craft-repair, Other-service, Sales, **Exec-managerial**, **Prof-specialty**, Handlers-cleaners, **Machine-op-inspect**, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, **Armed-Forces**
Marital-status	Married-civ-spouse, **Divorced**, Never-married, **Separated**, Widowed, **Married-spouse-absent**, Married-AF-spouse
Capital-loss	Continuous

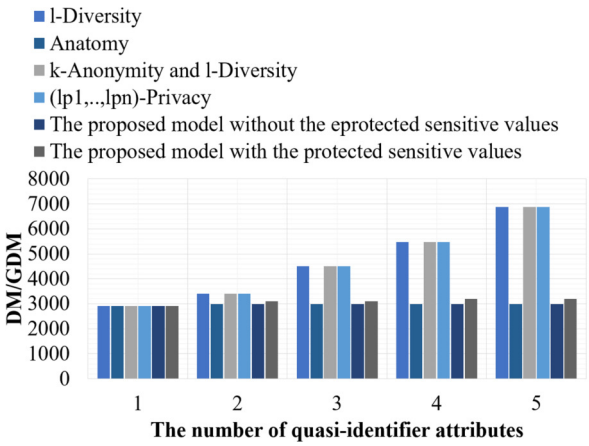

Fig.1: *The effectiveness is based on the number of quasi-identifier attributes.*

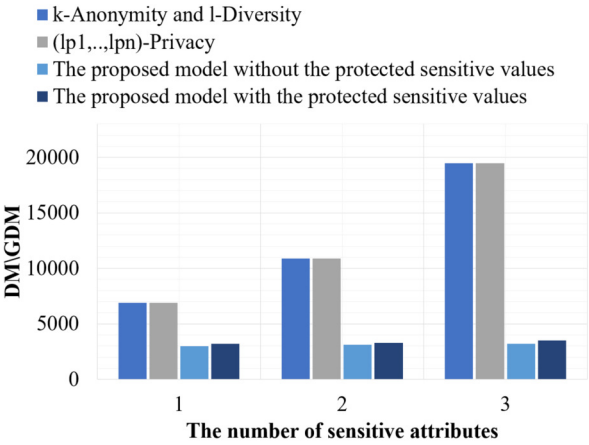

Fig.2: *The effectiveness is based on the number of sensitive attributes.*

$$\mathrm{GDM}(\mathrm{PAR}_{s_1}, PAR_{s_2}, \ldots, PAR_{s_b}) = \sum_{x=1}^{b} DM(PAR_{s_x}) \quad (6)$$

In the first experiment, we propose to evaluate the number of quasi-identifier attributes that can influence the size of partitions and equivalence classes. For the experiment, all quasi-identifier attributes are available in the experimental dataset by varying from 1 to 5 attributes. However, only Occupation is the sensitive attribute. The value of l for l-Diversity and Anatomy is 2. With k-Anonymity and l-Diversity, the value of k and l is 2. The value of $l^{p1}, \ldots, l^{pz}$ for $(l^{p1}, \ldots, l^{pz})$-Privacy is 2. With yjr proposed model, we give the value of $C_{Occupation}$ for the proposed model to be 0.5.

With the experimental results that are shown in Figure 1, we can see that the number of quasi-identifier attributes directly influences the penalty cost of DM and GDM, i.e., the number of quasi-identifier attributes directly affects the size of partitions and equivalence classes. In addition, the experiments further show that the proposed model and Anatomy have less effect on the penalty cost of DM and GDM than other models. That is, the datasets can satisfy the proposed privacy preservation constraints and Anatomy constraints by using their multiset of original values. Moreover, the experimental results show that the proposed model is based on considering the set of protected sensitive values; they have the penalty cost of DM and GDM to be more than the datasets without considering the set of protected sensitive values. That is because the set of protected sensitive values directly affects the size of partitions and equivalence classes, but we found that they have little effect.

In the second experiment, we evaluate the number of sensitive attributes that can influence the size of partitions and equivalence classes. In addition, l-Diversity and Anatomy are not considered in this experiment because they do not propose to address privacy violation issues in datasets that have multiple sensitive attributes. For the experiment, all quasi-identifier attributes are available in the experimental dataset. The number of sensitive attributes varies from 1 to 3. The value of k and l for k-Anonymity

and l-Diversity is 2. The value of $l^{p1}, \ldots, l^{pz}$ for $(l^{p1}, \ldots, l^{pz})$-Privacy is also 2. With the proposed model, we give the confidence of data re-identification for all nominal sensitive attributes to be 0.5, and the value of $R_{\text{Capital loss}}$ is 100.

