Data Science and Engineering
https://doi.org/10.1007/s41019-025-00283-0

RESEARCH PAPERS

Analysis and Multi-objective Protection of Public Medical Datasets from Privacy and Utility Perspectives

Samsad Jahan[1] · **Yong-Feng Ge**[1] · **Enamul Kabir**[2] · **Kate Wang**[3]

Received: 4 August 2024 / Revised: 1 November 2024 / Accepted: 26 January 2025

Abstract

In this era of big data, seamless distribution of healthcare information is crucial for improving patient care and advancing medical research, necessitating meticulous attention to preserving health data privacy. However, overly stringent protection measures can impede the efficient utilization of invaluable resources for medical research and personalized healthcare, posing a central challenge in balancing privacy protection with effective data utilization. This study aims to explore various methods used to protect the privacy of patients' health records, and evaluates their advantages and limitations. Additionally, it conducts an in-depth analysis of a public medical dataset concerning privacy protection, assessing the effectiveness of *k*-anonymity and *l*-diversity privacy criteria and examining the influence of quasi-identifier (QID) attributes on privacy preservation. The study showcases techniques to achieve privacy standards, including generalization and suppression. Furthermore, it introduces a novel approach that utilizes the genetic algorithm (GA) and a non-dominated sorting technique to maximize both privacy and utility in health data through multi-objective optimization. After examining the results, this paper offers a guide for data owners on selecting attributes for medical data publication and choosing suitable privacy preservation strategies. Through the exploration of the GA and the non-dominated sorting approach, this paper suggests that the proposed GA can offer promising non-dominated solutions to the issue of health data privacy in the era of data-driven healthcare. A combination of these algorithms can enhance privacy protection and provide healthcare professionals and researchers with essential knowledge, ultimately benefiting patient care and ensuring a more secure database system.

Keywords Data privacy · *k*-anonymity · *l*-diversity · Genetic algorithm · Multi-objective optimization

1 Introduction

The dissemination of research data is crucial for transparency, reproducibility, and collaboration in scientific pursuits, facilitating verification and improvement of current knowledge [15, 37, 53]. The preservation of patient privacy in medical data is essential due to the sensitive nature of personally identifiable information (PII) and the extensive medical history involved [11, 17, 24, 50]. The unauthorized utilization or reidentification of such data might result in substantial consequences, jeopardizing patient confidentiality, trust, and legal responsibility [40, 49, 54, 55, 57, 58]. Recent data breach occurrences have heightened concerns about the necessity for comprehensive privacy protection methods that reconcile privacy with utility.

Conventional privacy-preserving measures such as *k*-anonymity [46], *l*-diversity [32], and *t*-closeness [29] provide the foundational privacy protection of data, but they are inadequate to protect against advanced and complex

Enamul Kabir and Kate Wang have contributed equally.

✉ Samsad Jahan
samsad.jahan@live.vu.edu.au

✉ Yong-Feng Ge
yongfeng.ge@vu.edu.au

Enamul Kabir
enamul.kabir@usq.edu.au

Kate Wang
kate.wang@rmit.edu.au

[1] Institute for Sustainable Industries and Liveable Cities, Victoria University, 70/104 Ballarat Rd, Footscray, VIC 3011, Australia

[2] School of Mathematics, Physics and Computing, University of Southern Queensland, 487-535 West St, Toowoomba, QLD 4350, Australia

[3] School of Health and Biomedical Sciences, RMIT University, 30 Janefield Drive, Bundoora, VIC 3083, Australia

Published online: 01 April 2025

threats. For instance, while k-anonymity shields data from identity disclosure by hiding it from $k-1$ other records, it fails to thwart linkage attacks. On the other hand, l-diversity protects attributes from being exposed and makes sure that sensitive attributes (SAs) are spread out evenly among groups [38]. However, it cannot prevent the similarity and skewness attacks [23, 29]. Setting a threshold for SAs and effectively managing the distribution of sensitive values within each equivalence class are necessary to achieve t closeness. Achieving this degree of effectiveness can prove challenging since it depends on delicate variables that lie outside the jurisdiction of publishers [29]. A different form of k-anonymity, referred to as p-sensitive k-anonymity, is capable of protecting against identity exposure while still allowing attribute information to be revealed [5]. However, this method is vulnerable to similarity attacks, which have the ability to identify specific places. To tackle this problem, Sun et al. developed the p^+-sensitive k-anonymity and (p, α) -sensitive k-anonymity models. These models are improvements on the p-sensitive k-anonymity model. They are meant to lower the risk of similarity attacks and the possible exposure of attributes [45].

Differential Privacy (DP) offers customizable levels of privacy protection for sensitive data [10]. However, the introduction of noise into the dataset compromises transparency and utility. Numerous studies have focused on the methods to maintain privacy while increasing data utility [43]. Soria-Comas proposed a method that integrates k-anonymity with DP, aiming to improve the utility of differentially private responses [41]. Liu et al. came up with a new way to protect privacy while still giving useful data for data mining. Their method uses conditional probability distribution and machine learning to keep private information safe while still being useful for data mining [30]. But, the trade-off between data usefulness and privacy protection in DP remains an unresolved issue [12, 18, 19]. Furthermore, most privacy-preserving techniques involve generalizing data, suppressing it, or adding noise to anonymize it. These techniques, however, have the potential to significantly reduce data quality and utility.

Given the limitations of conventional approaches, we aim to explore the utilizations of evolutionary algorithms (EAs) with conventional approaches like k -anonymity and want to see their scope in optimization fields to produce near-optimal privacy and utility solutions. Therefore, we formulate our research questions as follows: (i) How can we enhance privacy standards and strike a balance between privacy and utility? (ii) How effectively can genetic algorithms (GAs) be applied to produce near optimal solutions for privacy and utility?

Recently, the use of EAs in privacy-preserving data publishing has emerged as a new dimension that can be very helpful in finding optimal solutions for the trade-offs between privacy and utility issues. EAs, drawing inspiration from natural selection, utilize fundamental operators like selection, crossover, and mutation to efficiently solve diverse optimization problems [8, 14, 28, 35, 52]. These algorithms, renowned for their effectiveness in real-world optimization tasks, employ techniques such as generalization, suppression, and perturbation to sanitize data while preserving its utility [27, 33]. As privacy preservation becomes increasingly critical, various approaches, including set-based EAs and safe federated data-driven methods, have been proposed to address privacy concerns in multi-objective optimization problems [16, 20, 31, 59]. In light of this, we proposed in a recent study that maximizing both privacy and utility can be addressed by formulating a multi-objective optimization problem [22].

GA is a powerful tool for solving complex optimization problems, even in large, intricate search spaces. By using GA to address multi-objective optimization problems, it is possible to find solutions that maximize both privacy and utility. Therefore, it is our natural expectation that it GA can be useful in achieving privacy-utility trade-off. In this article, we first examine the existing privacy measures for preserving medical data and discuss how to achieve higher privacy criteria, specifically k-anonymity and l-diversity. Secondly, our focus is on utilizing GA to find a near-optimal solution for the privacy-utility trade-off. In order to resolve these concerns, our contribution to this research work comprises the following:

1. a comprehensive analysis of anonymization methods for safeguarding public medical data, coupled with an assessment of the privacy protection standards using k-anonymity and l-diversity models for specific combinations of Quasi-identifier (QID) and SA;
2. a demonstration of achieving higher privacy criteria through attribute generalization and record suppression;
3. the formulation of privacy utility trade-off as a multi-objective optimization problem followed by the development of an algorithm that integrates a GA and a non-dominated sorting approach to identify the non-dominated solutions.

