

ANALYZING EFFECTS OF PRIVACY MODELS ON SMART HEALTHCARE DATA

ELANUR ÖZMEN¹, ASWANI KUMAR CHERUKURI^{2*} AND ANNAPURNA
JONNALAGADDA³

ABSTRACT. Privacy concerns in smart healthcare arise due to the collection, storage, and sharing of personal health data through various interconnected devices and networks, coupled with the widespread use of data-gathering tools facilitated by advancing technology. This article focuses on the data collected in smart healthcare and the privacy concerns that arise from it. With the increasing adoption of technology in healthcare, there is a growing need to ensure the protection of sensitive health data. Anonymization techniques such as k-anonymity, l-diversity, and t-closeness are essential tools for safeguarding personal information. This paper discusses the importance of these techniques in maintaining privacy and security in the health environment and analyzes the privacy preserving in healthcare databases. It has been concluded that combinations of privacy models used have been successful in ensuring privacy of health data.

Keywords. K-Anonymity, L-Diversity, Privacy Preserving, Sensitive Attributes, Smart Healthcare, T-Closeness

2020 Mathematics Subject Classification. Primary 68T01; Secondary 68T07

1. INTRODUCTION

With the development of technology, collecting, processing, storing, and analyzing data has become easier. However, these conveniences have made it more difficult to protect the privacy of data. The increase in data requires more responsibility for protecting the privacy of sensitive data. The exposure of sensitive data can harm an individual or organization. Protecting sensitive data is crucial not only for maintaining privacy but also for preventing fraud, identity theft, and other harmful actions. To protect data, appropriate security measures such as encryption, access controls, and data backups should be taken. Following best practices for data processing and storage, regularly updating software, and conducting regular security audits are also important. Additionally, organizations

Date: Received: Apr 25, 2025; Accepted: May 16, 2025.

* Corresponding author

© The Author(s) 2025. This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of the licence, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

must comply with relevant laws such as the General Data Protection Regulation (GDPR) [30] or the Health Insurance Portability and Accountability Act (HIPAA).

Anonymization is an important and very common technique used to protect sensitive data. This technique includes different anonymization techniques. This can involve removing or altering names, addresses, dates of birth, and other personal information from data sets. Anonymization is used in various fields, including healthcare, finance, and marketing, to protect sensitive information while still allowing for data analysis and research. However, it is important to note that complete anonymization is not always possible, and there is always a risk of re-identification through data linkage or other means.

The main topic addressed in this article is the protection of privacy by applying three different privacy models to a healthcare dataset using the ARX-Data Anonymization Tool. A small healthcare dataset was created by removing descriptive information such as patient names, which is typically removed from healthcare datasets on the Internet, and replacing it with random information. Firstly, smart healthcare systems were introduced, and the privacy issues in these systems were examined. Then, the identification of sensitive data in a dataset was discussed, as correctly identifying sensitive data is the first step in protecting it and is therefore very important. In addition, the concept of protecting privacy was addressed because learning the scope of this concept in detail is important for the correct application of the concept to protect data. Then, anonymization methods were examined, and k-anonymity, one of the most commonly used anonymization techniques, was examined. However, since k-anonymity is not sufficient for the protection of sensitive data, l-diversity and t-closeness privacy models were examined together with this technique. The use of these three models separately and together was discussed, and a comparison was made among them based on the scores shown by ARX.

T-anonymity is a technique used in data anonymization that involves modifying or concealing personally identifiable information (PII) so that individuals cannot be identified within a dataset. The concept of K-anonymity is based on the idea that a record in a dataset is considered anonymous if it cannot be distinguished from at least K-1 other records. This technique enables the sharing of datasets without compromising the privacy of individuals within the data, making it useful for research purposes. L-diversity is a privacy protection technique used for data anonymization. It is an extension of the K-anonymity technique and is used to ensure that sensitive attributes have at least L well-represented values. Well-represented values are selected based on their frequency in the dataset and must be numerous enough to make it difficult to extract sensitive information. L-diversity provides an additional level of privacy protection beyond K-anonymity by making it harder for an attacker to extract sensitive information even if they are able to identify a specific record in the dataset. T-closeness is a privacy protection technique used to ensure that any sensitive attribute in a dataset is distributed similarly to the distribution of all other attributes. This technique works by controlling the distribution of an attribute in a certain subset of the

dataset. T-closeness can be seen as a combination of L-diversity and K-anonymity techniques.

The main contributions of this research are as follows:

- (1) The lifecycle of data in smart healthcare systems has been examined and privacy issues have been investigated.
- (2) The identification of sensitive data has been thoroughly discussed.
- (3) Privacy models have been examined in the context of privacy protection and elaborated upon in detail through the use of ARX-Anonymization Tool.
- (4) Privacy techniques have been explored through healthcare scenarios.

The remainder of this article is structured as follows. In Section 2, privacy concerns surrounding smart healthcare systems are discussed, followed by Section 3, which details the identification of sensitive attributes within the data. Chapter 4 expounds upon the meaning of privacy preservation, while Section 5 delves into the concept of anonymization. Section 6 provides an overview of various privacy models, while Section 7 outlines the experimental outcomes. Lastly, Section 8 presents a thorough analysis of the experimental results.

2. PRIVACY ISSUES ON THE SMART HEALTHCARE ENVIRONMENT

Smart systems refer to automated systems that utilize advanced technologies such as artificial intelligence (AI), machine learning, and the Internet of Things (IoT) to perform complex tasks and make decisions without human intervention. These systems are designed to analyze and respond to their environment in real-time, adapting to changes to improve their performance and optimize their functionality. Smart systems are typically composed of several components, including sensors, actuators, communication devices, and control systems. These components work together to collect data from the environment, process it, and respond accordingly. For example, a smart HVAC system might use temperature and humidity sensors to monitor the indoor environment and adjust the heating and cooling accordingly. These systems are used in various industries, such as healthcare, transportation, manufacturing, and energy, to improve efficiency, reduce costs, and enhance safety. In this article, smart healthcare systems were mentioned.

Smart healthcare systems are a type of smart system that utilize advanced technologies such as artificial intelligence, machine learning, and the Internet of Things to improve the delivery and quality of healthcare services. These systems are designed to collect and analyze health data from patients and healthcare providers in real-time, enabling more personalized and efficient healthcare. Smart healthcare systems can be used for a variety of applications, including remote patient monitoring, telemedicine, predictive analytics, and drug discovery. For example, remote patient monitoring systems use wearable sensors and other connected devices to track patients' vital signs and activity levels, allowing healthcare providers to monitor their health and intervene if necessary. Telemedicine systems enable patients to consult with healthcare providers remotely, reducing

the need for in-person visits and improving access to care for patients in remote or underserved areas.

The use of smart healthcare systems has the potential to improve healthcare outcomes and reduce costs, as well as enhance patient experience and satisfaction. However, the adoption of these systems also raises concerns about data privacy and security, as well as the potential for bias and discrimination in the algorithms used to analyze health data.

