


Article

Adapting Geo-Indistinguishability for Privacy-Preserving Collection of Medical Microdata

Seungmin Song and Jongwook Kim * 

Department of Computer Science, Sangmyung University, Seoul 03016, Republic of Korea; ssmredssm@naver.com

* Correspondence: jkim@smu.ac.kr

Abstract: In the era of the Fourth Industrial Revolution, the increasing demand for data collection and sharing for analysis purposes has raised concerns regarding privacy violations. Protecting individual privacy during the collection and dissemination of sensitive information has emerged as a critical concern. In this paper, we propose a privacy-preserving framework for collecting users' medical microdata, utilizing geo-indistinguishability (Geo-I), a concept based on well-known differential privacy. We adapt Geo-I, originally designed for protecting location information privacy, to collect medical microdata while minimizing the reduction in data utility. To mitigate the reduction in data utility caused by the perturbation mechanism of Geo-I, we propose a novel data perturbation technique that utilizes the prior distribution information of the data being collected. The proposed framework enables the collection of perturbed microdata with a distribution similar to that of the original dataset, even in scenarios that demand high levels of privacy protection, typically requiring significant perturbations to the original data. We evaluate the performance of our proposed algorithms using real-world data and demonstrate that our approach significantly outperforms existing methods, ensuring user privacy while preserving data utility in medical data collection.

Keywords: medical microdata privacy; data collection; differential privacy; geo-indistinguishability



Citation: Song, S.; Kim, J. Adapting Geo-Indistinguishability for Privacy-Preserving Collection of Medical Microdata. *Electronics* **2023**, *12*, 2793. <https://doi.org/10.3390/electronics12132793>

Academic Editor: Andrei Kelarev

Received: 22 May 2023

Revised: 20 June 2023

Accepted: 21 June 2023

Published: 24 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the advent of the Fourth Industrial Revolution, numerous fields are experiencing significant changes. Among these changes, the most prominent is the massive generation of data in diverse areas through intelligent information technologies. For example, each mobile phone user generates a large amount of data through daily activities such as messaging, making calls, taking photos, and conducting searches. According to projections, by the end of 2028, the global monthly mobile data traffic is anticipated to reach an estimated total of 453 exabytes [1].

Currently, personal data are viewed as valuable assets in the market. This is because by analyzing individual data, companies can gain valuable insights that ultimately enhance their competitiveness. Consequently, there has been a significant increase in the demand for data collection and sharing for analysis purposes. This increase has been fueled by recent advancements in data analytics methods, such as deep learning and machine learning, which necessitate large quantities of training data. However, there are concerns that the indiscriminate collection or distribution of user data may lead to privacy violations. Many data analytics organizations collect extensive personal information and share it with external parties, often without the user's awareness, leading to serious privacy infringements. For instance, Netflix released anonymized movie ratings data from 500,000 subscribers during its Netflix Prize contest. Although the data were anonymized, it was still possible to identify individual identities, leading to significant controversy [2].

A similar issue arises with medical data. In the age of data-driven decision making, collecting and analyzing diverse medical information is crucial for modern healthcare and medical research. By analyzing extensive medical datasets, researchers can identify

patterns, trends, and associations that contribute to advancements in disease prevention, diagnosis, treatment, and management [3]. Furthermore, the application of advanced analytical methods, such as machine learning and deep learning, has accelerated the potential for groundbreaking discoveries in medical research. These techniques can help uncover previously unnoticed connections or patterns within vast amounts of data [4].

However, concerns have arisen regarding potential privacy violations due to the indiscriminate collection or dissemination of users' medical data. Especially since medical data contain highly sensitive information, indiscriminately collecting and disseminating such data can lead to serious societal issues and may even develop into legal issues. For example, in 2019, Google collaborated with Ascension, the second-largest healthcare system in the United States, on the Nightingale project, which collected patient data without securing users' consent [5]. The gathered data included sensitive details, such as diagnoses and hospitalization records, sparking considerable controversy.

Over the past few decades, considerable efforts have been made to protect the privacy of individuals when collecting and disseminating sensitive data. These efforts have primarily focused on two areas. The first is the establishment of regulations for collecting and sharing personal data. For instance, the European Union's 2018 General Data Protection Regulation (GDPR) [6] mandates companies to acquire explicit consent before collecting, processing, or storing user data. The second area involves developing various methods for protecting individuals' privacy during data collection and sharing. Common approaches include anonymization techniques [7,8] and cryptographic mechanisms [9,10]. Recently, differential privacy (DP) [11] has emerged as the de facto standard for privacy-preserving computations. DP is a mathematical framework that introduces a controlled amount of noise to the data, making it probabilistically difficult to identify individual users. These privacy-preserving methods not only protect user privacy during data collection and distribution but also ensure a certain degree of data utility. Consequently, data analysts can conduct a range of analyses using privacy-protected data.

Despite these efforts, privacy-preserving data collection still faces significant challenges, particularly in the context of collecting medical microdata, where ensuring user privacy often compromises data utility. Thus, in this paper, we aim to develop a method for collecting users' medical microdata in a privacy-preserving manner. We utilize geo-indistinguishability (Geo-I) [12–15], a concept based on the well-established DP and recently recognized as the standard privacy definition for protecting location data in LBS (location-based service), to ensure user privacy during the medical data collection process.

1.1. Motivation

Consider the motivational example depicted in Figure 1, where a data analytics organization aims to collect patients' medical microdata, such as each user's disease, for analytical purposes. However, as medical information is highly sensitive, users are hesitant to share their disease information with the data analytics organization.

A possible solution is to utilize DP, which is widely considered as the standard for privacy-preserving data collection. As shown in Figure 1, each user first perturbs their medical information using the perturbation mechanism of DP and then provides the perturbed data to the data analytics organization. This process alleviates users' concerns about privacy breaches, as data perturbation is performed on the user's end, guaranteeing that the original medical data remain undisclosed to third parties. However, this solution also reduces the data utility of the collected information due to the user-side data perturbation, leading to decreased accuracy in the analysis results obtained from the perturbed datasets. Consequently, there is a need for a mechanism that can protect user privacy during the collection of medical microdata while minimizing the reduction in data utility.

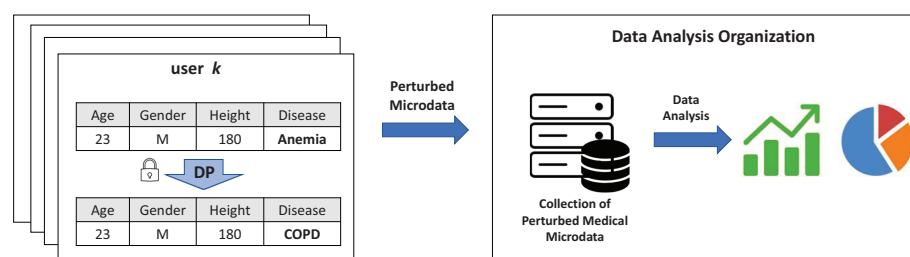


Figure 1. Motivational example in which disease information corresponds to sensitive information.

1.2. Contributions

The main contributions of this paper are summarized as follows:

- We develop a privacy-preserving framework to collect medical microdata from each user while ensuring privacy. Specifically, we adapted Geo-I, which was originally designed for protecting location information privacy to collect medical microdata, which are a form of text data. In particular, this is the first attempt to utilize Geo-I for the privacy-preserving collection of medical microdata.
- To address the reduced data utility of collected datasets caused by Geo-I's perturbation mechanism, we introduce a new data perturbation method for Geo-I that utilizes prior distribution information of the data to be collected. The primary advantage of leveraging a prior distribution of the collected data during the process of perturbing the original microdata is that it enables the collection of perturbed microdata of which distribution is similar with that of the original dataset.
- Furthermore, we evaluate the performance of our proposed algorithms using real-world data. The results demonstrate that our approach significantly outperforms existing methods. Especially, the experiment results confirm that our proposed method can maintain the data utility of the collected datasets, even in scenarios demanding high levels of privacy protection, which typically necessitate considerable perturbations to the original data.