This article is organized as follows: Sect. 2 presents the problem statement, Sect. 3 outlines the anonymization approaches used in medical data, Sect. 4 discusses the analysis and protection of data, Sect. 5 introduces our proposed algorithm, and finally, Sect. 6 provides our conclusions.

2 Problem Definition

The primary aim of privacy-preserving data publishing is to modify the original dataset in a way that maintains high data utility while safeguarding privacy. For example,

consider a scenario where a hospital authority wishes to analyze the patterns of a specific disease and plans to release its information for research purposes. Their goal is to ensure maximum privacy protection for the original dataset D by applying techniques such as generalization and suppression to QID attribute combinations. This results in the creation of an anonymized dataset P that aims to achieve optimal privacy and utility.

To address this challenge, we approach it as a multi-objective problem by defining two anonymization objectives.

Definition 1 (Anonymization Objective 1) To maximize privacy requirements for an anonymized dataset P given a utility degree UD such that $(AD(P)) \geq k)$. AD is the anonymity degree assessed by k-anonymity, while UD indicates the utility degree for an anonymous dataset.

Definition 2 (Anonymization Objective 2) To maximize the utility $(UD(P) \geq UD_{threshold})$ for a given AD on an anonymized dataset P. $UD_{threshold}$ represents the data utility's threshold value.

The utility of P depends on its Transparency Degree (TD) [13].

$$\mathrm{TD}(P) = \sum_{i \in P} \mathrm{TD}(i) \tag{1}$$

$$\mathrm{TD}(i) = \sum_{v_g \in i} \mathrm{TD}(v_g) \tag{2}$$

where i represents the specific record in the set P; v_g is the generalized value in record i. TD value of v_g is estimated as follows:

$$\mathrm{TD}(v_g) = \frac{1}{|v_g|} \tag{3}$$

where $|v_g|$ is the number of domain values that are descendants of v_g.

Given the anonymization objectives mentioned above, we can more precisely describe the multi-objective optimization issue using Pareto dominance [36], where AD and UD are simultaneously optimized.

Definition 3 (Pareto Dominance) Let $\mathcal{R}_i$ and $\mathcal{R}_j$ represent two anonymization solutions. Then $\mathcal{R}_i$ is said to have Pareto dominance [56] over $\mathcal{R}_j$, denoted by $\mathcal{R}_i \succ \mathcal{R}_j$, if and only if:

$$\begin{cases} \forall m = 1, 2 \ f_m(\mathcal{R}_i) \geq f_m(\mathcal{R}_j) \\ \exists m = 1, 2 \ f_m(\mathcal{R}_i) > f_m(\mathcal{R}_j). \end{cases} \tag{4}$$

where f_1 represents AD, and f_2 represents UD.

This definition regards non-dominated solutions as the most optimal solution for addressing multi-objective optimization problems.

3 Approaches to Anonymizing Medical Data

This section provides a concise overview of prominent anonymization methods, including k-anonymity, l-diversity, and DP, which are effective in preserving the privacy of medical data.

3.1 *k*-anonymity

Many data custodians, such as governmental bodies and healthcare institutions, contend that omitting specific information such as name, address, and phone number will ensure data anonymity. Nevertheless, correlating data with other published datasets, such as voter registries, may result in the erosion of anonymity. Incorporating noise into the dataset may yield erroneous statistical outcomes [3, 51]. To address these issues Samarati demonstrated the application of k-anonymity in safeguarding data privacy using methods such as generalization and suppression. In addition, the authors established the notion of minimum generalization, which ensures that the published data retains its characteristics while maintaining k-anonymity [39]. The k-anonymity paradigm, introduced by Sweeney in 2002, has gained significant popularity for safeguarding individual privacy owing to its straightforwardness and efficacy [46].

Definition 4 (k-anonymity) A dataset is said to satisfy k-anonymity if it contains a minimum of k records for each possible combination of QID attributes.

k-anonymity is a strategy for protecting the identity and confidential data of individuals within a dataset. It assures that each element of the data set cannot be distinguished from at least $k-1$ other entries using specific identifying characteristics or QIDs. By doing this, it helps to ensure the confidentiality and protection of the individuals whose data are being studied [46]. Literature encompasses various forms of k-anonymity, including the k-join-anonymity model [42], cluster-based anonymity [3], k-anonymity in DP [41], and the microaggregation sorting framework [25], among other examples. Despite the widespread adoption of k-anonymity as a method of data anonymization, it is important to recognize its limitations and vulnerabilities. Therefore, the pros and cons of k-anonymity are listed below:

3.1.1 Pros

- It gives protection against identity disclosure, a severe breach where the attacker identifies an individual within a dataset [1].
- This method requires less computational cost than other anonymization methods, such as cryptographic methods [38].
- This method is well known for its simplicity. It provides enhanced protection when incorporated with clustering techniques.

3.1.2 Cons

- It is susceptible to several types of privacy breaches, including membership disclosure, where an attacker can infer an individual's presence in a dataset; and attribute disclosure [2, 4], where SAs of individuals can be inferred, even if their identities remain obscured.
- When there is not enough dissimilarity in the SAs, it can generate a cluster that might disclose information and compromise privacy.

3.2 *l*-diversity

To tackle the problems related to uniformity and prior knowledge, it is imperative to use a more comprehensive privacy framework known as *l*-diversity. Advancing beyond *k*-anonymity, *l*-diversity secures sensitive data by ensuring that each SA contains a minimum of *l* indistinguishable values, thereby reducing the risk of attribute disclosure attacks. These attacks exploit dataset patterns to deduce specific sensitive information, posing a significant risk. By diminishing the chance of re-identification and improving data privacy, *l*-diversity introduces an essential layer of protection [32]. Unlike simpler models, the *l*-diversity approach specifically targets SAs and enhances privacy by ensuring that each group contains at least *l* distinct values. This method not only focuses on the diversity within these groups but also considers QIDs to mitigate privacy breaches associated with SAs. The definition of *l*-diversity is as follows:

Definition 5 (*l*-diversity) A dataset is considered to satisfy the *l*-diversity if, for each QID group, there is a minimum of *l* well-represented values for the SA.

The advantages and weaknesses of *l*-diversity is given below:

3.2.1 Pros

- Prevents the risk of attribute disclosure and provides better privacy protection than *k*-anonymity.
- It can provide robust protection against background information attacks by implying a diverse distribution of SAs.

3.2.2 Cons

- Achieving *l*-diversity is more complex than *k*-anonymity. It needs careful manipulation, which leads to increased computational costs [7].
- Sometimes this method depends on the range of SAs, if the number is lower than the privacy parameter *l*, some fictitious data are added to achieve the *l* standard, which may give bias results of the analysis.
- Despite the implementation of *l*-diversity, there is a potential vulnerability in which the SA values can still be exposed through skewness attacks and similarity attacks [2, 29, 44].

The data shown in Table 1 exemplify the characteristics of a 3-anonymous and 2-diverse dataset. Each set of data, categorized by the QIDs, contains a minimum of three identical records. Additionally, within each set, there are a minimum of two distinct pieces of information for the SA. Nevertheless, it is crucial to acknowledge that the efficacy of *l*-diversity is contingent upon the spectrum of values for the SA.