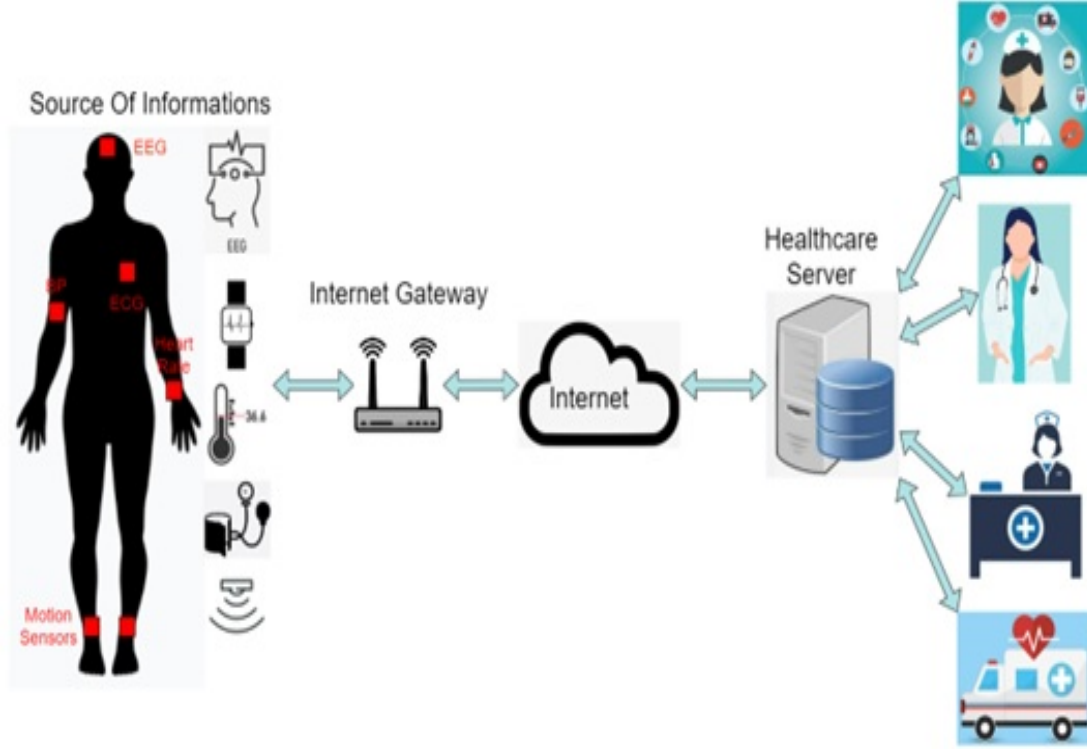


FIGURE 1. Smart healthcare scenario

The physical world layer in healthcare systems refers to the tangible and concrete aspects of healthcare delivery, such as medical facilities, equipment, and supplies. This layer includes devices such as sensors, control systems, communication tools, and the physical environment in which healthcare services are delivered, such as hospitals, clinics, and other healthcare facilities. The gateway layer in healthcare systems refers to the technology infrastructure that allows different healthcare systems and applications to communicate with each other, share data, and exchange information securely. It acts as a bridge between different healthcare systems, enabling the seamless flow of data and information across different platforms.

The gateway layer in healthcare systems refers to the technology infrastructure that allows different healthcare systems and applications to communicate with each other, share data, and exchange information securely. It acts as a bridge between different healthcare systems, enabling the seamless flow of data and information across different platforms. The gateway layer is a critical component

of healthcare systems, as it allows healthcare providers to access and exchange patient data quickly and securely. It enables different healthcare systems to interoperate, which means that patient data can be shared between different systems and applications, even if they were developed by different vendors. The gateway layer is also essential for ensuring data security and compliance with privacy regulations, such as HIPAA. It provides authentication and authorization mechanisms to ensure that only authorized users can access patient data, and it encrypts data transmissions to protect patient privacy.

In Smart healthcare systems, the source of information is humans. Information that is required is collected from the human body through sensors located in the physical layer. Gateway layers act as a bridge between the physical layer and the cloud layer. The gateway layers route data from the storage unit in the physical layer and other incoming data to the data storage and processing units in the cloud layer. In this context, health services and applications in the cloud layer can process and analyze the data. The cloud layer and the healthcare server system are connected through a network infrastructure that allows secure data transfer between the two systems. The healthcare server system typically hosts critical healthcare applications and services such as electronic health records (EHRs), picture archiving and communication systems (PACS), and clinical decision support systems (CDSS). These systems generate and store large amounts of patient data that must be accessible and secure. Healthcare organizations typically use secure network protocols such as virtual private networks (VPNs) and firewalls to connect the cloud layer and the healthcare server system. VPNs provide a secure connection between the cloud layer and the healthcare server system by encrypting data transmitted over the network. Firewalls help prevent unauthorized access to the network by monitoring incoming and outgoing traffic. Only authorized personnel such as nurses, doctors, emergency medical technicians, secretaries, etc., are allowed to access the information on the healthcare servers.

Privacy should be ensured wherever sensitive data is involved. Privacy issues are a major concern for healthcare systems. At the physical layer the reason is because sensitive health data is often transmitted wirelessly, making it susceptible to eavesdropping and interception. Additionally, unauthorized access to medical devices, such as pacemakers and insulin pumps, can have severe consequences for patients. Therefore, privacy and security must be a top priority in the development and implementation of smart healthcare systems to ensure patient trust and safeguard sensitive information. In the gateway layer, privacy issues include data breaches, unauthorized personnel accessing sensitive patient information, and data being compromised if not properly secured. Healthcare systems may use a server to store healthcare data. These servers may be vulnerable to attacks due to security vulnerabilities, leading to the theft of sensitive healthcare information. To ensure the security of servers, strong encryption, secure authentication, monitoring, and update processes are necessary.

3. SENSITIVE ATTRIBUTE IN DATA

Sensitive data can be defined as classified data that needs to be protected with various cybersecurity measures and cannot be accessed by unauthorized individuals or third-party components without privileged access permission. Such data may include sensitive or personal information such as personal identity, financial status, medical condition, ethnic origin, religion, gender, and sexual orientation. The storage of sensitive data in electronic or physical form does not change its nature. In both cases, sensitive data must be carefully protected against cyber threats.

Sensitive data is legally protected in many countries and its processing is subject to strict rules. Legal regulations, such as the General Data Protection Regulation (GDPR) in the European Union (EU), which contains rules on the processing of personal data, make the protection of sensitive data an important issue.

According to GDPR, sensitive personal data, in other words, special categories of personal data, include many different elements such as [30]:

- Race
- Political opinions, ethnic origin, religion, philosophical beliefs, sect
- Style of dress
- Membership in associations, foundations, and unions
- Health information, sexual life
- Criminal convictions, security
- Genetic and biometric data

When determining the sensitivity of data and how it should be categorized, it's crucial to consider the potential impact on confidentiality, integrity, and availability. If sensitive data is revealed, it could result in reputational damage for your organization and threaten the privacy and security of your customers. Additionally, tampering or alteration of sensitive data can lead to misinformation and damage trust in your business. Loss or unavailability of sensitive data can also cause legal and financial consequences and inconvenience for customers. Therefore, it's important to classify sensitive data based on its level of sensitivity and potential impact, and put in place appropriate security measures such as encryption, access controls, and monitoring systems to detect and respond to security breaches.

It is possible to examine sensitive data in terms of different types and levels of security. Sensitive data types are discussed under four different headings, while data sensitivity levels are discussed under three different headings.

3.1. Categories of Sensitive Data. Sensitive data can be categorized based on their levels of sensitivity. Low data sensitivity refers to data that can be disclosed without posing any significant problems for individuals, private organizations, or government agencies. Access to this type of data is usually unrestricted, and it is generally considered as public information that anyone can access. Data in the medium sensitivity category are subject to contracts involving two or more

parties, and the disclosure of such data can result in minimal damage to organizations. Examples of data in this category include student records, IT service information, building plans, and travel information.

High data sensitivity refers to confidential data that can expose organizations to various cyber attacks and may lead to GDPR sanctions. Protected health data, IT security information, social security numbers, and controlled unclassified information are some examples of data falling into this category. Restricted sensitive data is subject to a Non-Disclosure Agreement (NDA) to minimize legal responsibility. This type of data includes trade secrets, credit card details, intellectual property data, customer information, and education records [1].

3.2. Sensitive Data Levels. Highly sensitive data: This class includes special categories of personal data. In the event of a breach of highly sensitive data, serious negative consequences may occur. For example, such a breach could cause significant financial losses to organizations and lead to significant legal sanctions. Medium sensitivity data: Breaches of confidentiality of data used for internal purposes typically do not pose serious problems for organizations.

Low sensitivity data: This group includes publicly available information and poses a low level of concern for individuals, private companies, and government agencies. Access restrictions on this type of data are generally nonexistent or minimal [1].

Sensitive data were analyzed for ARX (<https://arx.deidentifier.org/>). Different colors are used for different features in ARX: Red for identifying attributes, yellow for quasi-identifying attributes, purple for sensitive attributes, green for insensitive attributes. In this, ID is selected as insensitive because it is not identifier in any way. Name is a direct identifier. Gender, age, and ZIP can be identifiers, so quasi-identifier was chosen. Diseases are sensitive data.

First, the attribute whose type is desired to be determined is selected, and then from this part, the type of the selected attribute is determined as shown in Figure 2.