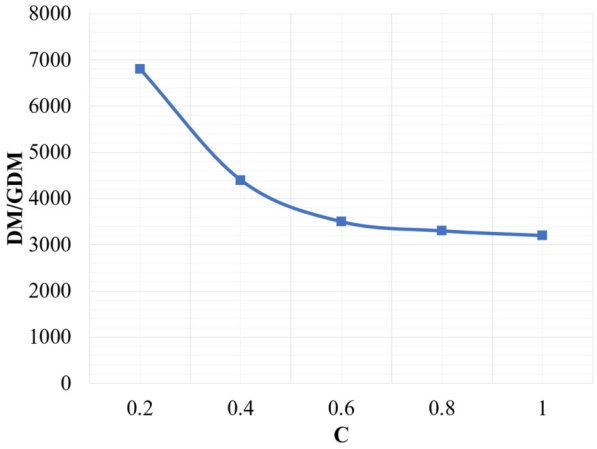

Fig.3: *The effectiveness is based on the confidence of data re-identifications.*

With the experimental results that are shown in Figure 2, we can conclude that the number of sensitive attributes also more influences the penalty cost of DM and GDM, i.e., the number of sensitive attributes directly influences the size of partitions and equivalence classes. Moreover, we can see that the penalty cost of DM and GDM of k-Anonymity and l-Diversity constraints and Anatomy constraints are the same because they are both extended versions of l-Diversity. Moreover, the experimental results show that the proposed model is less affected than the compared models. That is because each sensitive attribute of datasets is independently considered to satisfy privacy preservation constraints with the proposed model. Also, the experimental results show that the proposed model is based on considering the set of protected sensitive values; they have the penalty cost of DM and GDM to be more than the datasets without considering the set of protected sensitive values. That is because the set of protected sensitive values directly affects the size of partitions and equivalence classes, but we found that they have little effect.

In the third experiment, we evaluate the effect of the confidence of data re-identifications that can influence the penalty cost of DM and GDM. For the experiment, all quasi-identifier attributes are available in the experimental dataset. However, only Occupation is the sensitive attribute. The value of $C_{\text{Occupation}}$ varies from 0.2 to 1.0. With the experimental results that are shown in Figure 3, we can see that a higher value of $C_{\text{Occupation}}$ leads to a smaller size of partitions. The cause of the smaller size of partitions is that when increasing the value of $C_{\text{Occupation}}$, it is the number of protected sensitive values that affect the number of sensitive values that are available in each partition of the datasets.