3.3 Differential Privacy

Medical data comprises a wide range of highly sensitive and confidential information, such as diagnoses, genetic information, geographical data, and any other details related to health [26]. DP enhances the model's ability to withstand security breaches, preventing efforts to obtain accurate responses to falsified queries. It is the most popular privacy-preserving strategy that significantly enhances privacy by addressing all vulnerabilities associated with data anonymization methods without making any assumptions about the

Table 1 Example of a 3-anonymous and 2-diverse patient micro-dataset

No.	QID			SA
	Age	Zip Code	Gender	
1	<30	960**	Male	Heart Disease
2	<30	960**	Female	Heart Disease
3	<30	960**	Female	Cancer
4	3*	963**	Male	Heart Disease
5	3*	963**	Male	Diabetes
6	3*	963**	Male	Diabetes
7	>40	965**	Female	Diabetes
8	>40	965**	Male	Cancer
9	>40	965**	Female	Cancer

background knowledge of prospective adversaries [10]. By implementing DP methodologies, it is possible to ensure the confidentiality of patients' personal information while still enabling the analysis and research of healthcare data.

Definition 6 (ϵ-DP) In the context of two datasets M and N that are distinct in a single element, the ϵ-DP property holds true for a randomized function f if the following condition is satisfied for each subset S of its range:

$$P[f(M) \in S] \leqslant e^{\epsilon} P[f(N) \in S] \tag{5}$$

where the parameter ϵ governs the level of privacy protection.

According to this definition, when comparing two databases that vary by a single record, a differentially private method yields two randomized outputs with nearly identical probability distributions. This makes it unlikely that an opponent would be able to deduce the presence of a specific victim in the released database with high confidence [7].

DP mechanisms frequently employ the Laplace, Exponential, and Gaussian mechanisms to ensure anonymity. The Laplace mechanism is often used for numerical outputs, while the Exponential mechanism is better suited for non-numeric queries [48]. The Gaussian technique is effective for aggregating sensitive information, conducting private data analysis (e.g., regression analysis or clustering), and implementing machine learning applications [21]. The advantages and disadvantages of DP mechanism are listed below:

3.3.1 Pros

- DP provides strong privacy, and it does not require any assumption of background information of a potential adversary.
- It can preserve better utility for low-sensitive queries such as count, range, etc.
- It provides strong resilience to homogeneity, background information, and skewness attacks.

3.3.2 Cons

- A database with DP protection can provide inaccurate results for highly sensitive queries.
- The addition of noise can reduce the accuracy of the data for small datasets with a high privacy budget.
- The addition of noise in the dataset may lead to a loss of information.
- Determining the optimal privacy budget is not easily possible. A high privacy budget gives insufficient privacy, while a low budget degrades the data utility.

Table 2 provides the overall summary and comparison of the conventional privacy-preserving approaches mentioned in this study. Due to simplicity and less computational complexity of k-anonymity and l-diversity than DP, we analyze the privacy protection standard and how to achieve higher values of k and l in the next section.

4 Analysis and Protection

The objective of our research is to examine the privacy measures employed in publicly available medical data by quantifying the distribution of k and l values for particular combinations of QID and SA. To address privacy issues, we alter the combinations of QID and determine the associated k and l values, which we then assess in terms of their distribution. By analyzing the distribution of k and l, we calculate their means and compare them across various combinations of QID attributes. The aim of our study is to identify the traits that are most vulnerable to privacy breaches. We have observed that lower values of both k and l significantly increase the probability of identification. Subsequently, we compute the record suppression ratio for various values of k and l that demonstrate a balance between privacy and utility.

Table 2 Comparison of anonymization approaches on various aspects

Features	k-Anonymity	l-Diversity	DP
Privacy	No records can be identified from $k-1$ other records in the group	Guarantees privacy by distributing the data in l well-distributed SA	Guarantees privacy by mathematically proven formula
Vulnerability	Attribute Disclosure	Skewness and similarity attack	Resilient to all common attacks
Utility loss	Reduces utility due to generalization	Reduces data utility for achieving l diversity	Adding noise reduces utility
Computational complexity	Simple	Moderate	High complexity due to rigorous mathematical calculation
Privacy-utility trade-off	Depends on tuning the value of k	Depends on tuning l	Depends on tuning ϵ

4.1 Privacy Analysis

In our investigation, we utilize a publicly available dataset of hospital inpatient discharges provided by the New York State Department of Health.[1] From this dataset, we selected 1,020 records for our study. Initially, our focus was directed toward a specific set of characteristics known as QID 1. This set includes information such as {health service area, hospital county, operating certificate number, facility ID, facility name, age group, and zip code}. In this context, SA is considered to be the "CCS diagnosis description". For QID 1, the estimated distribution of k and l values yields an average of 24.238 and 8, respectively.

Subsequently, we investigate another QID attribute combination, labeled as QID 2, which comprises health service area, hospital county, age, zip code, gender, and race, with the SA remaining unchanged. QID 2 encompasses the most prevalent QID attributes. Analysis of QID 2 reveals mean values of $k = 13.9589$ and $l = 5.41$. This comparison clearly indicates that QID 1 offers superior privacy protection compared to QID 2. Further exploration involves observing the impact of removing individual QID attributes from QID 2. Removing the race attribute resulted in QID 3: {health service area, hospital county, age, zip code, gender}, with average k and l values of 16.70492 and 6.295082, respectively. Notably, there is a slight improvement in the average k and l values after removing the race attribute, suggesting its significance as a key QID within the dataset.

Similarly, removing the zip code from QID 2 led to QID 4: {health service area, hospital country, age, gender, race} with average k and l values of 26.12821 and 7.564103, respectively (Fig. 1). This indicates the importance of zip code as a significant QID, as evidenced by the increased averages.

The elimination of gender from QID 2 led to the creation of QID 5, which includes {health service area, hospital country, age, zip code, race}. The average values for k and l in QID 5 are 20.38 and 6.74, respectively. This illustrates that the dataset's level of privacy increases when gender is removed from the list of attributes. After removing the age from QID 2, the new combination, QID 6: {health service area, hospital country, zip code, gender, race}, results in average values of k and l equal to 37.74074 and 10.40741, respectively (Fig. 1). The research identifies age as the primary factor that affects the privacy of hospital inpatient discharge data. Hence, the disclosure of age information should be handled with more prudence, given its substantial influence as personally identifiable data.

[1] https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8.

4.2 Improving *k* and *l*

Despite the implementation of some de-identification measures in this dataset, our analysis reveals that it still lacks sufficient protection. There remains a risk that attackers can easily identify unique information. Therefore, enhancing the values of k and l is imperative for better protection. This enhancement can be achieved through attribute generalization and record suppression techniques [13, 16]. To illustrate attribute generalization, we provide a simple example and demonstrate record suppression through a small experimental study.

4.2.1 Attribute Generalization

Generalization is a method used to represent attribute values in a table more abstractly and facilitate the identification of tuples. This technique involves transforming attribute values into broader categories within a universal domain. To preserve data integrity, QID attributes, such as zip codes, can be generalized from a specific level (e.g. Z_0 with values like 04123, 04126) to a more general level (e.g. Z_1 with values like 04120, 04120). This transformation follows a "domain generalization hierarchy". If the table already meets the requirement of k-anonymity, the technique of k-minimum generalization can be applied to protect privacy while still retaining specific values in private tables [39, 46]. However, this method may need greater generality when dealing with outliers or tuples that appear less frequently than k times [38].