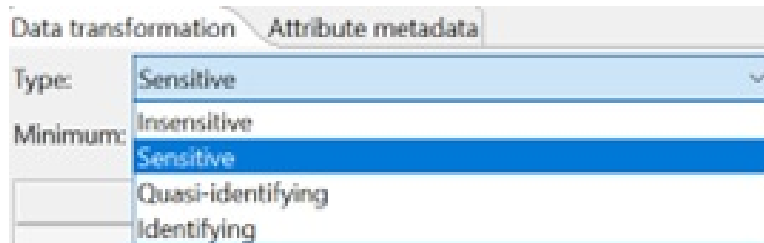


FIGURE 2. The part where the type of attributes is selected

Identifiers are data elements that directly identify individuals and should be eliminated as their use is undesirable. Quasi-identifiers are data elements that enable the linkage of different datasets. It is assumed that the data holder is aware of which attributes are quasi-identifiers. However, in practice, this can be challenging because it depends on the availability of other data in the universe. Sensitive attributes are those data elements that should not be linked to an

individual. The degree of sensitivity depends on the context. Misinterpreting the sensitivity of data elements can result in weaker anonymization than expected. Insensitive attributes are kept as they are [1].

4. PRESERVING PRIVACY

4.1. Privacy. Privacy enables individuals to have control over the collection, use, sharing, and storage of their personal information. Privacy also includes the right for an individual to live their private life without interference and with respect. Generally, privacy covers many different topics, such as the protection of an individual's private life, freedom of information, processing of personal data, digital rights, security, and cyber security.

Privacy is based on the principle that personal information must be collected and processed lawfully. Therefore, many countries have privacy laws that establish rules for the collection, use, and sharing of personal information. Privacy laws aim to prevent the illegal collection and processing of personal information and require that personal information not be shared or sold without explicit consent.

Today, especially in the digital world, privacy has become even more important. The internet and other technological tools have created new challenges and risks for the collection, processing, storage, and sharing of personal information. Therefore, privacy has become increasingly important for the security of personal information and the protection of individuals' private lives.

4.2. Privacy Preserving. Privacy preserving refers to all measures taken to protect individuals' private lives during the collection, processing, storage, and sharing of personal information. These measures aim to ensure that personal information is used only for designated purposes and in accordance with the law, and to prevent unauthorized access to this information. Privacy preserving is important because respecting individuals' privacy is crucial for human rights and fundamental freedoms. Personal information may include a variety of topics, such as an individual's identity, characteristics, habits, health status, financial situation, and other sensitive information. Unauthorized access or misuse of this information may violate individuals' privacy and cause harm to them.

In addition, the security of personal information is essential to ensure individuals' trust and safety. For example, when the security of sensitive information such as health records is compromised, serious consequences such as the dissemination of false information about individuals' health status or the misuse of health services for improper purposes may arise. In conclusion, privacy preserving is crucial for protecting individuals' private lives, avoiding violations of privacy, and ensuring the security of personal information. With the development of the digital world, privacy preserving measures have become even more important, and legal regulations and technological solutions are being developed in this regard [1].

With the ARX application, an attempt was made to preserve privacy by using privacy models on a health database that was created with random values.

TABLE 1. Healthcare database with random values

ID	Name	Gender	Age	ZIP	Disease
1	Allison	Female	21	47000	Flu
2	Lucas	Male	28	47008	Heart Disease
3	Nora	Female	31	47004	Flu
4	Eleanor	Female	35	47012	Gastritis
5	Daniel	Male	22	47073	Heart Disease
6	Madeline	Female	33	47098	Flu
7	Henry	Male	45	47031	Flu
8	Gabriella	Female	47	47031	High BP
9	Joseph	Male	42	47408	Hepatitis B
10	Michael	Male	36	47402	Gastritis
11	Hannah	Female	38	47630	Gastritis
12	Adam	Male	50	47689	High BP
13	Sofia	Female	29	47677	Hepatitis B
14	Nicholas	Male	32	47400	Hepatitis B
15	Emma	Female	41	47601	Heart Disease
16	Jackson	Male	25	47601	High BP

5. ANONYMIZATION TECHNIQUES

Anonymization is the process of making data so that it cannot be associated with any identifiable or determinable real person, even if it is matched with other data. Anonymization is the prevention of the identification of the person concerned by removing or changing all direct and indirect identifiers in a dataset, or by losing the characteristic of being distinguishable within a group or crowd in a way that cannot be associated with a real person.

All procedures for the deletion, destruction and anonymization of personal data are recorded, and these records are kept for at least three years, except for other legal obligations. The data controller is responsible for taking all necessary technical and administrative measures to anonymize personal data. If all the conditions for processing personal data have been eliminated, the data controller deletes, destroys or anonymizes the personal data requested by the data subject. The data controller responds to the data subject's request within thirty days at the latest and informs the data subject.

If it is possible to understand who the data belongs to by matching and supporting other data through monitoring over the data, it cannot be accepted that this data has been anonymized.

5.1. Commonly Used Techniques For Anonymization.

- **Masking** : It is the process of making personal data unidentifiable by deleting or masking certain areas. For example, masking occurs when a portion of a person's credit card number is replaced with asterisks (6698 **** 0006)

- **Aggregation/Cumulative Data Creation:** It refers to the process of aggregating data to reflect total values. For example, data on the number of female employees in a company being Z and 40% of them being university graduates and 60% having a master's degree have been anonymized.
- **Data Derivation:** It involves replacing detailed data with more general equivalents. For example, if the date of birth is replaced with the person's age, the data is anonymized through data derivation.
- **Data Mixing:** It refers to the process of mixing values in a dataset to eliminate the identifiability feature of individuals without causing harm to the overall utility. For example, data mixing occurs when values representing the ages of individuals in a class whose age average is required are exchanged with one another [2].
- **Data generalization:** It is the process of creating a model that expresses the characteristics of examples in a dataset in a more general way by generalizing their features. This process helps to make the dataset more understandable by revealing similarities and differences between examples. Data generalization is particularly important for data privacy and security. Each example in a dataset may contain sensitive information, and the leakage of this information can cause personal privacy violations. Data generalization can help protect sensitive information by allowing analysis of examples based on more general features. Generalization allows us to achieve k-anonymity used in the project.
- **Adding noise to data:** It is a technique used for privacy protection. In this technique, personal information is protected by adding random noise to data while maintaining its accuracy and reliability. This allows operations such as data mining and data analysis to be performed while preventing the disclosure of personal information. For example, adding random noise to age values in each record in a dataset prevents the exact age of each person from being known. In this way, age data can still be used and analyzed, but the privacy of personal information is protected. Adding noise to data is particularly useful for sensitive data sets such as medical and health data. These data sets contain personal information such as individuals' health status, medical history, and medication use. Protecting this data is important for preserving personal privacy and fulfilling legal requirements. Adding noise to data enables the use and analysis of this sensitive data while helping to protect privacy.

5.2. Generalization Hierarchy. The generalization hierarchy is a privacy protection technique that allows data sets to be anonymized at different levels. In this technique, different attributes in the data set are anonymized at different levels depending on their sensitivity. This way, the privacy of more sensitive attributes can be better protected, while less sensitive attributes can still be analyzed.

The generalization hierarchy is usually applied at three different levels. The top level is called complete anonymity and all sensitive attributes are anonymized. At this level, all individuals in the data set are considered the same and individual

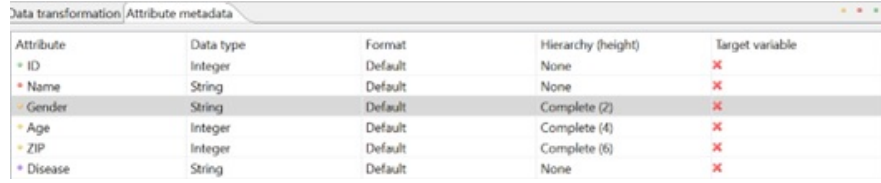
identification is not possible. However, anonymizing the data set in this way can also make data analysis more difficult.

The middle level is called partial anonymity and some sensitive attributes in the data set are anonymized while others are left open. This way, less sensitive attributes can be analyzed while the privacy of more sensitive attributes is still protected.

The bottom level represents a data set with no anonymization. In this case, the sensitive attributes of the data are completely open and each individual can be identified.

The generalization hierarchy is an important technique among the many techniques used to protect data privacy. However, this technique also has some disadvantages and can make data analysis more difficult.

After determining the sensitivity levels of the attributes in Arx, a generalization hierarchy should be created for each quasi-identifier attribute. Because these attributes will be anonymized during k-anonymization. Anonymization will be performed according to the created hierarchy. The user determines the type and levels of the hierarchy. Correctly determining the hierarchy is of great importance for the efficiency of anonymization. Since the hierarchy type will vary depending on the data type, it is also important to determine the data type correctly. Figure 3 below shows the area where the data type can be changed.