The rest of this paper is structured as follows: Section 2 discusses related work, and Section 3 provides background information. Section 4 describes the proposed method. The performance of the proposed technique is evaluated in Section 5, and conclusions are drawn in Section 6.

2. Related Work

2.1. Privacy-Preserving Text Analysis and Collection

Text data are frequently used as input for various data analysis tasks. However, inappropriate utilization of these types of data could result in significant privacy concerns, including the prediction of patients' health conditions using their clinical records [16]. Consequently, it is essential to carefully manage sensitive information. Privacy preservation in text analysis has been widely researched and explored in the literature [17]. One common approach involves identifying sensitive terms, such as personally identifiable information, within a document and replacing them with more generic terms [18–20].

Recently, various methods have been investigated to protect user text data by perturbing them during the data collection phase to ensure privacy preservation. Pretrained contextualized language models have been shown to improve the efficiency of numerous natural language processing tasks. However, the effectiveness of existing text sanitization methods remains limited due to the complexity of high-dimensional text representation. To address this issue, Yue et al. [21] developed a privacy-preserving natural language processing pipeline that tackles privacy concerns by generating sanitized text documents directly. They sanitize public data before training the model, as they enable the model to work with sanitized queries more effectively, thus enhancing accuracy. Additionally, recent studies have explored novel techniques for safeguarding text data privacy by manipulating the data during the collection process [22–24].

Feyisetan et al. [25] recently proposed a privacy-preserving mechanism for publishing sensitive text data using Geo-I. In their approach, for a given word in the text data, they first compute their vector representation in the embedding space, denoted as x . They then apply a calibrated noise parameter, N , which is designed to be sensitive to the global metric. The perturbed vector v is obtained by adding the noise parameter to x , resulting in $v = x + N$. Finally, the original word is replaced with another word whose embedding is closest to the perturbed vector v . Our proposed method has similarities with the approach in [25] in its use of Geo-I. However, in contrast to [25], our method in this paper takes advantage of prior distribution information of the data to be collected, aiming to improve the data utility of the collected datasets.

2.2. Geo-Indistinguishability and Its Applications

Geo-I has gained significant attention in various LBS applications due to its ability to protect location privacy. In mobile crowdsourcing, where workers must share their locations with mobile crowdsourcing servers to allocate sensing tasks to the nearest workers, Geo-I is employed to obfuscate workers' true locations, thereby protecting their location privacy [26–29]. Location-based social networking platforms assist people in connecting with each other, but they also pose a threat to location privacy. Therefore, various studies have suggested using Geo-I to protect the privacy of users in such platforms [30,31]. To mitigate the risk of exposing sensitive location data in ride-sharing services such as Uber, Waze, and Lyft, several studies have proposed scheduling schemes that utilize Geo-I to protect the location information of ride-sharing users who are required to share their current and destination locations with the service providers [32,33]. Geo-I has been utilized in estimating density distribution. One example of this is EGeoIndis [34], which is a vehicle-location privacy protection framework that utilizes Geo-I to estimate traffic density and to protect vehicle location privacy. Chen et al. [35] relied on Geo-I to collect the locations of voluntary participants with COVID-19 symptoms in a privacy-preserving manner. The location data collected under Geo-I were used to construct a COVID-19 vulnerability map.

3. Background

3.1. Differential Privacy

DP is based on the assumption that there is a trusted aggregator or curator who serves as a central intermediary between data contributors (i.e., data owners) and data users [11]. DP is commonly used in two different contexts: non-interactive and interactive settings. In the non-interactive setting, the trusted aggregator collects raw data from individual data owners, computes aggregate statistics based on the collected data, introduces random noise to the true aggregate statistics to produce perturbed aggregate statistics, and then publishes these perturbed statistics to data users [36,37]. In the interactive setting, when using DP, the trusted curator receives a query from a data user, computes the true result of the query using the original database, adds random noise to the true result to generate a perturbed result, and then returns this perturbed result to the data user [38,39]. DP can be formally defined as follows:

Definition 1. (ϵ -DP) A randomized algorithm \mathcal{A} satisfies ϵ -DP, if and only if for (1) any two neighboring datasets, D_1 and D_2 , and (2) any output O of \mathcal{A} , the following is satisfied:

$$\Pr[\mathcal{A}(D_1) = O] \leq e^\epsilon \times \Pr[\mathcal{A}(D_2) = O].$$

Two datasets, D_1 and D_2 , are considered neighboring if they differ by only one record. Here, \Pr represents the probability, with its probability space defined by the random outputs generated by the mechanism \mathcal{A} . This definition indicates that, given any output of \mathcal{A} , an adversary with arbitrary background knowledge cannot confidently determine whether the input of \mathcal{A} is D_1 or D_2 . Here, the parameter ϵ acts as a privacy budget, controlling the level of privacy. In other words, smaller values of ϵ provide stronger privacy

protection with more noise, while larger values of ϵ result in weaker privacy guarantees with less noise in the true result.

3.2. Geo-Indistinguishability

As DP has emerged as a widely recognized standard for privacy-preserving computation, many approaches try to apply the concept of DP for protecting location data. Among these, Geo-I has recently emerged as the de facto privacy definition for protecting location data. Geo-I can provide strong location privacy protection against adversaries with arbitrary background knowledge, leading to its extensive adoption in various location-based applications [12–15]. Geo-I can be formally defined as follows:

Definition 2. (ϵ -Geo-I) Assume that \mathcal{X} represents a set of possible user locations. Let K be a randomized mechanism that probabilistically generates a perturbed location from a user's true location. Then, a randomized mechanism, K , satisfies ϵ -Geo-I, if and only if for (1) all $x, x' \in \mathcal{X}$ and (2) any output location, $y \in \mathcal{X}$, the following is satisfied:

$$K(x)(y) \leq e^{\epsilon \cdot d(x, x')} \times K(x')(y). \quad (1)$$

Here, $d(x, x')$ is the distance metric, such as Euclidean or Manhattan distance between x and x' .

There are two primary methods for implementing Geo-I. The first approach is the Laplace mechanism, which involves adding Laplace-distributed noise to the user's actual data. While this method is simple, it may introduce a large amount of noise during the perturbation process, which leads to reduced data utility. Alternatively, the optimization mechanism is a more effective technique, as it results in less perturbation to the user's actual location data compared to the Laplace method, leading to higher data utility.

In the optimization mechanism, the LBS server first calculates the obfuscation matrix, M , by solving the following linear programming problem.

$$\begin{aligned} \min : & \sum_{x, y \in \mathcal{X}} \pi_x \cdot M[x, y] \cdot d(x, y) \\ \text{s.t.} : & M[x, y] \leq e^{\epsilon \cdot d(x, x')} \times M[x', y] \quad x, x', y \in \mathcal{X} \\ & \sum_{y \in \mathcal{X}} M[x, y] = 1 \quad x \in \mathcal{X} \\ & M[x, y] > 0 \quad x, y \in \mathcal{X} \end{aligned} \quad (2)$$

Here, π represents the prior probability distribution of user locations, which can be obtained using the existing historical data. $M[x, y]$ represents the probability of a true location x randomly generating a perturbed location y (i.e., $M[x, y] = K(x)(y)$). Once M is computed, it is disseminated to LBS users. Upon receipt of M , each user perturbs their true location based on the probabilities contained in M and sends the perturbed location along with a service request to the LBS server. Throughout this process, the true location of each user remains undisclosed, as the perturbation of location data is performed within his/her mobile device.