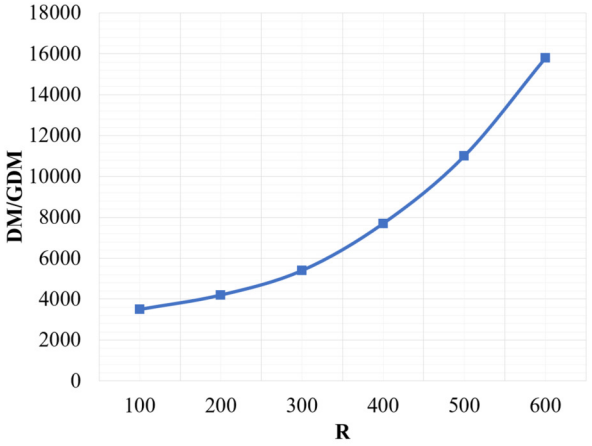

Fig.4: *The effectiveness is based on the range of data bounding.*

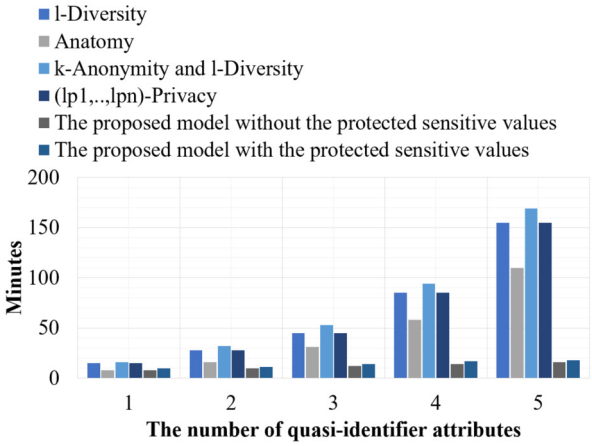

Fig.5: *The efficiency is based on the number of quasi-identifier attributes.*

In the fourth experiment, we evaluate the effect of the range of data bounding that can influence the penalty cost of DM and GDM. For the experiment, all quasi-identifier attributes are available in the experimental dataset. However, only Capital-loss is the sensitive attribute. The value of $R_{\text{Capital loss}}$ varies from 100 to 600. With the experimental results that are shown in Figure 4, we can see that more value of $R_{\text{Capital loss}}$ leads to a larger size of partitions. The cause of the larger size of partitions is when increasing the value of $R_{\text{Capital loss}}$, which is the number of sensitive values that are available in each partition of the datasets.

4.2.2 Efficiency

In the fifth experiment, we propose to show the efficiency, the execution time, of the proposed model and its compared models that are used for transforming datasets to satisfy privacy preservation constraints such that it is based on the number of quasi-identifier

attributes. In addition, this experiment is set up to be the same as the first experiment. With the experimental results that are shown in Figure 5, we can see that the number of quasi-identifier attributes has more influence on the execution time for transforming datasets to satisfy privacy preservation constraints, i.e., more quasi-identifier attributes lead to using more execution time for transforming datasets to satisfy privacy preservation constraints. Moreover, we can see that the compared models use the execution time for transforming datasets to be more than the proposed model. The cause of using more execution time of the compared models is that the search space for considering the tuples into an appropriate partition and equivalence class of datasets increases when the number of quasi-identifier attributes is increased. Moreover, with l-Diversity, k-Anonymity and l-Diversity, and $(l^{p1},\ldots,l^{pz})$-Privacy, they further have the execution time about ascertaining an appropriate less specific value for generalizing the specified group of unique quasi-identifier values to be indistinguishable. The reason why the number of quasi-identifier attributes has little effect on the execution time of the proposed model is that the proposed model partitions the tuples of datasets by only considering the sensitive values.

In the sixth experiment, we propose to show the execution time of the proposed model and its compared models that are used to transform datasets for satisfying privacy preservation constraints such that it is based on the number of sensitive attributes. In addition, this experiment is set up to be the same as the second experiment. With the experimental results that are shown in Figure 6, we see that the number of sensitive attributes also has more influence on the execution time for transforming datasets to satisfy privacy preservation constraints, i.e., more sensitive attributes lead to using more execution time for transforming datasets to satisfy privacy preservation constraints. The cause of more using the execution time is when the number of sensitive attributes that are increased, all experimental models are privacy preservation models that are based on considering the number of distinct sensitive values. Moreover, we further see that the number of sensitive values is more influence on the execution time of the compared models than the proposed model. That is because aside from the number of distinct sensitive values, the compared models have the execution time for ascertaining an appropriate less specific value for generalizing the specified group of unique quasi-identifier values to be indistinguishable. Moreover, with k-Anonymity and l-Diversity, it further has a parameter (i.e., k) that must be considered in its privacy preservation constraints. For this reason, k-Anonymity and l-Diversity uses the execution time to be more than other experimental models. Also, the reason why the proposed model is more efficient than other experimental models, it is that the proposed model does not have any execution time for ascertaining an appropriate less specific value for generalizing the specified group of unique quasi-identifier values to be indistinguishable.