Attribute	Data type	Format	Hierarchy (height)	Target variable
ID	Integer	Default	None	✗
Name	String	Default	None	✗
Gender	String	Default	Complete (2)	✗
Age	Integer	Default	Complete (4)	✗
ZIP	Integer	Default	Complete (6)	✗
Disease	String	Default	None	✗

FIGURE 3. The part where attribute details can be seen and edited

ARX includes different generalization hierarchies for different types of attributes. A few of these hierarchies have been addressed in this work [31].

5.2.1. Order-Based Hierarchies. Order-based hierarchy was used when creating the generalization hierarchy for gender. Order-based hierarchy is a privacy protection technique used in data anonymization. In this technique, data is anonymized based on the order of attributes. The idea is to anonymize the data in such a way that the original order of the data is preserved, while at the same time the identity of the individuals in the dataset is protected.

For example, consider a dataset that contains information about the age, gender, and occupation of individuals. An order-based hierarchy for this dataset could be created by first sorting the data based on age, then gender, and finally occupation. The data can then be anonymized by replacing each value in the dataset with the value in the same position in the hierarchy. This ensures that the original order of the data is preserved, while at the same time the identity of the individuals is protected.

Order-based hierarchy is particularly useful in cases where the order of attributes is important for data analysis. By preserving the original order of the

TABLE 2. Generalization hierarchy for gender

Level-0	Level-1
Female	{Female, Male}
Male	{Female, Male}

data, it allows for more accurate analysis of the data while still protecting the privacy of the individuals in the dataset.

However, it is important to note that the effectiveness of order-based hierarchy depends on the specific dataset and the privacy risks associated with it. As with any anonymization technique, careful consideration must be given to the specific use case and potential privacy risks before applying order-based hierarchy [7]. In the database in Table 1, the order-based hierarchy is selected for the gender attribute and this hierarchy is shown in Table 2.

5.2.2. Masking-Based Hierarchies. Anonymization using masking is a privacy-preserving technique that ensures the confidentiality of sensitive data while maintaining its usefulness for research and analysis purposes. This technique involves hiding or masking certain attributes of the data while keeping others visible. Masked attributes can be replaced with fake or random values, like using the initials of a person’s name instead of their full name or obscuring the last digits of a Social Security Number.

Masking techniques can be applied to various types of data, including structured and unstructured data, such as text, audio, and visual data. The effectiveness of masking depends on the level of masking used and the type of data being protected.

There are different types of masking techniques, such as random, deterministic, and permutation masking. Random masking involves replacing sensitive data with random values, while deterministic masking uses a fixed algorithm to mask the data. Permutation masking involves re-ordering the values of a dataset to preserve its statistical properties while obscuring sensitive information.

One popular method of anonymization with masking is k-anonymity, which involves grouping similar data points together to create a minimum group size of k. Within each group, the sensitive attributes are masked or replaced with random values to ensure that the group cannot be identified based on any individual attribute. This technique helps to prevent re-identification attacks on the data.

Although anonymization with masking can be effective in protecting sensitive data, it is essential to note that it is not foolproof. Researchers and analysts must take adequate precautions when handling masked data to prevent potential privacy breaches or re-identification attacks [7, 8]. In the database in Table 1, the masking-based hierarchy is selected for the ZIP Code attribute and this hierarchy is shown in Table 3.

5.2.3. Interval-Based Hierarchies. Anonymization with interval-based techniques involves replacing numerical values with intervals or ranges to protect sensitive data. This method is commonly used when the exact value of a data point is

TABLE 3. Generalization hierarchy for ZIP Code

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
47000	4700*	470**	47***	4****	*****
47004	4700*	470**	47***	4****	*****
47008	4700*	470**	47***	4****	*****
47012	4701*	470**	47***	4****	*****
47031	4703*	470**	47***	4****	*****
47073	4707*	470**	47***	4****	*****
47098	4709*	470**	47***	4****	*****
47400	4740*	474**	47***	4****	*****
47402	4740*	474**	47***	4****	*****
47408	4740*	474**	47***	4****	*****
47601	4760*	476**	47***	4****	*****
47630	4763*	476**	47***	4****	*****
47677	4767*	476**	47***	4****	*****
47689	4768*	476**	47***	4****	*****

not necessary but the general category is still relevant. For instance, age can be represented as an interval rather than the exact age.

The technique involves dividing the data into intervals and replacing the original values with the corresponding intervals. The size and granularity of the intervals can be adjusted to balance privacy and utility concerns. Data can be grouped by similar characteristics to create larger intervals and further protect privacy.

One popular method is k-interval anonymity, which involves grouping similar data points together to create a minimum group size of k and replacing the values with the corresponding interval. This technique can help prevent re-identification attacks on the data.

Interval-based anonymization techniques can be used to protect numerical data such as ages, income levels, or zip codes. They can be applied to both structured and unstructured data, including text, audio, and visual data. While they can effectively protect sensitive data, there is a potential loss of accuracy or utility in the anonymized data that should be considered [11, 12]. In the database in Table 1, the interval-based hierarchy is selected for the age attribute and this hierarchy is shown in Table 4.

6. PRIVACY MODELS

6.1. K-Anonymity. K-anonymity is a privacy protection technique used in data anonymization. It involves masking or modifying personally identifiable information (PII) so that individuals are not identifiable within a dataset. The concept of K-anonymity is based on the idea that, in a dataset, a record is considered anonymous if it cannot be distinguished from at least K-1 other records. The value of K is chosen based on the sensitivity of the data being anonymized, and the desired level of privacy protection.

TABLE 4. Generalization hierarchy for age

Level-0	Level-1	Level-2	Level-3
21	[20,30[	[20,40[	*
22	[20,30[	[20,40[	*
25	[20,30[	[20,40[	*
28	[20,30[	[20,40[	*
29	[20,30[	[20,40[	*
31	[30,40[	[20,40[	*
32	[30,40[	[20,40[	*
33	[30,40[	[20,40[	*
35	[30,40[	[20,40[	*
36	[30,40[	[20,40[	*
38	[30,40[	[20,40[	*
41	>=41	>=41	*
42	>=41	>=41	*
45	>=41	>=41	*
47	>=41	>=41	*
50	>=41	>=41	*

To achieve K-anonymity, data is transformed by generalizing or suppressing some of the attributes or fields of a dataset. For example, a dataset containing age, gender, and zip code could be generalized by replacing specific ages with age ranges and replacing exact zip codes with less specific geographic regions. It ensures that an individual's identity is protected, as their data is mixed in with that of other individuals, making it difficult to identify any individual from the dataset.

One of the advantages of K-anonymity is that it allows for the sharing of datasets without compromising the privacy of individuals within the data, making it useful for research purposes, where sensitive data needs to be shared with third parties. However, K-anonymity is not a foolproof method of data anonymization, as it can be vulnerable to attacks such as the homogeneity attack, where an attacker can use external information to identify individuals within a dataset. Overall, K-anonymity is a useful technique for protecting individual privacy in large datasets, but it is important to consider its limitations and potential vulnerabilities [4]. K-anonymity is applied when each record in a dataset containing sensitive data is identical to at least K other records. This means that a record can be seen as part of a group with others, making it more difficult to be individually identified. The formula for K-anonymity is:

$$K - anonymity = mincount(e)(e : QI)$$

Here, e represents non-sensitive attributes, and QI represents the attributes that contain sensitive information that could reveal identities. The term count(e) represents the number of records with the value e . For example, if a dataset contains attributes such as name, age, and postal code, the postal code can be considered a sensitive attribute. In this case, K-anonymity is achieved when there

are at least K records with the same name and age. For instance, if $k = 3$, for each postal code value, each record in the dataset should have at least three other records with the same name and age attributes [4].

6.2. L-Diversity. L-diversity is a formal privacy model that addresses the limitations of k-anonymity. Although k-anonymity ensures that a group of individuals cannot be distinguished based on their sensitive attributes, it does not provide protection against re-identification attacks on the distribution of those attributes. L-diversity addresses this limitation by requiring that each group of k individuals has l distinct sensitive attribute values, thereby ensuring that the sensitive values are not concentrated within the group [6].

Several variations of l-diversity exist, including entropy l-diversity, recursive l-diversity, and multi-attribute l-diversity. Entropy l-diversity measures the entropy of the sensitive attribute values within a group, and requires that it exceeds a certain threshold. Recursive l-diversity applies l-diversity recursively to the quasi-identifiers, creating a hierarchy of levels of l-diversity. Multi-attribute l-diversity extends l-diversity to multiple sensitive attributes.