4. Privacy-Preserving Framework for Collecting Medical Microdata

In this section, we introduce our privacy-preserving framework to collect users' medical microdata. The proposed framework, illustrated in Figure 2, consists of two parties: a data-collection server and individual users:

- **Data-collection server:** The data-collection server generates an obfuscation matrix designed for perturbing actual medical microdata under the ϵ -Geo-I and distributes them to each user. During the generation of this obfuscation matrix, the server utilizes prior distribution information derived from the available historical data.

- Individual user: When users receive the obfuscation matrix, they first perturb their own medical microdata based on the probabilities embedded in the obfuscation matrix and then send the perturbed data to the data-collection server.

During this entire process, each user’s sensitive medical microdata remain confidential, as the process of perturbing the data is conducted within their own device. In the following subsections, we explain each step in detail.

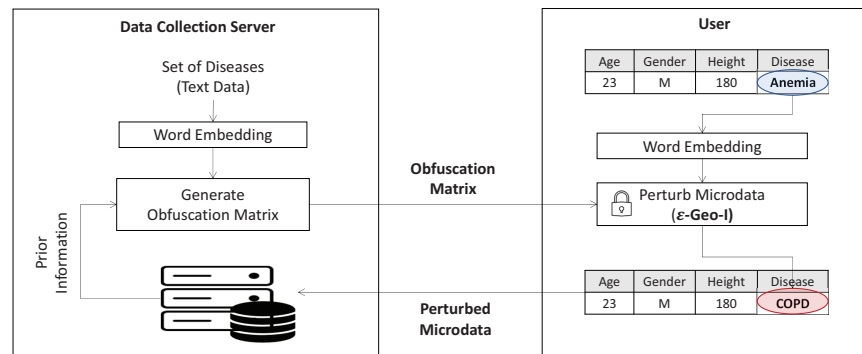


Figure 2. Overview of the proposed privacy-preserving framework for collecting medical microdata.

4.1. Preliminary

In this subsection, we introduce the notation necessary to explain our proposed method. Let $U = \{u_1, u_2, \dots, u_n\}$ be a set of users who agree to share their sensitive medical microdata with a data aggregator for analytical purposes. Here, $u_i \in U$ represents the i -th user. However, due to a lack of complete trust with the data aggregator, users instead provide perturbed data obtained using ϵ -Geo-I.

Let $S = \{s_1, s_2, \dots, s_m\}$ denote the set of distinct medical microdata. Furthermore, let $s_{u_i} \in S$ represent the medical microdata of the i -th user, $u_i \in U$. For the sake of simplicity, we consider a scenario in which the data aggregator collects a single microdatum from each user. However, we note that the approach proposed in this paper is equally applied to the situation where several microdata are collected from individual users.

4.2. Data-Collection Server

The data-collection server side processing consists of two phases. The first phase utilizes a word embedding to represent each microdatum as a vector in a high-dimensional space. In the second phase, an obfuscation matrix, which will be distributed to each user in order to perturb their original microdata, is generated. This subsection provides a detailed explanation of these two phases.

4.2.1. Vector Space Representation of Medical Microdata

Originally, Geo-I was designed to protect users’ location information in a two-dimensional space for LBSs. Consequently, this is not directly applicable to medical microdata, which are a type of text data. To address this, we first need to represent each microdatum as a vector in a high-dimensional space. Word embedding, such as Word2Vec [40], BERT [41], and GloVe [42], is a natural language processing technique that aims to represent words, phrases, or other text-based data as continuous vectors in a high-dimensional space. The primary objective of word embeddings is to capture the semantic meaning of words, making them easily understandable and processable for machine learning algorithms. In this paper, we employ Word2Vec to represent medical microdata as a vector in a high-dimensional space. Formally, given $s_i \in S$, let v_i be the corresponding t -dimensional vector representation obtained using Word2Vec.

4.2.2. Computation of Obfuscation Matrix

There are two approaches to achieve ϵ -Geo-I: the Laplace mechanism [12] and the optimization mechanism [13,14]. Although the Laplace mechanism is straightforward, it is well known for introducing significant noise into the original data, which leads to perturbed locations with reduced utility. In contrast, the optimization mechanism can produce perturbed locations with higher utility compared to the Laplace mechanism, as it takes advantage of a prior data distribution. However, the optimization mechanism is often regarded as inefficient due to the need to solve an expensive linear program. This issue is further magnified in our case, as unlike location data represented in a two-dimensional space, vectors derived from word embedding techniques usually are represented in a significantly high-dimensional space. As a result, directly applying the optimization mechanism to our problem is impractical.

One potential solution is to utilize dimensionality reduction techniques, such as principal component analysis [43], to transform the vector generated with word embedding techniques into a two-dimensional representation, followed by the application of the optimization technique. However, as will be shown in the experiment section, this method results in perturbed microdata with lower utility due to the accuracy loss caused by the reduction in dimensionality from a high-dimensional to a two-dimensional space.

In this paper, we introduce a novel perturbation mechanism to achieve ϵ -Geo-I that can be efficiently applied to vectors represented in a high-dimensional space. The proposed method is inspired by the perturbation mechanism in [29]. However, unlike the mechanism in [29], which does not use prior data distribution, our approach utilizes a prior data distribution to generate perturbed data with higher utility, similar to the optimization mechanism. Given a set of all medical microdata, $S = \{s_1, s_2, \dots, s_m\}$, that the data aggregator aims to collect from users, let $V = \{v_1, v_2, \dots, v_m\}$ represent the set of corresponding vectors obtained using Word2Vec in the previous phase. The obfuscation matrix is then defined as an $m \times m$ matrix, O , where $O[i, j]$ represents the probability that a perturbed microdatum s_j is randomly generated from the true microdatum s_i . Here, $O[i, j]$ is defined as follows:

$$O[i, j] = Pr(s_j | s_i) = \frac{\psi(p_{s_j}) \cdot e^{-\frac{\epsilon}{2} \cdot d(v_i, v_j)}}{\sum_{s_k \in S} \psi(p_{s_k}) \cdot e^{-\frac{\epsilon}{2} \cdot d(v_i, v_k)}} \quad (3)$$

Here, $d(v_i, v_j)$ represents the distance between two vectors, v_i and v_j . Moreover, p_{s_j} denotes the prior distribution information regarding the likelihood that microdatum s_j appears in the entire dataset, and ψ represents an arbitrary monotone increasing function that accepts s_j as input. According to Equation (3), by using the monotone increasing function that takes s_j as input, if microdatum s_i occurs more frequently than microdatum s_j in the original dataset, then s_i will also appear more frequently than s_j in the perturbed dataset. Once computing the obfuscation matrix O , the server disseminates it to all users.

We note that our approach shares similarities with the optimization mechanism in the way it relies on the obfuscation matrix to perturb sensitive data of users. Moreover, a prior data distribution is utilized when generating this obfuscation matrix. However, it is widely known that, as explained in Section 3.2, the optimization mechanism can be highly inefficient, as it requires solving a costly linear programming problem to generate the obfuscation matrix [44]. This issue is further magnified in our case, as unlike location data represented in a two-dimensional space, medical microdata are represented in a significantly high-dimensional space due to the word-embedding techniques. Yet, our proposed approach is highly efficient, as it eliminates the need for solving an expensive linear programming problem, making it suitable for collecting data that need to be represented in a significantly high-dimensional space.