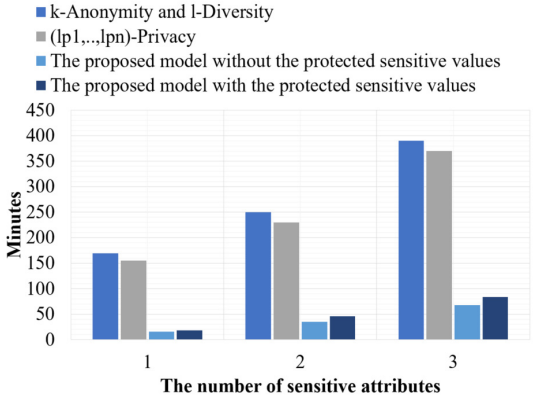

Fig.6: *The efficiency is based on the number of sensitive attributes.*

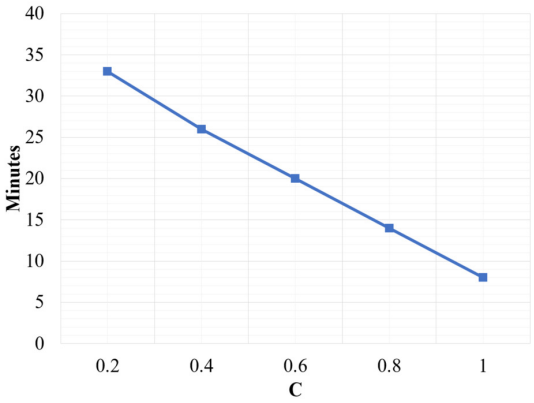

Fig.7: *The efficiency is based on the confidence of data re-identifications.*

In the seventh experiment, we propose to show the execution time of the proposed model that is used for transforming datasets to satisfy privacy preservation constraints such that it is based on the confidence of data re-identification. In addition, this experiment is set up to be the same as the third experiment. With the experimental results that are shown in Figure 7, we see that more confidence of data re-identification leads to less execution time for transforming datasets to satisfy the privacy preservation constraints of the proposed model. The cause of less execution time with more confidence of data re-identification is the size of partitions and the number of considered protection-sensitive values. More confidence of data re-identification leads to a smaller number of protected sensitive values that must be considered in each partition of datasets, or a smaller search space for considering the proper tuples into the appropriate partition of datasets.

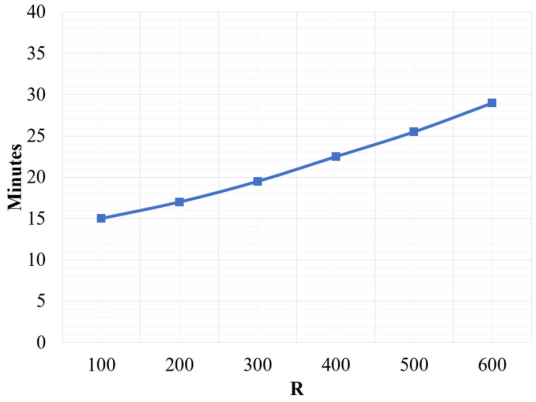

Fig.8: *The efficiency is based on the range of data bounding.*

In the final experiment, we propose to show the execution time of the proposed model that is used for transforming datasets to satisfy privacy preservation constraints, such that it is based on the data bounding. In addition, this experiment is set up to be the same as the fourth experiment. With the experimental results that are shown in Figure 8, we see that when the value of data bounding is increased, the execution time for transforming the datasets to the satisfaction of privacy preservation constraints also increases. That is because the value of data bounding also influences the search space for considering the proper tuples into the appropriate partition of datasets, i.e., a higher value of data bounding often leads to a larger search space for considering the proper tuples into the appropriate partition of datasets.

5. CONCLUSION

In this work, we illustrate the limitations of the existing privacy preservation models, i.e., they still have privacy violation issues and data utility issues that must be improved. To address these vulnerabilities of these models, a new privacy preservation model is proposed in this work, such that it is based on the confidence of data re-identification and the data bounding. That is, before datasets are released, their tuples are partitioned by the confidence of data re-identification and the data bounding. The confidence of data re-identification is the privacy preservation constraint for the non-sensitive attributes. The data bounding is the confidence range of data re-identification for the continuous sensitive attributes. From the experimental results, they indicate that the proposed model is an effective and efficient privacy preservation model, i.e., it is more effective and efficient than the compared models. Moreover, datasets are satisfied by the privacy preservation constraint of the proposed model, which can guarantee the confidence and data bounding of data re-identification.

AUTHOR CONTRIBUTIONS

Conceptualization, Surapon Riyana; methodology, Surapon Riyana and Noppamas Riyana; software, Surapon Riyana; validation, Surapon Riyana, Kittikorn Sasujit, Nigran Homdoung, and Noppamas Riyana; formal analysis, Surapon Riyana; investigation, Surapon Riyana; data curation, Surapon Riyana; writing—original draft preparation, Surapon Riyana; writing—review and editing, Kittikorn Sasujit, Nigran Homdoung, and Noppamas Riyana; visualization, Surapon Riyana, Kittikorn Sasujit, Nigran Homdoung, and Noppamas Riyana; supervision, All authors have read and agreed to the published version of the manuscript.