L-diversity has been applied to a variety of data, such as healthcare data, social network data, and location data. It has also been combined with other privacy models, such as t-closeness, to provide more robust privacy guarantees. L-diversity is a privacy model that aims to preserve the distribution of sensitive data, in contrast to the k-anonymity model that only guarantees that a group of k individuals cannot be distinguished based on their sensitive attributes. L-diversity also requires that each group of k individuals must have distinct sensitive attribute values. This ensures that the sensitive values are not overly concentrated within the group, thereby increasing the protection against re-identification attacks [8, 13].

6.3. T-Closeness. T-closeness is a privacy model that aims to prevent attribute disclosure by ensuring that the distribution of sensitive attribute values in a group of individuals is not too far from the distribution in the overall population [4].

T-closeness builds upon k-anonymity and l-diversity by addressing their limitations. While k-anonymity ensures that each group of individuals is indistinguishable based on their quasi-identifiers, and l-diversity ensures that each group has a diverse range of sensitive attribute values, neither guarantees that the distribution of sensitive values in a group reflects the distribution in the overall population.

T-closeness requires that the distribution of sensitive attribute values in a group should be no more than a certain distance (i.e., closeness) from the distribution in the overall population. The distance is calculated using a distance metric, such as the Kullback-Leibler divergence or the earth mover's distance. The value of t in t-closeness represents the maximum allowed distance between the distributions, and it is typically set to a small value to ensure strong privacy guarantees.

There are several variations of t-closeness, including bounded distance t-closeness, multi-dimensional t-closeness, and weighted t-closeness. Bounded distance t-closeness limits the distance metric to a certain bound, to prevent outliers from affecting the overall distance. Multi-dimensional t-closeness extends t-closeness

to multiple sensitive attributes. Weighted t-closeness assigns different weights to different sensitive attribute values, based on their importance or sensitivity.

T-closeness has been applied to various types of data, including healthcare data, social network data, and location data. It has been shown to provide stronger privacy guarantees than k-anonymity and l-diversity, and can be used in combination with them to provide even stronger privacy guarantees [5, 14]. This model requires that the distribution of sensitive attributes of records forming a k-anonymity group is under a certain limit when compared to a reference distribution. The limit is known as the t-parameter.

The *t*-closeness formula is as follows: $D(P_i||P^*) \leq t$

Here, P_i represents a sensitive attribute in a dataset, P^* is the reference distribution, and the *D* function measures the distance between two probability distributions. *t* is the limit value set as the t-parameter. The *t*-closeness model can be applied to various types of data. It has been used for healthcare data, geographic data, social media data, financial data, and many other types of data [15]. An overview of the literature on these 3 privacy models is given in Table 5.

Table 5: Summary of Literature

S.No	Article	Technique	Database	Result
1	t-Closeness: Privacy beyond k-anonymity and l-diversity [5]	t-Closeness	Healthcare database	The article presents an experimental study involving the application of t-closeness, k-anonymity, and l-diversity techniques to a dataset. The results demonstrate that the t-closeness model provides better privacy protection than the other two models.
2	Quantifying the Effects of Anonymization Techniques Over Micro-Databases [32]	Generalization, bucketization, randomization, and combination	Micro databases	The results showed that the effectiveness of the anonymization techniques varied depending on the specific micro-database and the chosen anonymity level. The authors concluded that a combination of different anonymization techniques could provide better privacy protection for micro-databases.

Table 5 – continued from previous page

S.No	Article	Technique	Database	Result
3	Data privacy preservation algorithm with k-anonymity	k-Anonymity	Healthcare database	The results show that the algorithm is effective in protecting the privacy of sensitive data while still maintaining data utility. It was concluded that the proposed algorithm can be used as a tool for healthcare data sharing while preserving data privacy.
4	t-Closeness: A New Privacy Metric for Data Publishing [14]	t-Closeness	More on healthcare databases and social security databases	This metric is suggested as an approach that can be used in addition to other techniques such as k-anonymity and l-diversity to preserve the privacy of data. In this study, the effectiveness of t-closeness was examined on different datasets, and the results showed that it outperformed other privacy techniques.
5	Multi-attribute k-anonymity with l-diversity [13]	k-Anonymity, l-diversity	More on healthcare databases and social security databases	The paper introduces Multi-Attribute k-Anonymity with l-Diversity, a technique that uses k-anonymity and l-diversity to protect privacy and prevent attribute disclosure. Real-world datasets are used to evaluate the technique, considering information loss and data utility. Results indicate that this technique surpasses the traditional k-anonymity and l-diversity methods in preserving privacy and maintaining data utility.

K-anonymity, l-diversity, and t-closeness are three different privacy models proposed to protect the confidentiality of healthcare databases. The reason why I chose these three models is that they are proven and effective methods for preserving the privacy of health data.

K-anonymity, l-diversity, and t-closeness models work together to provide a higher level of privacy protection for healthcare databases. Their combination ensures that each model complements the other, resulting in a more secure and reliable method for preserving the confidentiality of health data. Therefore, the combination of k-anonymity, l-diversity, and t-closeness is considered a trustworthy method for protecting the privacy of health data.

6.4. Different Anonymization techniques and privacy models: Anonymization techniques are methods that aim to securely share data by removing personal information from its original state. During the process of anonymization, certain criteria are used to reduce the decrease of the usefulness of the data. Techniques such as removal of identifying information, merging data to enhance anonymity, and adding noise are employed. However, it is important to note that anonymization does not completely eliminate the personal information contained within the original data. While anonymization techniques aim to facilitate secure data sharing, they do not offer an absolute guarantee of privacy.

Privacy models, on the other hand, are mathematical models used to protect private data. While they also assist in anonymizing data, they are designed to safeguard the privacy of the data. Privacy models such as K-anonymity, L-diversity, and T-closeness are used to preserve the original state of the data while ensuring its protection. These models employ various criteria during the anonymization process to safeguard private data. In summary, anonymization techniques and privacy models serve different purposes in ensuring data privacy and security. Anonymization techniques aim to facilitate secure data sharing, while privacy models use mathematical models to protect private data.

7. EXPERIMENTAL RESULTS

7.1. Experimental Setup. The database used in this study is a database containing personal healthcare data. Healthcare database was chosen because the increasing data collection, storage, and processing with advancing technology have started to create privacy issues for healthcare data, which is very important in terms of privacy. This database was created with random values because typically, sensitive information is not available in databases shared on the internet, and as a result, observing privacy techniques remains superficial. Therefore, a database was created including sensitive data. The database has 6 different attributes, ID, name, gender, age, ZIP Code, and disease. After examining the examples, it was understood that the use of privacy models for each database may not be comfortable, so these random values were carefully selected to ensure the optimum performance of the privacy models.

7.2. Experimental Results for K-Anonymity. When applying k-anonymity to the database, the sensitivity of the ID was chosen as insensitive, name as

direct identifier, and gender, age, and zip code as quasi-identifiers. Diseases are also considered quasi-identifier because k-anonymity alone does not protect sensitive data. If sensitive data is available, l-diversity or t-closeness should also be applied. Since only k-anonymity will be shown in this section, the disease was also chosen as quasi-identifier and a generalization hierarchy was created for the diseases. In the database in Table 1, the order-based hierarchy is selected for the disease attribute and this hierarchy is shown in Table 6.

TABLE 6. Generalization hierarchy for disease

Level-0	Level-1	Level-2	Level-3
Flu	Upper Respiratory Tract	Contagious	*
Hepatitis B	Liver	Contagious	*
Gastritis	Stomach	Non-infectious	*
Heart Disease	Heart	Non-infectious	*
High BP	Heart	Non-infectious	*

The table and groups formed when 2-anonymity is applied to the table given in Table 1 are shown in Figure 4. The groups are formed in such a way that they cannot be distinguished from at least k-1 other records. The ID has not been changed since it is insensitive, while the Name identifier has been concealed. The quasi-identifier values have been modified using generalization hierarchies to provide the most optimal result according to the calculations performed by ARX.