4.2.3. Estimation of Prior Distribution

The perturbation mechanism proposed in Section 4.2.2 utilizes prior information about the probability of each microdatum appearing in the entire dataset. This prior

information can be derived from a collection of historical data. However, in our case, we collect perturbed microdata from users instead of the true data, which makes it challenging to accurately compute the prior information from the perturbed microdata collection. Therefore, in this subsection, we present a method for effectively estimating the prior distribution using the collection of perturbed microdata.

Let DB represent the collection of historical perturbed microdata collected and maintained by the data-collection server. Additionally, for a given medical microdatum $s_i \in S$, let us assume that $cnt(s_i, DB)$ represents the number of occurrences of s_i in DB . Then, p_{s_i} , denoting the probability of microdatum s_i occurring in the dataset to be collected, can be estimated as $\frac{cnt(s_i, DB)}{|DB|}$. However, this straightforward approach cannot accurately estimate p_{s_i} because it does not consider the effect of the Geo-I perturbation mechanism.

A more effective solution is to leverage the probabilistic mapping information between the true and perturbed microdata encoded in the obfuscation matrix O . According to the definition of the obfuscation matrix O , for all $s_j \in S$, $O[s_i, s_j]$ represents the probability that a perturbed microdatum s_j (corresponding to the microdata received by the server from a user) is randomly generated from the user's actual microdatum s_i . Hence, by utilizing the mapping probability information between perturbed and true microdata, p_{s_i} are estimated as

$$p_{s_i} = \frac{\sum_{s_j \in S} (O[s_i, s_j] \times cnt(s_j, DB))}{|DB|} \tag{4}$$

In other words, when computing p_{s_i} , this approach takes into account the probabilities encoded in the perturbation matrix O that the perturbed microdata $s_j \in S$ are randomly generated from the true location s_i .

4.3. User-Side Processing

After receiving the obfuscation matrix from the server, each user perturbs their actual microdata according to the probabilities contained within the matrix. To be more precise, let us assume that the true microdata of user u_i are $s_k \in S$. User u_i randomly generates the perturbed microdata in S , based on the probabilities in the k -th row of the obfuscation matrix (i.e., $O[k, j]$ where $1 \leq j \leq m$). Note that, according to Equation (3), the sum of each row in the obfuscation matrix equals 1 (i.e., $\sum_{1 \leq j \leq m} O[k, j] = 1$). Users then send the perturbed microdata to the data-collection server. We note that the data perturbation process occurs on the user side, guaranteeing that users' true microdata are not exposed to external parties, thereby protecting the privacy of the user's medical microdata.

4.4. Privacy Analysis

In this subsection, we perform a privacy analysis of the proposed method.

Theorem 1. *Given the privacy budget ϵ , the proposed method satisfies ϵ -Geo-I.*

Proof. By the definition of ϵ -Geo-I, given $s_i, s_j, s_x \in S$ and their corresponding vector representations v_i, v_j and v_x , we need to prove the following:

$$Pr(s_j|s_i) \leq e^{\epsilon \cdot d(s_i, s_x)} \times Pr(s_j|s_x) \iff Pr(v_j|v_i) \leq e^{\epsilon \cdot d(v_i, v_x)} \times Pr(v_j|v_x) \tag{5}$$

Here, $Pr(v_j|v_i)$ represents the probability that a user sends the perturbed microdatum s_j to the data-collection server when their actual microdatum is s_i .

By Equation (3), $Pr(v_j|v_i)$ is computed as

$$Pr(v_j|v_i) = \frac{\psi(p_{s_j}) \cdot e^{-\frac{\epsilon}{2} \cdot d(v_i, v_j)}}{\sum_{s_k \in S} \psi(p_{s_k}) \cdot e^{-\frac{\epsilon}{2} \cdot d(v_i, v_k)}} \tag{6}$$

Using the triangular inequality, we obtain $d(v_i, v_j) + d(v_i, v_x) \geq d(v_x, v_j)$ and thus $-d(v_i, v_j) \leq d(v_i, v_x) - d(v_x, v_j)$. From this, we derive:

$$Pr(v_j|v_i) = \frac{\psi(p_{s_j}) \cdot e^{-\frac{\epsilon}{2} \cdot d(v_i, v_j)}}{\sum_{s_k \in S} \psi(p_{s_k}) \cdot e^{-\frac{\epsilon}{2} \cdot d(v_i, v_k)}} \leq \frac{\psi(p_{s_j}) \cdot e^{\frac{\epsilon}{2} \cdot (d(v_i, v_x) - d(v_x, v_j))}}{\sum_{s_k \in S} \psi(p_{s_k}) \cdot e^{-\frac{\epsilon}{2} \cdot d(v_i, v_k)}} \tag{7}$$

Similarly, using the triangular inequality, we have $d(v_i, v_k) \leq d(v_i, v_x) + d(v_x, v_k)$ and thus $-d(v_i, v_k) \geq -d(v_i, v_x) - d(v_x, v_k)$. From this, we obtain:

$$Pr(v_j|v_i) \leq \frac{\psi(p_{s_j}) \cdot e^{\frac{\epsilon}{2} \cdot (d(v_i, v_x) - d(v_x, v_j))}}{\sum_{s_k \in S} \psi(p_{s_k}) \cdot e^{-\frac{\epsilon}{2} \cdot d(v_i, v_k)}} \leq \frac{\psi(p_{s_j}) \cdot e^{\frac{\epsilon}{2} \cdot (d(v_i, v_x) - d(v_x, v_j))}}{\sum_{s_k \in S} \psi(p_{s_k}) \cdot e^{\frac{\epsilon}{2} \cdot (-d(v_i, v_x) - d(v_x, v_k))}} \tag{8}$$

Therefore, we have:

$$Pr(v_j|v_i) \leq e^{\epsilon \cdot d(v_i, v_x)} \cdot \frac{\psi(p_{s_j}) \cdot e^{-\frac{\epsilon}{2} \cdot d(v_x, v_j)}}{\sum_{s_k \in S} \psi(p_{s_k}) \cdot e^{-\frac{\epsilon}{2} \cdot d(v_x, v_k)}} = e^{\epsilon \cdot d(v_i, v_x)} \cdot Pr(v_j|v_x) \tag{9}$$

According to Equation (9), the proposed method satisfies ϵ -Geo-I. \square

4.5. Limitations

In this subsection, we discuss the limitations of the proposed approach, particularly in comparison with Laplace mechanism-based methods such as [25]. The method proposed in this paper necessitates the generation of an obfuscation matrix of size $m \times m$, where m represents the number of elements in the medical microdata set, S . Consequently, a large m could impose substantial overhead in terms of matrix generation and distribution to each individual user. Additionally, it is essential to identify and define all elements in S prior to generating the obfuscation matrix. These constraints imply that our proposed method is better suited to collect microdata specific to a particular domain rather than collecting general text data, which might include every word in a dictionary and thereby lead to an enormous word count. On the other hand, the method based on the Laplace mechanism does not rely on an obfuscation matrix, thereby enabling it to collect a wide range of data, even when the domain size is extensive.

5. Experiment

In this section, we present the experimental evaluation of the proposed method using real-world datasets. First, we describe the experimental setup, and then, we discuss the results obtained from the experiments.

5.1. Experimental Setup

We report the results for the following alternative methods:

- The Laplace mechanism-based approach, which corresponds to the method proposed in [25] adapted to our problem (*LM*);
- The approach discussed in Section 4.2.2 that utilizes dimensionality reduction techniques to convert the vector generated with word-embedding techniques into a two-dimensional representation, followed by the application of the optimization mechanism (*OM*);
- The approach that utilizes the perturbation mechanism proposed in [29], which does not use a prior distribution (*NP*);
- The proposed method that leverages prior distribution information of the data being collected (*PM*).