References

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[2] Y. Liang and R. Samavi, "Optimization-based k-anonymity algorithms," *Computers & Security,* vol. 93, p. 101753, 2020.

[3] Y. Canbay, "On the Complexity of Optimal k-Anonymity: A New Proof Based on Graph Coloring," in *IEEE Access*, vol. 12, pp. 94197-94204, 2024.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[5] F. Ashkouti, K. Khamforoosh, A. Sheikhahmadi and H. Khamfroush, "DI-Mondrian: Distributed improved Mondrian for satisfaction of the L-diversity privacy model using Apache Spark," *Information Sciences*, vol. 546, pp. 1–24, 2021.

[6] M. Jeon, O. Temuujin, J. Ahn and D. H. Im, "Distributed L-diversity using Spark-based algorithm for large resource description frameworks data," *The Journal of Supercomputing*, vol. 77, no. 7, pp. 7270–7286, 2021.

[7] K. Oishi, Y. Sei, J. Andrew, Y. Tahara and A. Ohsuga, "Algorithm to satisfy l-diversity by combining dummy records and grouping," *Security and Privacy*, vol. 7, no. 3, p. e373, 2024.

[8] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, Turkey, pp. 106-115, 2007.

[9] W. Ren, K. Ghazinour and X. Lian, "*kt*-Safety: Graph Release via k-Anonymity and t-Closeness," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 9102–9113, 2022.

[10] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proc. 32nd Int.*

Conf. on Very Large Data Bases (VLDB), pp. 139–150, 2006.

[11] X. He, Y. Xiao, Y. Li, Q. Wang, W. Wang and B. Shi, "Permutation anonymization: Improving anatomy for privacy preservation in data publication," in *New Frontiers in Applied Data Mining: PAKDD 2011 International Workshops, Shenzhen, China, May 24–27, 2011, Revised Selected Papers 15*, Springer, pp. 111–123, 2012.

[12] S. Riyana, N. Riyana and W. Sujinda, "An anatomization model for farmer data collections," *SN Computer Science*, vol. 2, no. 5, p. 353, 2021.

[13] S. Riyana, "Achieving Anatomization Constraints in Dynamic Datasets," *ECTI-CIT Transactions*, vol. 17, no. 1, pp. 27–45, Feb. 2023.

[14] Q. Zhang, N. Koudas, D. Srivastava and T. Yu, "Aggregate Query Answering on Anonymized Tables," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, Turkey, pp. 116-125, 2007.

[15] S. Riyana, N. Riyana and S. Nanthachumphu, "Enhanced (k, e)-anonymous for categorical data," in *Proceedings of the 6th International Conference on Software and Computer Applications*, pp. 62–67, 2017.

[16] R. Chi-Wing Wong, J. Li, A. W.-C. Fu and K. Wang, "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 754–759, 2006.

[17] B. C. M. Fung, M. Cao, B. C. Desai and H. Xu, "Privacy protection for RFID data," in *Proceedings of the 2009 ACM symposium on Applied Computing*, pp. 1528–1535, 2009.

[18] M. Rafiei, M. Wagner and W. M.P. van der Aalst, "TLKC-privacy model for process mining," in *International Conference on Research Challenges in Information Science*, Springer, pp. 398–416, 2020.

[19] S. Riyana, "($l^{p1}, \ldots, l^{pn}$)-Privacy: privacy preservation models for numerical quasi-identifiers and multiple sensitive attributes," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 1, pp. 9713–9729, 2021.

[20] C. Dwork, "Differential privacy," in *Proc. Int. Colloquium on Automata, Languages, and Programming*, Springer, pp. 1–12, 2006.

[21] K. Wei *et al.*, "Federated Learning With Differential Privacy: Algorithms and Performance Analysis," in *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454-3469, 2020.

[22] J. Dong, A. Roth and W. J. Su, "Gaussian differential privacy," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 84, no. 1, pp. 3–37, 2022.

[23] S. Yaseen *et al.*, "Improved Generalization for Secure Data Publishing," in *IEEE Access*, vol. 6, pp. 27156-27165, 2018.