Name	Gender	Age	ZIP	Disease	Name	Gender	Age	ZIP	Disease
Allison	Female	21	47000	Flu	*	Female	[20, 30]	47***	Contagious
Sofia	Female	29	47677	Hepatitis B	*	Female	[20, 30]	47***	Contagious
Nora	Female	31	47004	Flu	*	Female	[30, 40]	47***	Contagious
Madeline	Female	33	47098	Flu	*	Female	[30, 40]	47***	Contagious
Henry	Male	45	47031	Flu	*	Male	>=41	47***	Contagious
Joseph	Male	42	47408	Hepatitis B	*	Male	>=41	47***	Contagious
Emma	Female	41	47601	Heart Disease	*	Female	>=41	47***	Non-infectious
Gabriella	Female	47	47031	High BP	*	Female	>=41	47***	Non-infectious
Eleanor	Female	35	47012	Gastritis	*	Female	[30, 40]	47***	Non-infectious
Hannah	Female	38	47630	Gastritis	*	Female	[30, 40]	47***	Non-infectious
Daniel	Male	22	47073	Heart Disease	*	Male	[20, 30]	47***	Non-infectious
Lucas	Male	28	47008	Heart Disease	*	Male	[20, 30]	47***	Non-infectious
Jackson	Male	25	47601	High BP	*	Male	[20, 30]	47***	Non-infectious
Michael	Male	36	47402	Gastritis	*	*	*	*	*
Adam	Male	50	47689	High BP	*	*	*	*	*
Nicholas	Male	32	47400	Hepatitis B	*	*	*	*	*

FIGURE 4. Tables and groups after applying 2-anonymity

There are 3 patients in the table who cannot be grouped. These patients could not be grouped because there are no other records that they could be grouped with. For instance, if the completely hidden patients were not hidden, their values would appear as follows, respectively:

10, *, Male, [30,40[, 47***, Non-infectious 12, *, Male, >=41, 47***, Non-infectious 14, *, Male, [30,40[, 47***, Contagious

In this case, when looking at the table, there are no $k-1$ (2-1) values that can be grouped with these records. Therefore, these records must be hidden in order to protect privacy.

When looking at this data, for example, a person looking at the first group could understand that there are two women between the ages of 20 and 30 in this region who have a contagious disease. This would result in a privacy breach,

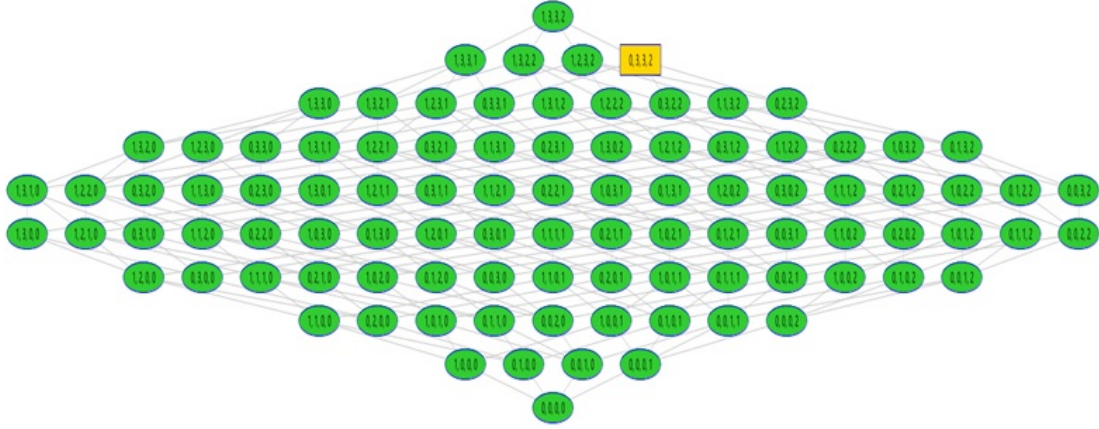


FIGURE 7. Explore Results screen for 3-anonymity

L-Diversity focuses on sensitive data. The 2-diversity applied version of Table 1 is shown in Figure 8.

Name	Gender	Age	ZIP	Disease	Name	Gender	Age	ZIP	Disease
Gabriella	Female	47	47031	High BP	*	Female	>=41	47***	High BP
Emma	Female	41	47601	Heart Disease	*	Female	>=41	47***	Heart Disease
Allison	Female	21	47000	Flu	*	Female	[20, 30[	47***	Flu
Sofia	Female	29	47677	Hepatitis B	*	Female	[20, 30[	47***	Hepatitis B
Nora	Female	31	47004	Flu	*	Female	[30, 40[	47***	Flu
Madeline	Female	33	47098	Flu	*	Female	[30, 40[	47***	Flu
Eleanor	Female	35	47012	Gastritis	*	Female	[30, 40[	47***	Gastritis
Hannah	Female	38	47630	Gastritis	*	Female	[30, 40[	47***	Gastritis
Henry	Male	45	47031	Flu	*	Male	>=41	47***	Flu
Joseph	Male	42	47408	Hepatitis B	*	Male	>=41	47***	Hepatitis B
Adam	Male	50	47689	High BP	*	Male	>=41	47***	High BP
Daniel	Male	22	47073	Heart Disease	*	Male	[20, 30[	47***	Heart Disease
Lucas	Male	28	47008	Heart Disease	*	Male	[20, 30[	47***	Heart Disease
Jackson	Male	25	47601	High BP	*	Male	[20, 30[	47***	High BP
Nicholas	Male	32	47400	Hepatitis B	*	Male	[30, 40[	47***	Hepatitis B
Michael	Male	36	47402	Gastritis	*	Male	[30, 40[	47***	Gastritis

FIGURE 8. Tables and groups after applying 2-diversity

As can be seen, 2-diversity groups patients with at least two different diseases. This prevents a single sensitive attribute from being prevalent in a group and therefore prevents re-identification. In Figure 8, only 2-diversity has been applied. However, when 2-diversity is applied after 2-anonymity, the same result is obtained. This is because, for this example, 2-diversity already satisfies the 2-anonymity condition.

As can be seen in Figure 9, the optimal result consists of 0 anonymity levels for gender, 1 anonymity level for age, and 3 anonymity levels for ZIP code.

In Figure 10, 3-anonymity and 2-diversity have been applied to the tables. Anonymization was performed using 0 anonymity levels for gender, 3 anonymity levels for age, and 2 anonymity levels for ZIP code.

7.4. Experimental Results for T-Closeness. When applying t-closeness to the database, the sensitivity of the ID was chosen as insensitive; name as direct identifier; gender, age, and zip code as quasi-identifiers and disease as sensitive.



FIGURE 9. Explore Results screen for 2-diversity

Name	Gender	Age	ZIP	Disease	Name	Gender	Age	ZIP	Disease
Allison	Female	21	47000	Flu		Female	*	470**	Flu
Nora	Female	31	47004	Flu		Female	*	470**	Flu
Madeline	Female	33	47098	Flu		Female	*	470**	Flu
Eleanor	Female	35	47012	Gastritis		Female	*	470**	Gastritis
Gabriella	Female	47	47031	High BP		Female	*	470**	High BP
Henry	Male	45	47031	Flu		Male	*	470**	Flu
Daniel	Male	22	47073	Heart Disease		Male	*	470**	Heart Disease
Lucas	Male	28	47008	Heart Disease		Male	*	470**	Heart Disease
Joseph	Male	42	47408	Hepatitis B		Male	*	474**	Hepatitis B
Nicholas	Male	32	47400	Hepatitis B		Male	*	474**	Hepatitis B
Michael	Male	36	47402	Gastritis		Male	*	474**	Gastritis
Sofia	Female	29	47677	Hepatitis B		Female	*	476**	Hepatitis B
Hannah	Female	38	47630	Gastritis		Female	*	476**	Gastritis
Emma	Female	41	47601	Heart Disease		Female	*	476**	Heart Disease
Adam	Male	50	47689	High BP		Male	*	476**	High BP
Jackson	Male	25	47601	High BP		Male	*	476**	High BP

FIGURE 10. Tables and groups after applying 3-anonymity and 2-diversity

T-closeness focuses on sensitive data. The version of Table 1 with 0.7-closeness applied is shown in Figure 11. As the T value decreases, the protection increases but the data loss also increases.