In the case of *PM*, we employ a linear function with its slope set to 1 as the monotone increasing function $\psi(\cdot)$ in Equation (3). For each approach, we use Word2Vec to represent microdata as a vector in a 300-dimensional space for the word-embedding process. In

particular, we leverage word vectors from a pretrained Word2Vec model that was trained on Wikipedia data using fastText. These 300-dimensional vectors were created using the skip-gram model, as described in [45].

The following two datasets are used for our evaluation:

- **MIMIC-III:** The first dataset is the MIMIC-III database [46]. This open-source database contains anonymized health data from more than 46,000 patients who were admitted to intensive care units (ICUs) in the United States from 2001 to 2012. We specifically use the admission data from this database, which consist of 58,976 records.
- **Wikipedia Disease:** For the second dataset, we first collect disease data from the “Lists of diseases” page on Wikipedia. These data are arranged in a tree structure with a maximum depth of 4. The names of the diseases utilized in the experiments are located at the leaf nodes of this tree, which altogether account for 61 nodes. Then, we randomly generate datum for 61,000 patients, each associated with one disease from the list of 61 diseases.

In the experiment, we compare the four alternative approaches using the average distance between the true and perturbed microdata:

$$Dist = \frac{1}{n} \times \sum_{i=1}^n d(s_{u_i}, s'_{u_i}) \tag{10}$$

Here, $s_{u_i} \in S$ denotes the true microdatum of the i -th user, u_i , while $s'_{u_i} \in S$ represents the perturbed microdatum of the same user. As defined in Section 4.1, n is the number of users. Additionally, $d(s_{u_i}, s'_{u_i})$ denotes the distance between s_{u_i} and s'_{u_i} . In the experiment, we employ two different methods to measure the distance between s_{u_i} and s'_{u_i} . In the first method, we first represent the microdatum in vector space using Word2Vec and then measure the Euclidean distance between the vectors, denoted as $Dist_{ed}$. In the second method, we calculate the tree distance between two microdatum, s_{u_i} and s'_{u_i} , using the disease tree obtained from the “Lists of diseases” page on Wikipedia, represented as $Dist_{tree}$.

5.2. Results and Discussion

Figure 3 illustrates the impact of varying privacy budgets, ϵ , on the average distance between true and perturbed microdata. In the experiments, ϵ varies from 0.5 to 2.0, and the *Wikipedia Disease* dataset is used. Across all methods, the average distance reduces as the privacy budget increases from 0.5 to 2.0. This occurs because a lower ϵ value in Geo-I provides more robust privacy protection by introducing larger perturbations to the true microdata, consequently reducing the utility of the collected data. In contrast, a higher ϵ value results in smaller perturbations to the actual microdata, offering weaker privacy protection but maintaining a higher data utility.

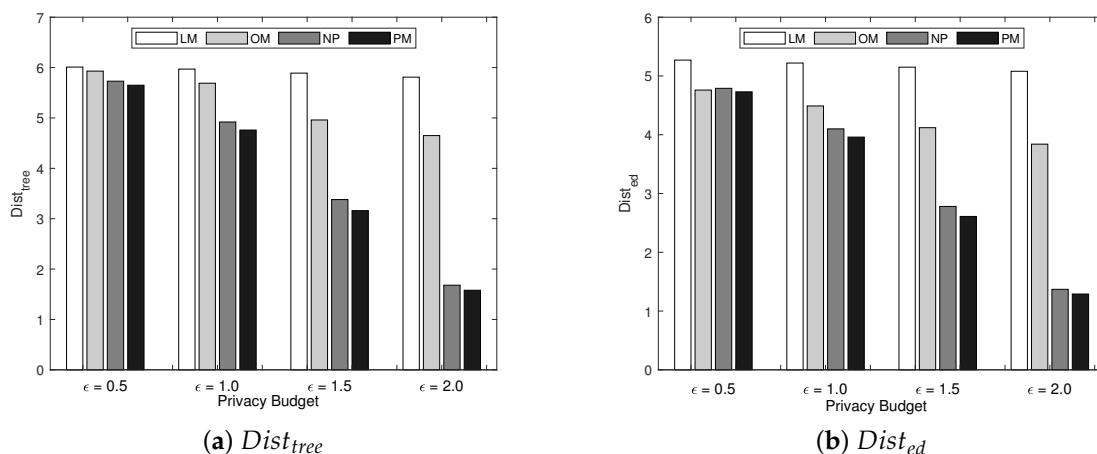


Figure 3. The average distance between the true and perturbed microdatum for varying ϵ .

Figure 3 indicates that the Laplace mechanism-based method, *LM*, shows the lowest performance in terms of data utility when compared to other techniques. This is attributed to the fact that the Laplace mechanism-based method is known for introducing significant noise during the perturbation phase, consequently leading to a decrease in data utility. Among the three alternatives, *OM*, *NP*, and *PM*, that utilize the server-generated obfuscation matrix to perturb user microdata, *NP* and *PM* significantly outperform *OM*. This is because *OM* generates perturbed microdata with decreased utility, as it experiences accuracy loss when dimensions are reduced from a high-dimensional space to a two-dimensional one. The figure also demonstrates that the proposed method, *PM*, which utilizes prior distribution information of the data collected, surpasses all other methods across all privacy levels. These experimental results verify that the proposed method, *PM*, effectively leverages the prior distribution information of the collected data.

In Figure 4, we make a comparison between the distribution of the true microdata collection and that of the perturbed microdata collection obtained using the three alternative methods *LM*, *NP*, and *PM*. We note that the *OM* results are absent from the figure because it shows the worst performance among the methods utilizing the server-generated obfuscation matrix. In Figure 4, *OG* stands for the results derived from the original dataset. In the figure, the x-axis indicates the index number for each disease, while the y-axis denotes the number of datum associated with each disease. As can be seen in the figure, the proposed method, *PM*, shows a distribution most similar to that of the original dataset. The reason for this is that *PM* utilizes a prior distribution of the collected data during the perturbation of the original microdata, which enables the collection of perturbed microdata whose distribution is highly similar to that of the original dataset.