Consider the case of a 19-year-old female living in Allegany with a zip code of 96040, who has been diagnosed with a mental disorder, and whose family desires to protect the privacy of her personal details. Due to the distinctiveness of her information in the dataset (see Table 3), she can be easily recognized. Using attribute generalization in this scenario can improve the values of k and l. By obfuscating the zip code, gender, and age, her personal data becomes less distinguishable in the publicly available dataset. An example demonstrating attribute generalization and its impact on improving k and l values is illustrated in Table 3.

In this scenario, the process of attribute generalization is applied to three attributes - age, zip code, and gender. Age is classified into groups such as 'less than 30', '3*', and 'greater than 40 '. The zip code is generalized by hiding the last two digits. Gender is generalized and referred to as 'individual'.

4.2.2 Record Suppression

The main purpose of this method is to hide the complete tuple t from the public dataset since it includes sensitive and non-sensitive data. However, excluding particular

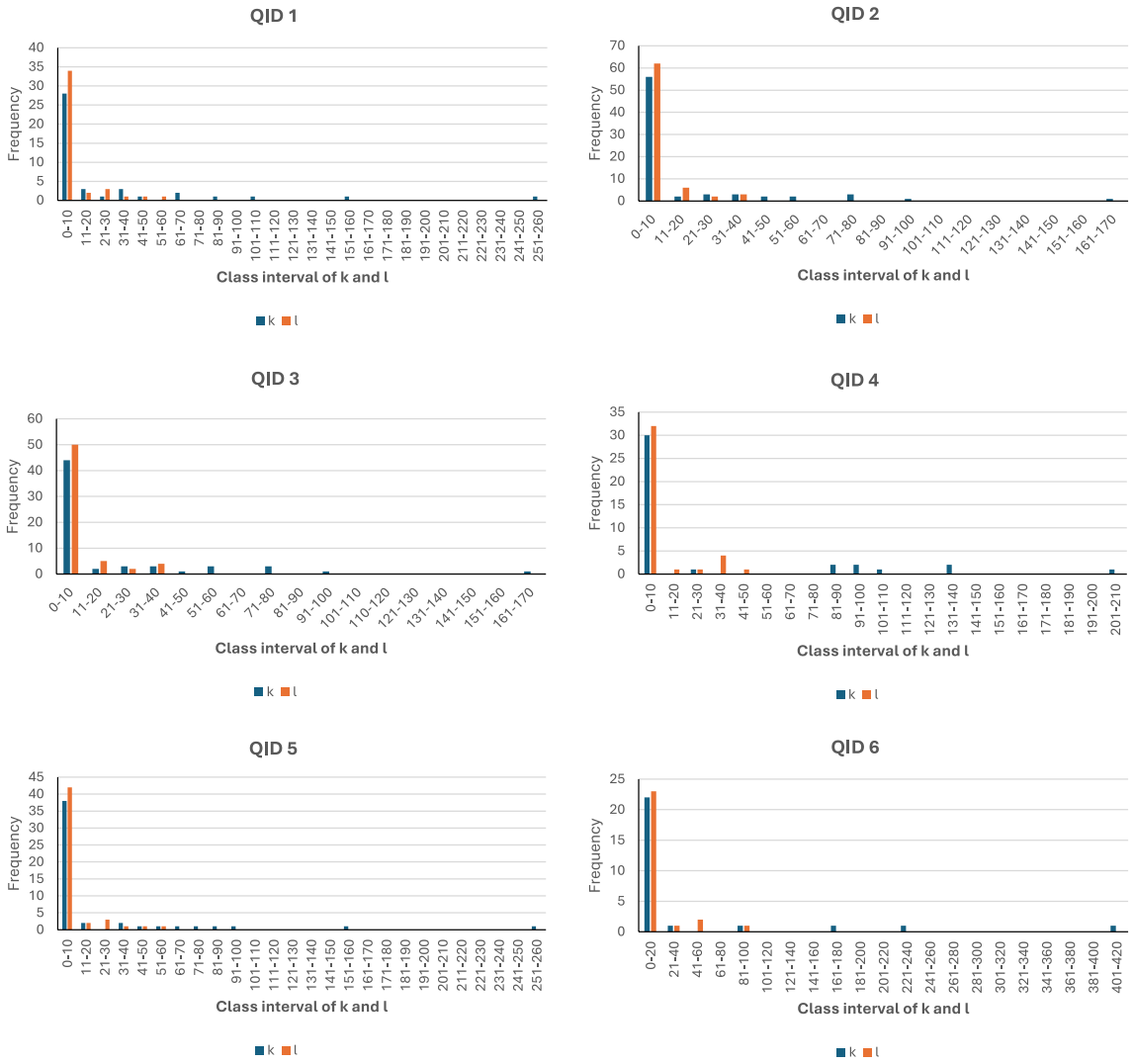

Fig. 1 Distribution of k and l for different QID attributes

Table 3 Example of attribute generalization

No.	QID			SA
	Age	Zip Code	Gender	
1	<30	960**	Individual	Heart Disease
2	<30	960**	Individual	Heart Disease
3	<30	960**	Individual	Cancer
4	<30	960**	Individual	Mental illness
5	3*	963*	Individual	Heart Disease
6	3*	963*	Individual	Diabetes
7	3*	963*	Individual	Diabetes
8	>40	965**	Individual	Diabetes
9	>40	965**	Individual	Cancer
10	>40	965**	Individual	Cancer

information from the dataset could potentially undermine its precision [6].

Our study focuses on improving the values of k and l through record suppression. Higher values of k and l lead to decreased identifiability of the information, thus enhancing the privacy of the data. It is crucial to increase the value of k and l to strengthen privacy protections. Lower values of k and l indicate that the record is unique and easily identifiable within the publicly available data. By eliminating tuples that contain unique information, we can raise the values of k and l for our data to 2, potentially increasing the confidentiality of the data. To protect privacy, we conduct a small study to determine the frequency of suppression of records at various levels of k anonymity and l diversity. We eliminate all records that contained a combination of QID attributes with k and l values of 1, with the objective of increasing these values to 2. Next, we evaluate the percentage of suppressed

records, termed the record suppression ratio. Furthermore, we repeat the procedure for k values between 3 and 10 and l values between 2 and 8.

The fluctuation of the record suppression ratio depicted in Fig. 2 shows the trade-off between privacy and utility in this dataset. The figure clearly demonstrates that an increased record suppression ratio is directly correlated with a higher level of privacy. Consequently, as the quantity of records in the dataset declines, there is a corresponding decrease in the dataset's utility. Hence, it is imperative to optimize the anonymization method in order to maintain the highest possible level of utility when addressing privacy needs [16].

5 Non-dominated Sorting with Genetic Algorithm

In this section, we propose an algorithm that addresses maximizing both privacy and utility as a multi-objective optimization problem. This algorithm utilizes GA and a non-dominated sorting procedure to identify non-dominated solutions within the k-anonymity privacy protection model.