Name	Gender	Age	ZIP	Disease	Name	Gender	Age	ZIP	Disease
Gabriella	Female	47	47031	High BP		Female	>=41	47***	High BP
Emma	Female	41	47601	Heart Disease		Female	>=41	47***	Heart Disease
Allison	Female	21	47000	Flu		Female	[20, 30]	47***	Flu
Sofia	Female	29	47677	Hepatitis B		Female	[20, 30]	47***	Hepatitis B
Nora	Female	31	47004	Flu		Female	[30, 40]	47***	Flu
Eleanor	Female	35	47012	Gastritis		Female	[30, 40]	47***	Gastritis
Madeline	Female	33	47098	Flu		Female	[30, 40]	47***	Flu
Hannah	Female	38	47630	Gastritis		Female	[30, 40]	47***	Gastritis
Henry	Male	45	47031	Flu		Male	>=41	47***	Flu
Joseph	Male	42	47408	Hepatitis B		Male	>=41	47***	Hepatitis B
Adam	Male	50	47689	High BP		Male	>=41	47***	High BP
Lucas	Male	28	47008	Heart Disease		Male	[20, 30]	47***	Heart Disease
Daniel	Male	22	47073	Heart Disease		Male	[20, 30]	47***	Heart Disease
Jackson	Male	25	47601	High BP		Male	[20, 30]	47***	High BP
Nicholas	Male	32	47400	Hepatitis B		Male	[30, 40]	47***	Hepatitis B
Michael	Male	36	47402	Gastritis		Male	[30, 40]	47***	Gastritis

FIGURE 11. Tables and groups after applying 0.7-closeness

In Figure 11, only 0.7-closeness has been applied. However, when 0.7-closeness is applied after 2-anonymity, the same result is obtained. This is because, for this example, 0.7-closeness already satisfies the 2-anonymity condition.

Figure 12 shows a table where 2-anonymity and 0.6-closeness are applied. The T value has been decreased compared to the example in Figure 18, resulting in increased protection but also data loss.

Name	Gender	Age	ZIP	Disease	Name	Gender	Age	ZIP	Disease
Allison	Female	21	47000	Flu	*	Female	[20, 40[	47***	Flu
Nora	Female	31	47004	Flu	*	Female	[20, 40[	47***	Flu
Eleanor	Female	35	47012	Gastritis	*	Female	[20, 40[	47***	Gastritis
Madeline	Female	33	47098	Flu	*	Female	[20, 40[	47***	Flu
Sofia	Female	29	47677	Hepatitis B	*	Female	[20, 40[	47***	Hepatitis B
Hannah	Female	38	47630	Gastritis	*	Female	[20, 40[	47***	Gastritis
Henry	Male	45	47031	Flu	*	Male	>=41	47***	Flu
Joseph	Male	42	47408	Hepatitis B	*	Male	>=41	47***	Hepatitis B
Adam	Male	50	47689	High BP	*	Male	>=41	47***	High BP
Lucas	Male	28	47008	Heart Disease	*	Male	[20, 40[	47***	Heart Disease
Daniel	Male	22	47073	Heart Disease	*	Male	[20, 40[	47***	Heart Disease
Nicholas	Male	32	47400	Hepatitis B	*	Male	[20, 40[	47***	Hepatitis B
Michael	Male	36	47402	Gastritis	*	Male	[20, 40[	47***	Gastritis
Jackson	Male	25	47601	High BP	*	Male	[20, 40[	47***	High BP
Gabriella	Female	47	47031	High BP	*	*	*	*	High BP
Emma	Female	41	47601	Heart Disease	*	*	*	*	Heart Disease

FIGURE 12. Tables and groups after applying 2-anonymity and 0.6-closeness

8. EXPERIMENTAL ANALYSIS

The closer the score on this screen is to 0, the less likely the result is to be re-identified, meaning it is better. ARX calculates this score and selects the optimal box with the best score. As a result of applying 3-anonymity to the database in Table 1, the score was found to be greater than the score of 2-anonymity. Therefore, we can compare 2-anonymity as more optimal for this database. However, it should not be forgotten that data loss also occurred in 2-anonymity. Losses and gains in privacy should be measured carefully.

When 2-anonymity and 2-diversity are applied to the database, the resulting score is closer to 0 compared to when only k-anonymity is applied. Therefore, we can say that applying 2-anonymity and 2-diversity to this database provides a better result than only applying 2-anonymity or 3-anonymity. After applying 3-anonymity and 2-diversity to the database, the score of 0.48 was calculated, which is worse than the score of 2-anonymity and 2-diversity.

The score of the 0.7-closeness is 0.38, is greater than the score of 2-anonymity and 2-diversity. Therefore, based on this score, we can say that 2-anonymity and 2-diversity are better protection options for this database. It's important to note that these techniques may have different levels of effectiveness and may be more suitable for certain types of data or use cases. Additionally, some techniques may require additional measures to ensure the security and confidentiality of the data, such as secure communication protocols and access controls. Following Table 7 provides a list of different techniques and their relevance to different types of data.

K-anonymity is used to group individuals with similar attributes, while l-diversity is used to increase the diversity of the distribution of these groups. T-closeness controls the proximity of groups based on sensitive attributes. The k-anonymity method is more effective for large databases because it is easier to group unique identities due to the large number of records in such databases. These databases usually contain records of students, patients, or employees.

L-diversity is an anonymization method that provides more diversity of sensitive attributes within a group with different values. For example, if a group of people has different ethnicities, the l-diversity method ensures that the group is more diverse in terms of their ethnicities. L-diversity is a more effective method

TABLE 7. Relevance of different Privacy Techniques for Protecting Data

S.No	Techniques	Type of Data
1	K-Anonymity	Big databases, personal data, health data, geographic data, transaction data
2	L-Diversity	Sensitive data (medical records, ethnicity, etc.)
3	T-Closeness	Sensitive data (medical records, ethnicity, etc.)
4	Differential privacy	Personal data, health data, geographic data, social network data, financial information
5	Data masking	Financial data, health data, transaction data, personal data
6	Data mixing	Personal data, health data, geographic data

for heterogeneous data. This technique is particularly effective in protecting healthcare data, where many people have similar illnesses. This technique is effective regardless of the size of the database, given the sensitive nature of the data.

The T-closeness method is an anonymization method used when the distribution of sensitive data is close to each other in the database records. This method is more effective for anonymizing data in situations where the distribution of sensitive data is homogeneous. T-closeness is often used to protect financial data, and it is suitable for smaller databases due to the sensitivity of the data.

Large databases are generally anonymized using the k-anonymity method, while l-diversity and t-closeness methods can also be used in different scenarios. L-diversity method can be more effective, particularly when sensitive data is heterogeneous. Similarly, the t-closeness method can be used more effectively for anonymization in situations where the distribution of sensitive data is homogeneous.

In conclusion, the choice of anonymization method depends on the characteristics of the data. The k-anonymity method is the most commonly used method for large databases, but the l-diversity and t-closeness methods can also be used in different scenarios and for different purposes. Therefore, the characteristics of the data should be considered when selecting an anonymization method.

9. CONCLUSION

In conclusion, the development of smart healthcare environments has created significant privacy concerns that must be addressed to protect individuals' sensitive data. The European Convention on Human Rights recognizes the right to privacy under Article 8, which stipulates that "everyone has the right to respect for his private and family life, his home and his correspondence." As such, privacy is a fundamental human right, and the use of privacy-preserving techniques is essential to protect this right. Sensitive healthcare data can reveal personal information that should be protected to prevent potential harm to individuals. The use of privacy-preserving techniques such as k-anonymity, l-diversity, and

t-closeness can help mitigate the risk of privacy breaches and protect individuals' sensitive data.

However, the implementation of these techniques is not always straightforward, and their effectiveness depends on several factors, including the data set, the attack model, and the privacy risk model. Therefore, it is essential to develop appropriate policies and regulations to ensure the effective implementation of privacy-preserving techniques in smart healthcare environments. Smart healthcare environments provide an opportunity for the collection and analysis of vast amounts of sensitive data. However, the use of this data also raises significant privacy concerns. It is crucial to identify sensitive data elements and adopt privacy-preserving techniques to protect the privacy of individuals. With the increasing use of technology in healthcare, it is necessary to maintain a balance between the benefits of using this data and the potential harm that can result from a privacy breach.

K-anonymity, l-diversity, and t-closeness are three popular privacy-preserving techniques used to protect sensitive data in healthcare environments. K-anonymity ensures that each record in a dataset cannot be distinguished from at least $k-1$ other records. L-diversity, on the other hand, ensures that each sensitive attribute has at least l well-represented values in a dataset. T-closeness is a privacy metric that measures the distance between the distribution of a sensitive attribute in a dataset and its distribution in the overall population.