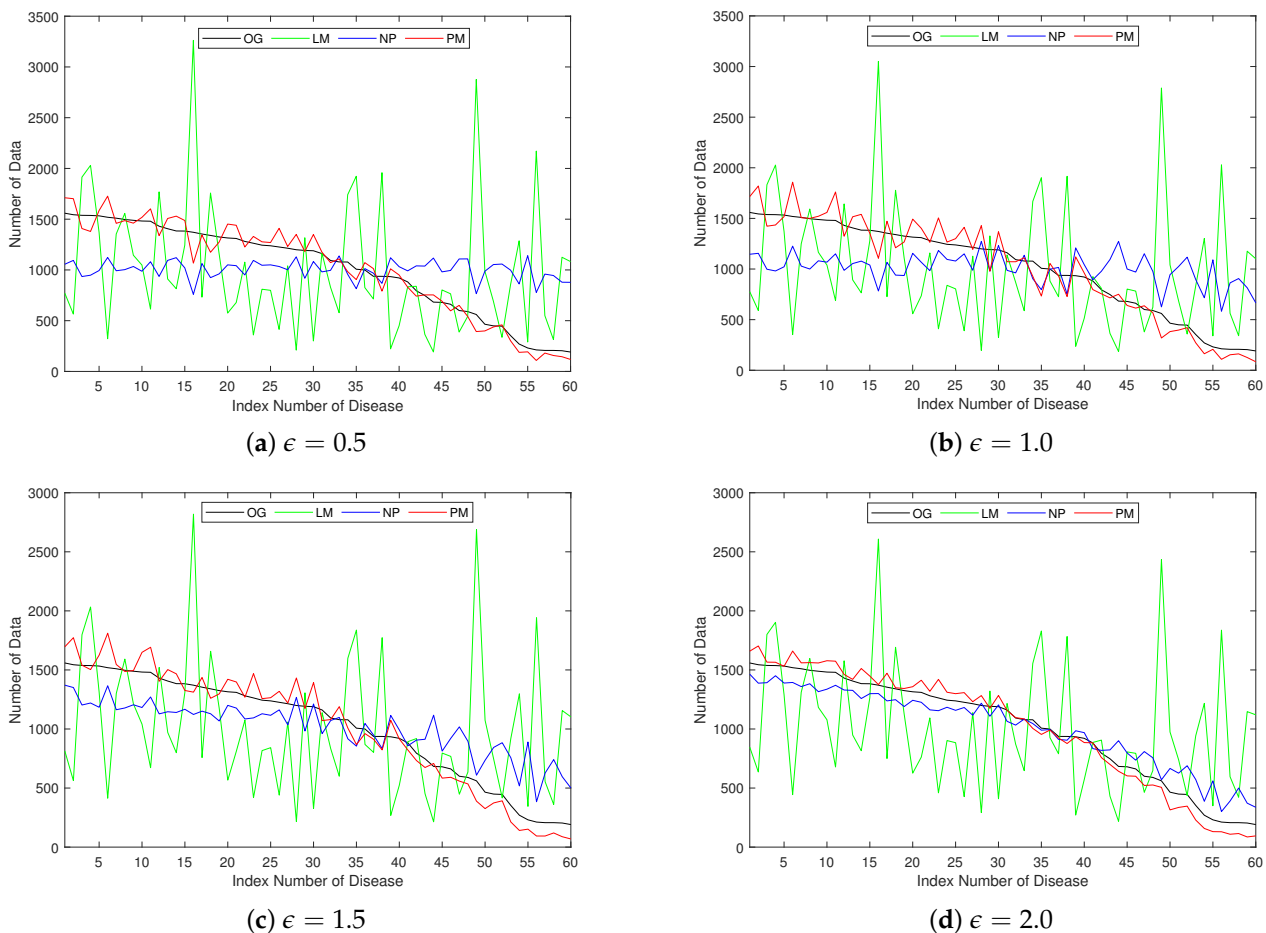


Figure 4. A comparison between the distribution of the true microdata collection and that of the perturbed microdata collection.

To further investigate the performance gap between the proposed method *PM* and the Laplace mechanism-based approach *LM*, we plotted the Euclidean distance between true and perturbed microdata for each individual data point in Figure 5. For visual clarity, the Euclidean distances of a subset of 6000 individual data points, extracted from the total pool of 61,000, are shown in this figure. In the figures, the x-axis denotes the index number for each sampled microdatum, while the y-axis represents the Euclidean distance between the true and perturbed microdata. Moreover, the red dots correspond to the results obtained using the proposed *PM*, while the blue dots represent the results obtained using *LM*. As shown in the figure, for the majority of cases, *PM* achieves better results compared to *LM*, which is in line with earlier results presented in Figure 3.

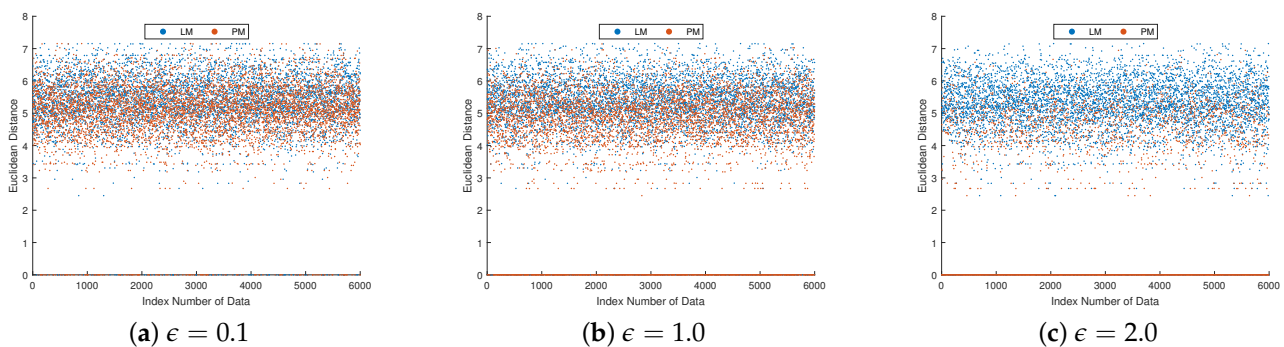


Figure 5. The Euclidean distance between each true and perturbed microdatum.

In order to evaluate the data utility of the collected microdata, we present the results of a data analysis task carried out using the collection of perturbed microdata collected through different methods. In the experiment, the MIMIC-III dataset was used. The analysis query used in this experiment is the following aggregation query:

```
SELECT month, count(*) FROM mimic WHERE comments such as '%newborn%' GROUP BY month
```

This query aims to calculate the number of patients admitted to ICUs on account of being newborns. In Figure 6, we report the results obtained using *LM*, *NP*, and *PM*. We note that the results of *OM* are not included in the figure, as it exhibits the worst performance among the methods utilizing the server-generated obfuscation matrix. Additionally, for comparative purposes, we also plot the results derived from the collection of true microdata, referred to as *OG*. In Figure 6, the x-axis corresponds to the month, while the y-axis represents the count of newborn patients admitted to ICUs. As the privacy budget increases, the results derived from the perturbed datasets become increasingly similar to the actual results obtained using the true datasets. Furthermore, the results computed using the data collected using our proposed technique, *PM*, exhibit a pattern most similar to those obtained from the original data. This validates that our proposed method is capable of enhancing the data utility of the collected datasets, while also protecting user privacy.

When the level of privacy protection decreases (represented by an increasing value of ϵ), the performance gap between *PM* and *NP* similarly reduces. This is because a decrease in privacy protection level induces smaller perturbations to the original microdata. As a result, the impact of employing a prior distribution in the process of perturbing the original microdata becomes less significant compared to scenarios with higher privacy protection levels. On the other hand, as the level of privacy protection increases (represented by a decreasing value of ϵ), the performance gap between *PM* and *NP* also increases. The primary reason is that *PM* employs a prior distribution of the collected data during the process of perturbing the original microdata. As a result, despite the large perturbations to the original microdata induced by a high level of privacy protection, the distribution of the perturbed dataset remains similar to the original dataset's distribution. In contrast, *NP* does not utilize a prior distribution when perturbing the original microdata. Hence, as the

level of privacy protection increases, leading to larger perturbations to the original data, the distribution of the perturbed dataset becomes significantly different from the original data’s distribution. This confirms that our proposed method can improve the data utility of the collected datasets, even in circumstances where a high degree of privacy protection is necessary.

Table 1 presents the results of additional aggregation queries, which are of the same type as the query used in Figure 6. For these experiments, in addition to the query that computes the monthly count of newborn patients admitted to the ICU, we also include four additional aggregation queries. These queries calculate the monthly number of patients admitted to ICUs due to bleed, coronary conditions, pneumonia, and sepsis, respectively. In the experiments, we calculate the absolute error, which is defined as $|cnt - cnt'|$. Here, cnt represents the query result derived from the collection of true microdata, whereas cnt' refers to the query result generated from the collection of perturbed microdata that are collected using the privacy-preserving methods *LM*, *NP*, and *PM*. In the experiments, ϵ is set to 2.0.