GA is a search metaheuristic within artificial intelligence (AI) inspired by the evolutionary processes of biological organisms. As part of the evolutionary algorithm family, GAs are designed to mimic natural processes, including inheritance, mutation, selection, and crossover [47]. Researchers use these algorithms to produce optimal or near-optimal solutions for multi-objective optimization problems. For example, Liu et al. introduced a safe federated data-driven evolutionary multi-objective optimization method. This approach ensures the protection of the original data and the newly generated solutions produced by optimizing the acquisition function [31]. The authors in [16] discussed the challenges associated with maintaining privacy while releasing data and used GA such as the Information-Driven Genetic Algorithm (ID-GA) as a solution. This algorithm incorporates an information-driven crossover operator, an information-driven mutation operator, and a two-dimensional selection operator, specifically developed to facilitate information sharing among anonymization solutions, facilitate information disclosure, and assess the characteristics of various solutions. Gong et al. focused on the complex issue of interval many-objective optimization problems (IMaOPs) and presented a set-based EA as a viable solution. Their approach converts the issue into a deterministic bi-objective problem and adds additional goals of hyper-volume and imprecision [20]. In our proposed algorithm, we use a non-dominated sorting approach during the population selection stage to provide non-dominated solutions. Our method begins by creating a population of random individuals, where each person has two vectors: Individual A represents attribute generalization, and Individual B represents record suppression. We assess each individual's fitness based on privacy and transparency. The following sections will cover the representations of individuals, crossover, and selection techniques.

5.1 Data Representation

A sample dataset and its anonymization solution are presented in Fig. 3. The figure displays three QIDs and four records. The figure also displays an anonymization solution, where vector 'A' signifies attribute generalization and vector 'B' signifies record suppression. Vector 'A' generalizes each QID attribute according to its level, while vector 'B' suppresses each record based on its corresponding value. A value of '0' indicates the deletion of the related record, while '1' indicates its retention.

Figure 3 shows that age, gender, and zip code are the three QIDs, and they are generalized according to levels 1, 2, and 3. The information for records 1, 2, and 3 will be released, while record 4 will be suppressed.

5.2 Crossover

We employ a crossover operator to recombine parents and produce new offspring. At each gene position, a random

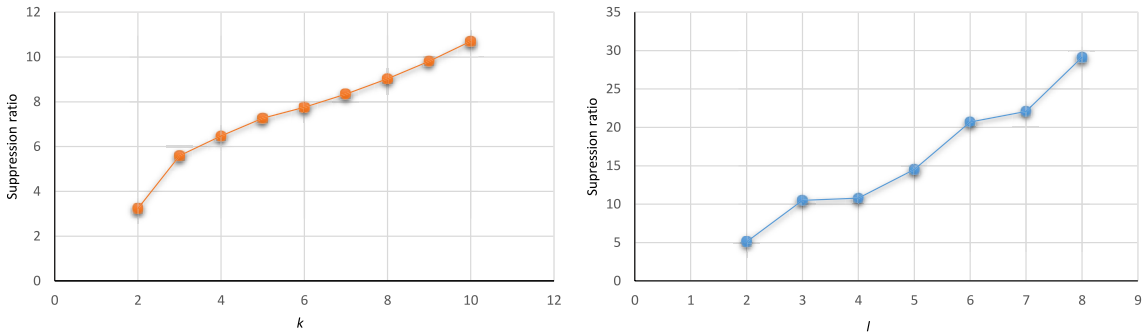

Fig. 2 Record suppression ratio for various k and l

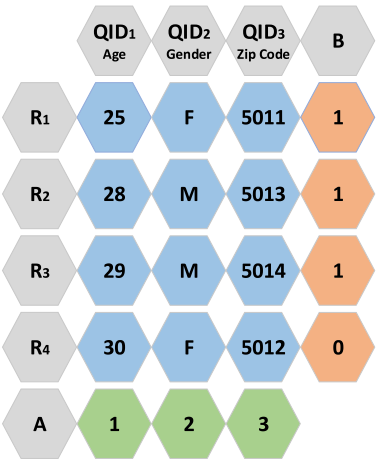

Fig. 3 Illustration of representation of a dataset with three QID attributes and four records

number between 0 and 1 is generated. If the number is lower than 0.5, we select the gene from the mother for the child; if it is 0.5 or greater, we choose the gene from the father. We repeat this process for all the positions of the genes, resulting in an offspring that combines genes from both parents.

In GAs, the crossover operator exchanges information between parent individuals. There are two methods to accomplish this. First, we randomly exchange the values of two '*A*' vectors, resulting in a mix of generalization levels between these two anonymization solutions. If the parents have met the required level of privacy preservation, their offspring are likely to do the same. Second, the system gathers the values of two '*B*' vectors. This implies that the disclosure of one record from an anonymization solution also releases the equivalent record in the offspring solution, potentially revealing additional information. Figure 4 provides an illustration of the crossover operator. Consider a population size of 30, divided into 15 groups, with each group consisting of two individuals: a father and a mother. For two generalization vectors, assume that A_1 is from the father individual and A_2 is from the mother individual. For the first gene position, the random number is less than 0.5; therefore, the information from the mother is chosen in the offspring. For the second gene position, the random number is greater than 0.5, so the information from the father is selected. At the third gene position, the random number is again less than 0.5, so the mother's information is chosen for the offspring. Regarding the suppression vectors, B_1 and B_2, if one parent releases information, it will also be released in the offspring. However, if both parents do not release the information, the record will remain unreleased in the offspring. This process explains how the offspring A_{1X2} and B_{1X2} are generated, as depicted in the figure.

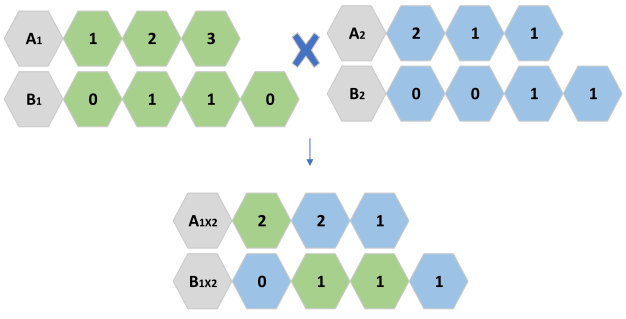

Fig. 4 An example of a crossover operator, which swaps the information in two anonymization solutions

5.3 Mutation

The mutation operator is a technique employed in genetic search processes to prevent entrapment at local optima. It randomly introduces new information into the process, which helps to differentiate an individual's chromosomes from those of the parent. Operating at the bit level, the mutation operator allows for the possibility that any bit may undergo alteration during the transfer of bits from the existing chromosome to the new one. This probability of mutation is typically denoted as the mutation probability, which is kept very low [34]. This paper randomly alters the information within vectors A and B, setting the mutation probability at 0.1. An illustration of the mutation operator is provided in Fig. 5. In this figure, following the crossover operation, the third position of the individual vector $A_{1\times2}$ is altered from 1 to 2 in the mutant version of A_1^* and the fourth position of $B_{1\times2}$ is modified from 1 to 0 in the mutant version B_1^*.