These three privacy models were tested using ARX which is a data anonymization tool for privacy protection techniques. The k-anonymity, l-diversity, and t-closeness techniques were tested with ARX to determine which techniques were more effective. The results may vary for different datasets, in this study, the combination of k-anonymity and l-diversity has yielded the most effective result for the used database, while the use of t-closeness has provided the second most effective result. However, a literature review reveals that t-closeness is generally more efficient.

REFERENCES

1. Kron.com.tr. (2022, March 9). Hassas veri türleri nelerdir? Nasıl korunabilirler? <https://kron.com.tr/hassas-veri-turleri-nelerdir-nasil-korunabilirler/>
2. Bicakci, K., Sahingoz, O. K., & Levi, A. (2011). Privacy and data protection impact assessment frameworks for RFID applications in healthcare environments. *Journal of medical systems*, 35(5), 977-987. [10.1007/s10916-010-9481-1](https://doi.org/10.1007/s10916-010-9481-1)
3. Sweeney, L. (2002). k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557-570. <https://doi.org/10.1142/S0218488502001648>
4. Li, N., Li, T., & Venkatasubramanian, S. (2007). T-closeness: Privacy beyond k-anonymity and l-diversity. 2007 IEEE 23rd International Conference on Data Engineering. [10.1109/ICDE.2007.367856](https://doi.org/10.1109/ICDE.2007.367856)
5. Kumar, P., & Kushwaha, R. S. (2019). Anonymization Techniques: A Review. *International Journal of Engineering and Advanced Technology (IJEAT)*, 8(5), 623-626. <https://www.ijeast.com/papers/310-316,Tesma410,IJEAST.pdf>
6. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkatasubramanian, M. (2006). l-diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 3. <https://dl.acm.org/doi/10.1145/1217299.1217302>

7. Lui, K. H., & Wong, R. C. (2014). Anonymization with Masking. In *Privacy-Preserving Data Publishing* (pp. 49-72). Springer. [10.1109/ACCESS.2020.3045700](https://doi.org/10.1109/ACCESS.2020.3045700)
8. El Emam, K., Dankar, F. K., & Neisa, A. (2009). Protecting Privacy using k-Anonymity. *Journal of the American Medical Informatics Association*, 16(1), 130-133. [10.1197/jamia.M2716](https://doi.org/10.1197/jamia.M2716)
9. Li, J., Zhang, S., & Li, Q. (2012). An Interval-Based Anonymization Method for Protecting Privacy of Continuous Data. *Journal of Computational Information Systems*, 8(9), 4029-4036. [10.1197/jamia.M2716](https://doi.org/10.1197/jamia.M2716)
10. Zhang, N., Hong, B., & Wang, F. (2015). A Survey of Interval-Based Anonymization Techniques. *International Journal of Information Security and Privacy (IJISP)*, 9(3), 35-53. [10.3844/jcssp.2020.194.201](https://doi.org/10.3844/jcssp.2020.194.201)
11. Chen, R., Fung, B. C., & Desai, B. C. (2012). Multi-attribute k-anonymity with l-diversity. *IEEE Transactions on Knowledge and Data Engineering*, 24(4), 651-663.
12. N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: A New Privacy Metric for Data Publishing. *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 5, pp. 615-627, 2011.
13. Kerschbaum, F. (2015). T-Closeness: A review. *arXiv preprint arXiv:1504.02529*.
14. Prasser, F., Kohlmayer, F., Lautenschläger, R., & Kuhn, K. A. (2014). ARX—A Comprehensive Tool for Anonymizing Biomedical Data. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2014, 984-993.
15. Thaseen, S., Cherukuri, A. K., Kopparapu, A., & Velu, G. (2021). Cross-Border E-Commerce Security Issues and Protections. In M. Hoque & R. Bashaw (Eds.), *Cross-Border E-Commerce Marketing and Management* (pp. 98-119). IGI Global. <https://doi.org/10.4018/978-1-7998-5823-2.ch005>
16. Preethi, G., Cherukuri, A.K. (2018). Privacy Preserving Hu's Moments in Encrypted Domain. In: Abraham, A., Muhuri, P., Muda, A., Gandhi, N. (eds) *Intelligent Systems Design and Applications. ISDA 2017. Advances in Intelligent Systems and Computing*, vol 736. Springer, Cham. https://doi.org/10.1007/978-3-319-76348-4_32
17. Cramer, C., & Rauschnabel, P. A. (2019). How smart is your smart product? A proposed scale and preliminary empirical validation. *Journal of Product Innovation Management*, 36(2), 230-249. [10.1111/jpim.12451](https://doi.org/10.1111/jpim.12451)
18. "Smart healthcare: A review on its development and potential." (2017). *Journal of Healthcare Engineering*, 2017, 1-15. [doi:10.1155/2017/1271706](https://doi.org/10.1155/2017/1271706)
19. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, 2, 3. [doi:10.1186/2047-2501-2-3](https://doi.org/10.1186/2047-2501-2-3)
20. Li, S., Li, P., & Liang, H. (2020). Cloud computing in healthcare: A systematic review. *IEEE Access*, 8, 22337-22349. [doi:10.1109/ACCESS.2020.2962821](https://doi.org/10.1109/ACCESS.2020.2962821)
21. Gao, Y., Li, X., & Jiang, X. (2021). The design of a smart healthcare system based on the Internet of things. *Journal of Physics: Conference Series*, 1809(1), 012016. [doi:10.1088/1742-6596/1809/1/012016](https://doi.org/10.1088/1742-6596/1809/1/012016)
22. Khalid, S., Usman, M., & Yaqoob, I. (2019). Healthcare IoT devices: A survey on secure healthcare data transmission. *Journal of Ambient Intelligence and Humanized Computing*, 10(1), 73-86. [doi:10.1007/s12652-018-0918-x](https://doi.org/10.1007/s12652-018-0918-x)
23. Zhang, W., & Gao, F. (2022). A comparative study of privacy preservation methods for big data in healthcare. *Journal of Medical Systems*, 46(1), 1-15.
24. Gao, J., Liu, Y., Chen, S., & Chen, X. (2021). A survey on privacy-preserving data publishing techniques. *Journal of Big Data*, 8(1), 1-34.
25. Li, Y., Li, X., Sun, X., & Li, Q. (2021). An Improved t-Closeness Method for Privacy Protection of Data. *Journal of Computer Science and Technology*, 36(5), 1015-1029.
26. Soria-Comas, J., Domingo-Ferrer, J., & Martínez-Balleste, A. (2019). On the limitations of t-closeness and l-diversity in privacy-preserving data publishing. *Journal of Computer and System Sciences*, 105, 61-73.

27. Agrawal, R., & Srikant, R. (2022). Privacy-preserving data mining. Encyclopedia of Database Systems, 1-9.
28. GDPR.eu. (n.d.). What is GDPR? Retrieved March 24, 2023, from <https://gdpr.eu/what-is-gdpr/>
29. ARX. (n.d.). ARX - Anonymization Tool Configuration. <https://arx.deidentifier.org/anonymization-tool/configuration/>
30. Sadhya, D., & Chakraborty, B. (2017). Quantifying the effects of anonymization techniques over micro-databases. IEEE Transactions on Information Forensics and Security, 12(3), 588-602. doi:10.1109/TIFS.2016.2636979
31. Menon, S. C., Prasad, S., & Soni, G. (2023). Preserving Privacy of Patients With Disabilities in the Smart Healthcare Systems. In Using Multimedia Systems, Tools, and Technologies for Smart Healthcare Services (pp. 191-205). IGI Global
32. Wang, W., Li, X., Qiu, X., Zhang, X., Zhao, J., & Brusica, V. (2023). A privacy preserving framework for federated learning in smart healthcare systems. Information Processing & Management, 60(1), 103167.
33. Puri, G. D., & Haritha, D. (2023). Implementation of Big Data Privacy Preservation Technique for Electronic Health Records in Multivendor Environment. International Journal of Advanced Computer Science and Applications, 14(2).

¹ DEPARTMENT OF COMPUTER ENGINEERING, KOCAELI UNIVERSITY, TURKIYE.
Email address: ce.elanurozmen@gmail.com

² SCHOOL OF COMPUTER SCIENCE ENGINEERING AND INFORMATION SYSTEMS, VELLORE INSTITUTE OF TECHNOLOGY, INDIA.
Email address: cherukuri@acm.org

² SCHOOL OF COMPUTER SCIENCE AND ENGINEERING, VELLORE INSTITUTE OF TECHNOLOGY, INDIA.
Email address: jannapurna@gmail.com