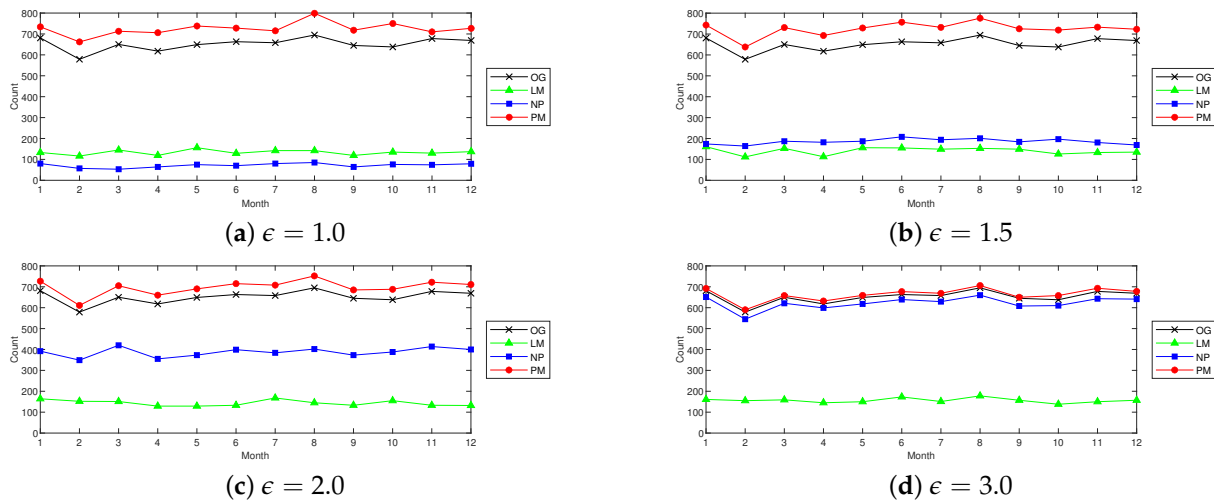


Figure 6. A comparison of aggregation query results computed from the perturbed dataset collected using *LM*, *NP*, and *PM*.

Table 1. A comparison of the absolute error of aggregation queries among *LM*, *NP*, and *PM* ($\epsilon = 2.0$).

Diagnosis	Method	Month												Average
		1	2	3	4	5	6	7	8	9	10	11	12	
newborn	<i>LM</i>	517	427	499	489	520	530	490	550	512	483	545	537	508.25
	<i>NP</i>	289	230	230	263	276	264	274	293	272	250	264	269	264.50
	<i>PM</i>	49	51	50	50	62	35	43	41	53	54	46	43	48.08
bleed	<i>LM</i>	184	187	175	190	188	148	206	194	198	205	179	175	185.75
	<i>NP</i>	116	126	113	128	115	107	133	144	135	152	132	119	126.67
	<i>PM</i>	61	65	81	37	57	81	71	64	72	66	59	58	64.33
coronary	<i>LM</i>	268	236	291	256	290	272	250	277	261	299	263	269	269.33
	<i>NP</i>	198	167	209	196	215	204	189	212	199	246	207	186	202.33
	<i>PM</i>	109	110	100	125	108	117	152	156	120	109	105	130	120.08
pneumonia	<i>LM</i>	206	208	216	181	177	168	156	149	150	155	177	214	179.75
	<i>NP</i>	162	158	151	133	125	111	98	127	104	117	136	158	131.67
	<i>PM</i>	50	45	19	58	42	56	93	53	62	78	33	15	50.33
sepsis	<i>LM</i>	149	115	163	144	143	163	153	159	130	141	147	160	147.25
	<i>NP</i>	104	70	114	89	102	105	93	104	85	91	91	107	96.25
	<i>PM</i>	54	85	72	56	63	67	40	49	86	56	50	43	60.08

As can be seen in Table 1, the proposed method *PM* performs better than other approaches across all aggregation queries. This is due to the fact that the method proposed in this paper enables the collection of perturbed microdata with a distribution that is more similar to that of the original dataset, compared with other approaches *LM* and *NP*. As a result, the aggregation query results obtained from the collection of perturbed microdata using the proposed approach exhibit a higher level of similarity with the true results compared to other approaches *LM* and *NP*.

6. Conclusions

In the era of the Fourth Industrial Revolution, the collection and analysis of large amounts of personal data, especially in sensitive fields such as healthcare, is inevitable. However, it is equally important to ensure that such endeavors respect individual privacy rights and do not lead to any inadvertent data breaches or misuse. This paper aimed to address this crucial challenge by proposing a new framework for collecting medical microdata in a privacy-preserving manner while maintaining data utility. We adapted the concept of Geo-I, a privacy-preserving method originally designed for LBSs, to the context of medical microdata. We introduced a novel data perturbation method for Geo-I that leverages prior distribution information of the data to be collected, in order to address the issue of reduced data utility that often arises from the use of privacy-preserving methods. Through comprehensive experiments conducted with real-world datasets, we demonstrated that our proposed method significantly outperforms existing ones in terms of maintaining data utility while ensuring privacy.

The findings of this paper have significant implications, especially in sensitive areas such as healthcare. It underscores the importance of privacy-preserving techniques that maintain data utility, as the healthcare sector increasingly depends on data-driven insights. The method proposed in this paper demonstrates outstanding performance in preserving privacy while maintaining data utility. This progress enables the use of sensitive medical data for analysis and predictions without compromising privacy or risking data breaches.

Author Contributions: Conceptualization, J.K.; methodology, S.S.; software, S.S.; validation, S.S.; formal analysis, J.K.; investigation, S.S.; resources, J.K.; data curation, J.K. and S.S.; writing—original draft preparation, S.S.; writing—review and editing, J.K.; visualization, S.S.; supervision, J.K.; project administration, J.K.; funding acquisition, J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a 2022 research grant from Sangmyung University (2022-A000-0166).

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ericsson Mobility Report. Available online: <https://www.ericsson.com/en/reports-and-papers/mobility-report> (accessed on 4 May 2023).
2. Narayanan, A.; Shmatikov, V. How to break anonymity of the Netflix prize dataset. *arXiv* **2007**, arXiv:0610105.
3. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, analysis and future prospects. *J. Big Data* **2019**, *6*, 54. [[CrossRef](#)]
4. Chen, M.; Hao, Y.; Hwang, K.; Wang, L.; Wang, L. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access* **2017**, *5*, 8869–8879. [[CrossRef](#)]
5. Schneble, C.O.; Elger, B.S.; Shaw, D.M. Google's project Nightingale highlights the necessity of data science ethics review. *EMBO Mol. Med.* **2020**, *12*, e12053. [[CrossRef](#)]
6. General Data Protection Regulation. Available online: <https://gdpr-info.eu/> (accessed on 18 April 2023).
7. Sweeney, L. *k*-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 557–570. [[CrossRef](#)]
8. LeFevre, K.; DeWitt, D.J.; Ramakrishnan, R. Incognito: Efficient full domain *k*-anonymity. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, MD, USA, 14–16 June 2005.