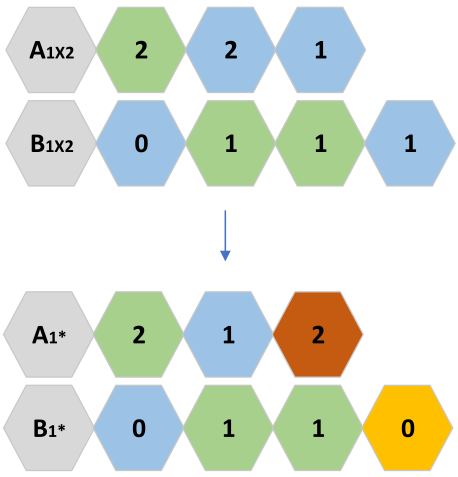

Fig. 5 An example of mutation operator, which changes two vectors in the solution independently

5.4 Non-dominated Sorting as Selection

The main idea of non-domination is to identify solutions A and B, classifying A as dominant if it outperforms B in all objectives or equal in at least one. This strategy aims to classify the solutions into various non-dominated fronts depending on the dominance relationships, which addresses the computational complexity of earlier approaches. The two major elements of this design are domination count and dominated set of solutions. The domination count represents the number of solutions that dominate a given solution, while the dominated set preserves the list of solutions that each solution dominates. The process starts with the initialization of an empty list for the dominated set of each solution and the setting of the domination count to zero. It subsequently compares each pair of solutions to ascertain dominance, thereby classifying them into non-dominated fronts. The domination count is reduced for each solution, and if it is zero, it is added to the next non-dominated front. The procedure is repeated until all solutions have been assigned. The overall total complexity of this procedure is $O(MN^2)$, where M is the number of objectives and N is the population size.

The proposed non-dominated sorting strategy follows the selection strategy given in NSGA-II [9]. It starts by initializing a population of individuals and ranked based on their fitness value. The population is then sorted according to the non-domination level, with the first front containing individuals not dominated by others, the second front containing individuals only dominated by the first front, and so on. Crowding distances are calculated within each front to promote better diversity. The algorithm randomly selects two individuals with better rank and crowding distance, then performs mutation and crossover operations. The next generation is created by combining the parents and offspring population, and the best individuals are selected based on non-domination rank and crowding distance for the next generation.

The selection criteria begin by initializing the domination count to zero for each individual in the population. This count helps us to track how many times each individual is dominated by others. Then, we iterate through two nested loops over all pairs of individuals in the population to determine which individual dominates others by comparing their fitness vectors. To check for non-dominance, we use the 'dominates' function. Additionally, we evaluate each anonymization method using two indicators, namely AD and TD, to assess its quality. We then sort the population based on various levels of non-domination. Specifically, for two individuals i and j, individual i dominates individual j in the following scenario:

1. the AD of individual i is greater than the AD of individual j and also the TD of individual i is greater than the TD of individual j;
2. the AD of individual i is equal to AD of individual j but the TD of individual i is greater than TD of individual j;
3. the AD of individual i is greater than the AD of individual j and the TD of individual i is equal to the TD of individual j.

When an individual i dominates another individual j, the domination count of j increases, and vice versa. The algorithm first calculates the domination count for each solution and then sorts the indices of the solutions in ascending order based on their domination counts and in descending sequence based on their privacy degree. Then, it maintains the population size by selecting a subset of indices for the next generation. Finally, it updates the population and fitness vector lists using the selected indices and conducts a size check to maintain the desired population size. This selection criteria helps us identify non-dominated solutions that balance AD and TD, ensuring the population satisfies privacy and utility requirements during updates.

Algorithm 1 Pseudo-code of Proposed GA

```
Set generation index A = 0
Initialize population P
Evaluate population P
while stopping criterion is not met do
    for each two parent individuals in P do
        Perform crossover operator on two individuals and generate offspring
        Execute mutation operator on offspring
        Evaluate offspring
    end for
    Perform non-dominated sorting to select population P
    A = A + 1
end while
Output non-dominated solutions
```

5.5 Overall Process of the Algorithm

Algorithm 1 is an optimization algorithm that begins by initializing a population of candidate solutions. Each individual in the population is evaluated on the basis of their fitness function, which assesses their effectiveness in solving the optimization problem. The algorithm then enters a loop that continues until a stopping criterion is met. In each iteration of the loop, the algorithm selects pairs of parent individuals from the population and applies a crossover operator to produce offspring. This operator exchanges genetic information between parents to generate new individuals. Subsequently, the offspring undergoes a mutation, which introduces small random changes to their genetic information. The fitness of the offspring is evaluated, and this process repeats for all pairs of parents.

After generating and evaluating offspring for all pairs of parents, a selection mechanism is employed to determine

individuals for the next generation. This algorithm utilizes the non-dominated sorting mechanism to rank individuals according to their dominance relationship with each other. The next generation is subsequently formed by selecting individuals from the current generation based on their rank and fitness. This evolutionary loop persists until the stopping criterion is satisfied. For the stopping criterion, we have set a maximum fitness evaluation number. The loop meets the stopping criterion once it reaches this number. Ultimately, the algorithm outputs the non-dominated solutions in the last generation.

5.6 Non-dominated Solutions

After applying the non-dominated sorting selection criterion in our proposed algorithm, we obtain non-dominated solutions. In Fig. 6, we can see that our algorithm can produce non-dominated solutions of Pareto front for various test cases. This indicates all the solutions are optimal. Within this figure, we showcase the non-dominated solutions for both AD and TD. Notably, in test case 4, we observe that for AD values surpassing 300, TD exhibits a significantly low value, approximately 250. In contrast, for lower AD values, specifically AD=1, TD reaches its peak around 2400. In test case 5, we notice a similar trend: the highest recorded AD level is 165, corresponding to a TD of 215, while the lowest AD level, at 1, corresponds to a TD of 1590. This pattern persists in other test cases as well, indicating that AD attainment aligns with the underlying data patterns. For example, in test case 11, only six pairs of AD and TD values are observed. We derive these non-dominated solutions by simultaneously fulfilling both objectives, ensuring that improvements in one do not compromise the other. For each testcase, in the top left of Fig. 6, it is also clear that one can achieve optimal utility when privacy is compromised. Conversely, the bottom right of each test case's figure clearly shows that optimal privacy can be achieved once the utility is no longer required. In between, we can find some near-optimal trade-off solutions that meet the user's preferences. In summary, our findings consistently reveal an inverse relationship between AD and TD: as AD increases, TD decreases, and conversely, as AD decreases, TD tends to increase.

5.7 Experimental Tools and Environment

This section outlines the test cases, parameter configurations, and details of the algorithm used in the experiment.

5.7.1 Test Cases

We conduct experiments on our method using 16 distinct test cases derived from a publicly available dataset provided by the New York State Department of Health.[2] The characteristics of these test cases, labeled as T_1 to T_{16}, are described in Table 4. In this context, A_n denotes the attribute number, QID_n denotes the QID attribute number, and R_n denotes the record number. For every test case, we set the privacy criterion k as 2.

5.7.2 Parameter Settings

The population size in our approach is fixed at 30, while the mutation rate is set at 0.1. The maximum number of fitness evaluations is determined by multiplying QID_n by R_n.

5.7.3 Algorithm Implementation

Our proposed algorithm is implemented in Python 3.11 (64-bit) on a Windows 10 system with an Intel(R) Core(TM) i5–8500 CPU@3.00 GHz and 8.00 GB RAM.

6 Conclusion

This article discusses measures taken to protect the privacy of a publicly accessible medical dataset. We have examined current privacy models, identified their strengths and weaknesses, and suggested ways to enhance privacy protection through techniques such as attribute generalization and record suppression. Additionally, we propose an algorithm that utilizes a GA and a non-dominated sorting technique to optimize both privacy and utility as a multi-objective optimization problem. Our experiment aims to assess the trade-off between privacy and utility by examining the equilibrium between information loss and privacy protection. The results demonstrate that our suggested algorithm produces solutions not dominated by any other in the population, effectively achieving the highest possible values for both privacy and utility. Incorporating DP alongside k-anonymity and l-diversity in the future could provide an additional level of protection for medical datasets. This framework has the potential to enable healthcare facilities to securely disseminate data while preserving individual privacy.