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

[25] F. Farokhi and H. Sandberg, "Ensuring privacy with constrained additive noise by minimizing Fisher information," *Automatica*, vol. 99, pp. 275–288, 2019

[26] Y. Hu, A. Hu, C. Li, P. Li and C. Zhang, "Towards a privacy protection-capable noise fingerprinting for numerically aggregated data," *Computers & Security*, vol. 119, p. 102755, 2022.

[27] R. Wang, Y. Zhu, C. C. Chang and Q. Peng, "Privacy-preserving high-dimensional data publishing for classification," *Computers & Security*, vol. 93, p. 101785, 2020.

[28] S. Riyana, S. Nanthachumphu and N. Riyana, "Achieving privacy preservation constraints in missing-value datasets," *SN Computer Science*, vol. 1, pp. 1–10, 2020.

[29] R. Wang, Y. Zhu, T. S. Chen and C. C. Chang, "Privacy-preserving algorithms for multiple sensitive attributes satisfying t-closeness," *Journal of Computer Science and Technology*, vol. 33, pp. 1231–1242, 2018.

[30] T. Kanwal *et al.*, "Privacy-preserving model and generalization correlation attacks for 1: M data with multiple sensitive attributes," *Information Sciences*, vol. 488, pp. 238–256, 2019.

[31] T. Kanwal *et al.*, "A robust privacy preserving approach for electronic health records using multiple dataset with multiple sensitive attributes," *Computers & Security*, vol. 105, p. 102224, 2021.

[32] T. Gal, Z. Chen and A. Gangopadhyay, "A Privacy Protection Model for Patient Data with Multiple Sensitive Attributes," *International Journal of Information Security and Privacy (IJISP)*, vol. 2, pp. 28–44, Jul. 2008.

[33] A. Dey, S. Biswas and L. Abualigah, "Efficient Violence Recognition in Video Streams using ResDLCNN-GRU Attention Network," *ECTI-CIT Transactions*, vol. 18, no. 3, pp. 329–341, Jul. 2024.

[34] W. Sae-Tang and A. Sirikham, "Image Steganography-based Copyright and Privacy-Protected Image Trading Systems," *ECTI-CIT Transactions*, vol. 17, no. 3, pp. 358–375, Aug. 2023.

[35] S. Riyana, K. Sasujit, N. Homdoung, T. Chaichana and T. Punsaensri, "Effective Privacy Preservation Models for Rating Datasets," *ECTI-CIT Transactions*, vol. 17, no. 1, pp. 1–13, Nov. 2022.

[36] Kruskal, "Searching, Merging, and Sorting in Parallel Computation," in *IEEE Transactions*

on Computers, vol. C-32, no. 10, pp. 942-946, Oct. 1983.

Surapon Riyana received a B.S. degree in computer science from Payap University (PYU), Chiangmai, Thailand, in 2005. Moreover, He further received a M.S. degree and a Ph.D. degree in computer engineering from Chiangmai University (CMU), Thailand, in 2012 and 2019 respectively. Currently, he is a lecturer in Smart Farming and Agricultural innovation Engineering (Continuing Program), School of Renewable Energy, Maejo University (MJU), Thailand. His research interests include data mining, databases, data models, privacy preservation, data security, databases, and the internet of things.

Nigran Homdoung received a B.S. degree in mechanical engineering from King Mongkut's University of Technology Thonburi (KMUTT), Thailand, in 2001. He received a M.Eng. in Energy Engineering from Chiang Mai University (CMU), Thailand, in 2007. Moreover, he received a D.Eng. In Mechanical Engineering from Chiang Mai University (CMU), Thailand, in 2015. His research interests include biomass technology (gasification and pyrolysis process) and application Internal combustion Engine to biofuels. machine learning, data science, and artificial intelligence.

Kittikorn Sasujit (Assistant Professor) received a B.Eng (Environmental Engineering) in 2004 from Rajamangala University of Technology Lanna, Thailand, and an M. Eng and Ph.D. (Energy Engineering) in 2008 and 2020, respectively, from Chiang Mai University, Thailand. His studies will include biomass technology, wind energy technology, NTP applications for biomass tar removal, and renewable energy.

Noppamas Riyana received a B.S. degree in computer science from Payap University (PYU), Thailand, in 2005. Moreover, she received a M.S. degree in business administration from Payap University (PYU), Thailand, in 2012. Currently, she is a computer technical officer at Maejo University (MJU), Thailand. Her research interests include data mining, databases, data models, and privacy preservation.