9. Mascetti, S.; Freni, D.; Bettini, C.; Wang, X.; Jajodia, S. Privacy in geo-social networks: Proximity notification with untrusted service providers and curious buddies. *Int. J. Very Large Data Bases* **2011**, *20*, 541–566. [[CrossRef](#)]
10. Popa, R.A.; Blumberg, A.J.; Balakrishnan, H.; Li, F.H. Privacy and accountability for location-based aggregate statistics. In Proceedings of the ACM conference on Computer and communications security, Chicago, IL, USA, 17–21 October 2011.
11. Dwork, C. Differential privacy. In Proceedings of the International Conference on Automata, Languages and Programming, Venice, Italy, 10–14 July 2006.
12. Andres, M.E.; Bordenabe, N.E.; Chatzikokolakis, K.; Palamidessi, C. Geo-indistinguishability: Differential privacy for location-based systems. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, Berlin, Germany, 4–8 November 2013; pp. 901–914.
13. Bordenabe, N.E.; Chatzikokolakis, K.; Palamidessi, C. Optimal geo-indistinguishable mechanisms for location privacy. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, New York, NY, USA, 3–7 November 2014; pp. 251–262.
14. Ahuja, R.; Ghinita, G.; Shahabi, C. A utility-preserving and scalable technique for protecting location data with geo-indistinguishability. In Proceedings of the International Conference on Extending Database Technology, Lisbon, Portugal, 26–29 March 2019; pp. 210–231.
15. Qiu, C.; Squicciarini, A.C. Location privacy protection in vehicle-based spatial crowdsourcing via geo-indistinguishability. In Proceedings of the IEEE International Conference on Distributed Computing Systems, Dallas, TX, USA, 7–10 July 2019; pp. 1061–1071.
16. Yao, L.; Mao, C.; Luo, Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 31–39. [[CrossRef](#)] [[PubMed](#)]
17. Hill, S.; Zhou, Z.; Saul, L.; Shacham, H. On the (In) effectiveness of Mosaicing and Blurring as Tools for Document Redaction. *Proc. Priv. Enhancing Technol.* **2016**, *2016*, 403–417. [[CrossRef](#)]
18. Cumby, C.; Ghani, R. A machine learning based system for semi-automatically redacting documents. *Proc. AAAI Conf. Artif. Intell.* **2011**, *25*, 1628–1635. [[CrossRef](#)]
19. Anandan, B.; Clifton, C.; Jiang, W.; Murugesan, M.; Camacho, P.P.; Si, L. t-Plausibility: Generalizing words to desensitize text. *Trans. Data Priv.* **2012**, *5*, 505–534.
20. Sanchez, D.; Batet, M. C-sanitized: A privacy model for document redaction and sanitization. *J. Assoc. Inf. Sci. Technol.* **2016**, *67*, 148–163. [[CrossRef](#)]
21. Yue, X.; Du, M.; Wang, T.; Li, Y.; Sun, H.; Chow, S.M. Differential privacy for text analytics via natural text sanitization. *arXiv* **2021**, arXiv:2106.01221.
22. Chen, H.; Mo, F.; Chen, C.; Cui, J.; Nie, J.Y. A customised text privatisation mechanism with differential privacy. *arXiv* **2022**, arXiv:2207.01193.
23. Carvalho, R.S.; Vasiloudis, T.; Feyisetan, O. BRR: Preserving privacy of text data efficiently on device. *arXiv* **2021**, arXiv:2107.07923.
24. Du, M.; Yue, X.; Chow, S.M.; Sun, H. Sanitizing sentence embeddings (and labels) for local differential privacy. In Proceedings of the ACM Web Conference, Austin, TX, USA, 30 April–4 May 2023; pp. 2349–2359.
25. Feyisetan, O.; Balle, B.; Drake, T.; Diethe, T. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In Proceedings of the International Conference on Web Search and Data Mining, Houston, TX, USA, 3–7 February 2020; pp. 178–186.
26. Wang, Z.; Hu, J.; Lv, R.; Wei, J.; Wang, Q. Personalized privacy-preserving task allocation for mobile crowdsensing. *IEEE Trans. Mob. Comput.* **2018**, *18*, 1330–1341. [[CrossRef](#)]
27. Yan, K.; Luo, G.; Zheng, X.; Tian, L.; Sai, A.M.V.V. A comprehensive location-privacy-awareness task selection mechanism in mobile crowd-wensing. *IEEE Access* **2019**, *7*, 77541–77554. [[CrossRef](#)]
28. Kim, J.W.; Edemacu, K.; Jang, B. Privacy-preserving mechanisms for location privacy in mobile crowdsensing: A survey. *J. Netw. Comput. Appl.* **2022**, *200*, 103315. [[CrossRef](#)]
29. Zhang, P.; Cheng, X.; Su, S.; Wang, N. Area coverage-based worker recruitment under geo-indistinguishability. *Comput. Netw.* **2022**, *217*, 109340. [[CrossRef](#)]
30. Ma, C.; Chen, C.W. Nearby friend discovery with geo-indistinguishability to stalkers. *Procedia Comput. Sci.* **2014**, *34*, 352–359. [[CrossRef](#)]
31. Huang, C.; Lu, R.; Zhu, H.; Shao, J.; Alamer, A.; Lin, X. EPPD: Efficient and privacy-preserving proximity testing with differential privacy techniques. In Proceedings of the IEEE International Conference on Communications, Kuala Lumpur, Malaysia, 22–27 May 2016; pp. 1–6.
32. Tong, W.; Hua, J.; Zhong, S. A jointly differentially private scheduling protocol for ridesharing services. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 2444–2456. [[CrossRef](#)]
33. Shi, D.; Ding, J.; Errapotu, S.M.; Yue, H.; Xu, W.; Zhou, X.; Pan, M. Deep Q-network-based route scheduling for TNC vehicles with passengers' location differential privacy. *IEEE Internet Things J.* **2019**, *6*, 5. [[CrossRef](#)]
34. Ren, W.; Tang, S. EGeoIndis: An effective and efficient location privacy protection framework in traffic density detection. *Veh. Commun.* **2020**, *21*, 100187. [[CrossRef](#)]

35. Chen, R.; Li, L.; Chen, J.J.; Hou, R.; Gong, Y.; Guo, Y.; Pan, M. COVID-19 vulnerability map construction via location privacy preserving mobile crowdsourcing. In Proceedings of the GLOBECOM 2020-2020 IEEE Global Communications Conference, Taipei, Taiwan, 7–11 December 2020.
36. Machanavajjhala, A.; Kifer, D.; Abowd, J.; Gehrke, J.; Vilhuber, L. Privacy: Theory meets practice on the map. In Proceedings of the IEEE International Conference on Data Engineering, Cancun, Mexico, 7–12 April 2008.
37. Xiao, X.; Wang, G.; Gehrke, J. Differential privacy via wavelet transforms. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 1200–1214. [[CrossRef](#)]
38. McSherry, F.D. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. *Commun. ACM* **2010**, *53*, 19–30. [[CrossRef](#)]
39. Xiao, X.; Bender, G.; Hay, M.; Gehrke, J. iReduct: Differential privacy with reduced relative errors. In Proceedings of the ACM SIGMOD International Conference on Management of data, Athens, Greece, 12–16 June 2011.
40. Jang, B.; Kim, I.; Kim, J.W. Word2vec convolutional neural networks for classification of news articles and tweets. *PLoS ONE* **2019**, *14*, e0220976. [[CrossRef](#)] [[PubMed](#)]
41. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
42. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014.
43. Mackiewicz, A.; Ratajczak, W. Principal components analysis (PCA). *Comput. Geosci.* **1993**, *19*, 303–342. [[CrossRef](#)]
44. Chatzikokolakis, K.; Elsalamouny, E.; Palamidessi, C. Efficient utility improvement for location privacy. In Proceedings of the Privacy Enhancing Technologies, Minneapolis, MN, USA, 18–21 July 2017; pp. 210–231.
45. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [[CrossRef](#)]
46. MIMIC-III Clinical Database. Available online: <https://physionet.org/content/mimiciii/1.4/> (accessed on 18 April 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.