[2] https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8.

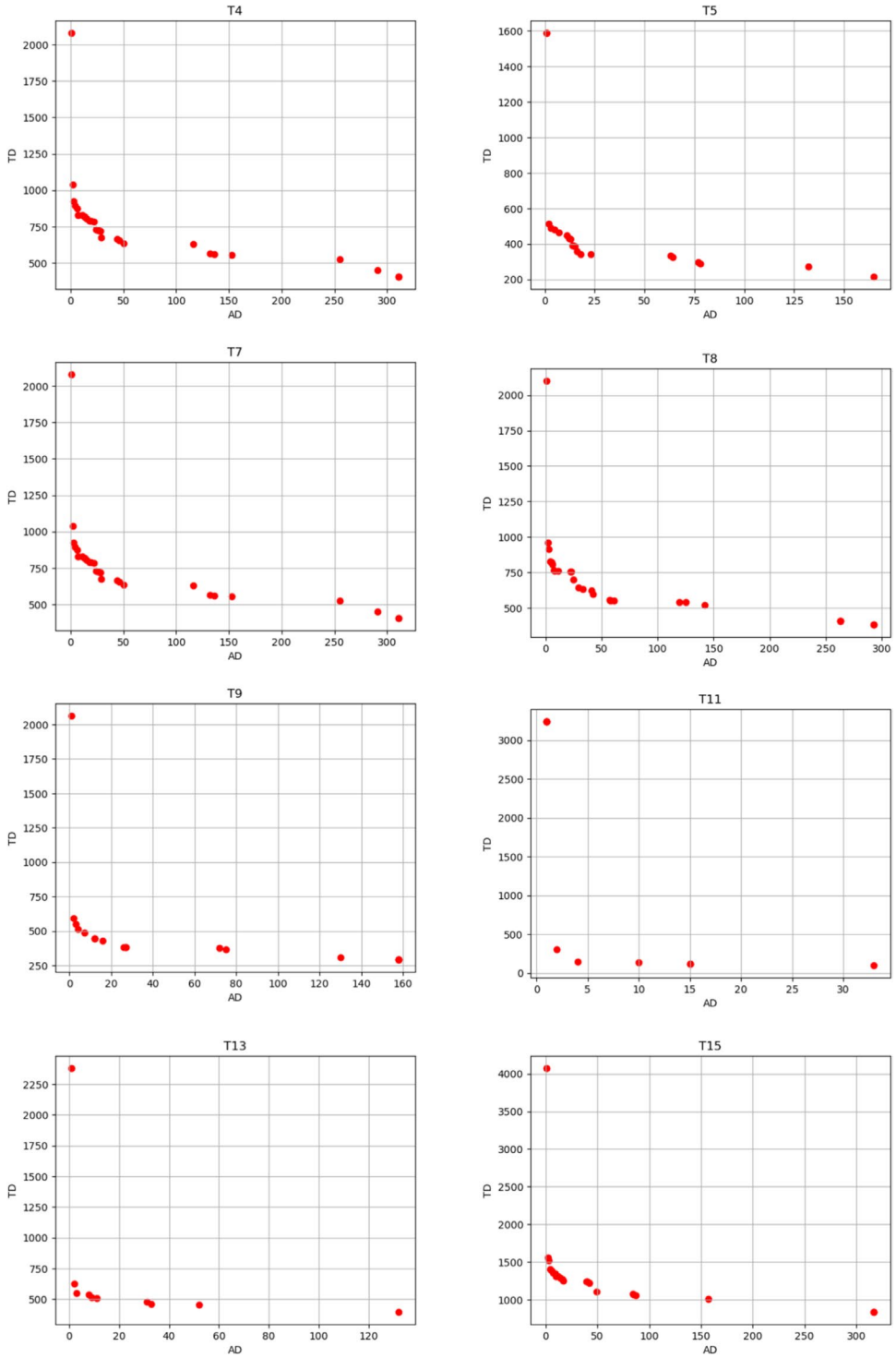

Fig. 6 Non-dominated solutions of our proposed algorithm

Table 4 Properties of 16 test cases [16]

Test cases	A_n	$QID\ n$	R_n
T_1	16	8	200
T_2	16	8	200
T_3	16	8	400
T_4	16	8	400
T_5	18	10	200
T_6	18	10	200
T_7	18	10	400
T_8	18	10	400
T_9	20	12	200
T_{10}	20	12	200
T_{11}	20	12	400
T_{12}	20	12	400
T_{13}	22	14	200
T_{14}	22	14	200
T_{15}	22	14	400
T_{16}	22	14	400

Data availability The data utilized in this paper is available at: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

Ethical approval Not applicable.

References

1. Anjum A, Raschia G (2017) Banga: an efficient and flexible generalization-based algorithm for privacy preserving data publication. Computers 6(1):1
2. Anjum A, Choo KKR, Khan A et al (2018) An efficient privacy mechanism for electronic health records. Comput Secur 72:196–211
3. Belsis P, Pantziou G (2014) A k-anonymity privacy-preserving approach in wireless medical monitoring environments. Pers Ubiquitous Comput 18:61–74
4. Bhuiyan MZA, Wang G, Choo KKR (2016) Secured data collection for a cloud-enabled structural health monitoring system. In: 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), IEEE, pp 1226–1231
5. Campan A, Truta TM, Cooper N (2010) P-sensitive k-anonymity with generalization constraints. Trans Data Priv 3(2):65–89
6. Carvalho T, Moniz N, Faria P, et al (2022) Survey on privacy-preserving techniques for data publishing. arXiv preprint arXiv:2201.08120
7. Chong KM (2021) Privacy-preserving healthcare informatics: A review. In: ITM Web of Conferences, EDP Sciences, p 04005
8. Deb K, Anand A, Joshi D (2002) A computationally efficient evolutionary algorithm for real-parameter optimization. Evolut Comput 10(4):371–395
9. Deb K, Pratap A, Agarwal S et al (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evolut Comput 6(2):182–197
10. Dwork C (2008) Differential privacy: A survey of results. In: Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5, Springer, pp 1–19
11. El Emam K, Dankar FK (2008) Protecting privacy using k-anonymity. J Am Med Inf Assoc 15(5):627–637
12. Ficek J, Wang W, Chen H et al (2021) Differential privacy in health research: a scoping review. J Am Med Inf Assoc 28(10):2269–2276
13. Fung BCM, Wang K, Chen R et al (2010) Privacy-preserving data publishing: a survey of recent developments. ACM Comput Surv 42(4):1–53. https://doi.org/10.1145/1749603.1749605
14. Ge YF, Cao J, Wang H et al (2021) Set-based adaptive distributed differential evolution for anonymity-driven database fragmentation. Data Sci Eng 6(4):380–391
15. Ge YF, Orlowska M, Cao J et al (2022) MDDE: multitasking distributed differential evolution for privacy-preserving database fragmentation. VLDB J. https://doi.org/10.1007/s00778-021-00718-w
16. Ge YF, Wang H, Cao J, et al (2022b) An information-driven genetic algorithm for privacy-preserving data publishing. In: International Conference on Web Information Systems Engineering, Springer, pp 340–354
17. Ge YF, Zhan ZH, Cao J et al (2022) DSGA: a distributed segment-based genetic algorithm for multi-objective outsourced database partitioning. Inf Sci 612:864–886. https://doi.org/10.1016/j.ins.2022.09.003
18. Ge YF, Bertino E, Wang H et al (2023) Distributed cooperative coevolution of data publishing privacy and transparency. ACM Trans Knowl Discov Data. https://doi.org/10.1145/3613962
19. Ge YF, Wang H, Bertino E et al (2023) Evolutionary dynamic database partitioning optimization for privacy and utility. IEEE Trans Dependable Secur Comput. https://doi.org/10.1109/tdsc.2023.3302284
20. Gong D, Sun J, Miao Z (2018) A set-based genetic algorithm for interval many-objective optimization problems. IEEE Trans Evolut Comput 22(1):47–60. https://doi.org/10.1109/tevc.2016.2634625
21. Hu J, Sun K, Zhang H (2022) Helmholtz machine with differential privacy. Inf Sci 613:888–903
22. Jahan S, Ge YF, Kabir E, et al (2023) Analysis and protection of public medical dataset: From privacy perspective. In: International Conference on Health Information Science, Springer, pp 79–90
23. Jain P, Gyanchandani M, Khare N (2016) Big data privacy: a technological perspective and review. J Big Data 3:1–25
24. Kabir ME, Wang H (2009) Conditional purpose based access control model for privacy protection. Proceedings of the

Twentieth Australasian Conference on Australasian Database-Volume 92:135–142
25. Kabir ME, Mahmood AN, Wang H et al (2020) Microaggregation sorting framework for k-anonymity statistical disclosure control in cloud computing. IEEE Trans Cloud Comput 8(2):408–417. https://doi.org/10.1109/tcc.2015.2469649
26. Kong L, Wang L, Gong W et al (2021) LSH-aware multitype health data prediction with privacy preservation in edge environment. World Wide Web 25:1793–1808
27. Li JY, Zhan ZH, Wang H et al (2020) Data-driven evolutionary algorithm with perturbation-based ensemble surrogates. IEEE Trans Cybern 51(8):3925–3937
28. Li JY, Du KJ, Zhan ZH et al (2022) Distributed differential evolution with adaptive resource allocation. IEEE Trans Cybern 53(5):2791–2804
29. Li N, Li T, Venkatasubramanian S (2006) t-closeness: privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering, IEEE, pp 106–115
30. Liu C, Chen S, Zhou S et al (2019) A novel privacy preserving method for data publication. Inf Sci 501:421–435. https://doi.org/10.1016/j.ins.2019.06.022
31. Liu Q, Yan Y, Ligeti P et al (2023) A secure federated data-driven evolutionary multi-objective optimization algorithm. IEEE Trans Emerg Top Comput Intell 8:191–205
32. Machanavajjhala A, Kifer D, Gehrke J et al (2007) l-diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov Data (TKDD) 1(1):3
33. Mandapati S, Bhogapathi RB, Chekka RB (2013) A hybrid algorithm for privacy preserving in data mining. Int J Intell Syst Appl 5(8):47–53
34. Mathew TV (2012) Genetic algorithm. Report submitted at IIT Bombay p 53
35. Mirjalili S (2018) Evolutionary Algorithms and Neural Networks. Springer
36. Ngatchou P, Zarei A, El-Sharkawi A (2005) Pareto multi objective optimization. In: Proceedings of the 13th international conference on, intelligent systems application to power systems, IEEE, pp 84–91
37. Patil DR, Pattewar TM (2022) Majority voting and feature selection based network intrusion detection system. EAI Endors Trans Scalable Info Syst 9(6):e6–e6
38. Rajendran K, Jayabalan M, Rana ME (2017) A study on k-anonymity, l-diversity, and t-closeness techniques. IJCSNS 17(12):172
39. Samarati P (2001) Protecting respondents identities in microdata release. IEEE Trans Knowl Data Eng 13(6):1010–1027
40. Singh R, Subramani S, Du J et al (2023) Antisocial behavior identification from twitter feeds using traditional machine learning algorithms and deep learning. EAI Endors Trans Scalable Inf Syst 10(4):e17–e17
41. Soria-Comas J, Domingo-Ferrer J, Sánchez D et al (2014) Enhancing data utility in differential privacy via microaggregation-based k-anonymity. VLDB J 23(5):771–794
42. Sowmiyaa P, Tamilarasu P, Kavitha S et al (2015) Privacy preservation for microdata by using k-anonymity algorthim. Int J Adv Res Comput Commun Eng 4(4):373–5
43. Sra P, Chand S (2024) A reinduction-based approach for efficient high utility itemset mining from incremental datasets. Data Sci Eng 9(1):73–87
44. Sun X, Li M, Wang H (2011) A family of enhanced (l, α)-diversity models for privacy preserving data publishing. Future Gener Comp Syst 27(3):348–356. https://doi.org/10.1016/j.future.2010.07.007
45. Sun X, Sun L, Wang H (2011) Extended k-anonymity models against sensitive attribute disclosure. Comput Commun 34(4):526–535
46. Sweeney L (2002) k-anonymity: a model for protecting privacy. Int J Uncertain, Fuzziness Knowl-based Syst 10(05):557–570
47. Tabassum M, Mathew K et al (2014) A genetic algorithm analysis towards optimization solutions. Inte J Dig Inf Wirel Commun (IJDIWC) 4(1):124–142
48. Vasa J, Thakkar A (2022) Deep learning: differential privacy preservation in the era of big data. J Comput Inf Syst 63:608–631
49. Venkateswaran N, Prabaharan SP (2022) An efficient neuro deep learning intrusion detection system for mobile adhoc networks. EAI Endors Trans Scalable Inf Syst 9(6):e7–e7
50. Vimalachandran P, Liu H, Lin Y et al (2020) Improving accessibility of the Australian my health records while preserving privacy and security of the system. Health Inf Sci Syst 8:1–9
51. Wan X, Han X (2024) Efficient top-k frequent itemset mining on massive data. Data Sci Eng 9:177–203
52. Wang C, Sun B, Du KJ et al (2023) A novel evolutionary algorithm with column and sub-block local search for sudoku puzzles. IEEE Trans Games 16(1):162–172
53. Wang H, Zhang Y, Cao J (2006) Ubiquitous computing environments and its usage access control. In: Proceedings of the 1st international conference on Scalable information systems, pp 6–es
54. Wang H, Jiang X, Kambourakis G (2015) Special issue on security, privacy and trust in network-based big data. Inf Sci 318:48–50
55. Wang H, Yi X, Bertino E et al (2016) Protecting outsourced data in cloud computing through access management. Concurr Comput: Pract Exp 28(3):600–615
56. Wang S, Wang H, Wei Z et al (2024) A pareto dominance relation based on reference vectors for evolutionary many-objective optimization. Appl Soft Comput 157:111505
57. Yin J, Tang M, Cao J et al (2022) Vulnerability exploitation time prediction: an integrated framework for dynamic imbalanced learning. World Wide Web 25:401–423
58. You M, Yin J, Wang H et al (2023) A knowledge graph empowered online learning framework for access control decision-making. World Wide Web 26(2):827–848
59. Zhang YH, Gong YJ, Gao Y et al (2019) Parameter-free Voronoi neighborhood for evolutionary multimodal optimization. IEEE Trans Evolut Comput 24(2):335–349