Linnæus University
Sweden

Bachelor Degree Project

Data mining in healthcare

- A security and privacy perspective

Author: Sara Vimark
Supervisor: Diego Perez
Semester: VT 2023
Subject: Computer Science

Abstract

Data mining has become an essential tool in various domains, including healthcare, for finding patterns and relationships in large datasets to solve business issues. However, given the sensitivity of healthcare data, safeguarding confidentiality and privacy to protect patient information is highly prioritized. This literature review focuses on security and privacy methods used in data mining within the healthcare field. The study examines various techniques employed to secure and preserve the privacy of healthcare data and explores their applications. The review addresses research questions about security and privacy techniques in healthcare data mining and their specific use cases. By summarizing the current state of security and privacy methods, this review aims to contribute to the knowledge base of data mining in healthcare and provide insights for future research. The results show that anonymization, cryptography, blockchain, differential privacy, and randomization techniques are the most prevalent methods. However, more research is needed to provide sufficiently secure methods that still preserve the data's utility.

Keywords: data mining, healthcare, security, privacy

Contents

1 Introduction

Data mining is the process of finding patterns, relationships, and predictions of future trends in large datasets (or "big data") that can help solve business issues through data analysis [1]. Due to its ability to find patterns in large amounts of data that can be used in decision-making, data mining has become a critical tool in various domains. One of the domains where data mining has been growing in use for several years is healthcare [1]. With the rise of digitalization and innovative techniques, vast amounts of patient data can now accurately be stored in computers. Information mining, given the exponential growth of electronic health records, presents enormous potential for advancements in healthcare. Efficiently filtering through the abundance of data to find useful patterns and information is challenging for healthcare service providers, and this is where data mining can be of tremendous value [48]. However, when performing data mining, the confidentiality of the data can easily be threatened if it is not correctly handled, secured, and stored. In healthcare, the data handled is often sensitive, and the need to protect the data to ensure patient information privacy is crucial. Healthcare data is known to be an attractive target for cyberattacks. Therefore, preserving the privacy of that data as well as securing it against attacks, is of high priority [1].

In this thesis, a literature review on security and privacy when performing data mining in the healthcare field is performed. The focus will be on the methods used to secure and protect the privacy of the data involved in data mining and for what purposes the methods are being used.

1.1 Background

Many organizations have switched from paper-based to electronic systems over the last few decades to improve job productivity and results. Electronic solutions offer considerable advantages to employers and employees, and due to this shift, massive amounts of data are being gathered daily via these computerized devices. Data owners are discovering the valuable information concealed in their datasets, and the data collected in any business is now viewed as a new source of crucial information that can immediately impact that organization's operational efficiency, deliver higher-quality results, and even eliminate wasteful expenditures that squander the organization's resources [2].

Healthcare institutions and organizations actively collect electronic health data using various methods, such as computer-based surveys, online insurance claims, and Electronic Health Records (EHR) or Electronic Medical Records (EMR) systems. Hospitals, clinics, and other healthcare providers are gathering substantial amounts of data due to deploying EHR systems [2]. EHR are digitally saved health records that provide information about a person's health, and this information consists of various components, such as demographic information, prescriptions, diagnoses, vital signs, vaccines, results of laboratory and radiology tests, concepts, and comments related to medicine, procedures, and treatment plans. Each time a customer or patient enters a hospital or healthcare facility, this information is logged. Overall, EHR systems raise the standard of medical care by providing a large amount of data that can be mined and used for improving healthcare services, such as better diagnosing of patients, better treatments, tailored customer service, and predictive analysis [51].

When utilizing data mining on health data, it is vital to consider the security and privacy aspects of this sensitive data. Damage can easily be caused by, for example, improper disclosure or loss of data integrity. Recent laws and regulations, including the Health Insurance Portability and Accountability Act (HIPAA) and the Data Protection Regulation of Europe (EU GDPR), give consumers legal rights over their personally identifiable health information and impose protection and restriction requirements on healthcare organizations. To lessen the potential harm to patients, their organizations, or themselves, data miners should have a fundamental awareness of healthcare information privacy and security. Furthermore, keeping the data properly secured and preserving the privacy of the patient's data while still keeping the quality of the data is considered a big challenge in healthcare data mining. The methods that are used for privacy and security purposes in the field need to be tweaked with this challenge in mind [51].

1.2 Related work

Several studies have been conducted on security and privacy in data mining, including data mining in the healthcare sector. Following are some of the most relevant papers found when researching the topic.

An overview of privacy-preserving data mining techniques is given in [49], and it discusses the underlying principles, benefits, and shortcomings of these techniques. Additionally, the paper categorizes the available privacy-preserving data mining techniques and discusses their notable benefits and drawbacks. The authors have comprehensively reviewed privacy-preserving data mining in healthcare. However, the paper was published in 2015 and needs to be updated, given the rapid advances in data mining. They also conclude that ‘significant enhancements for more robust privacy protection and preservation are affirmed to be mandatory’ [49, p. 1].

S.J. Gabriel *et al.* [48] wrote a survey on privacy-preserving data mining in healthcare, listing some techniques for preserving privacy when using data mining in healthcare. The authors discuss data mining and utility mining and how they are used to identify useful patterns. It describes how data mining impacts healthcare and how privacy can be preserved using different techniques.

Another survey was published in 2018 by K. Abouelmehdi *et al.* [47] on preserving security and privacy in big healthcare data. They present some of the security and privacy-preserving methods that can be used to protect data in the healthcare domain when dealing with big data. The authors evaluate how security and privacy issues arise when dealing with large amounts of healthcare data and explore possible solutions. It primarily focuses on the most current anonymization and encryption-based approaches that have been suggested and contrast their advantages and disadvantages. They conclude that privacy and security issues greatly hinder researchers in this field.

In [50], the authors review the most relevant privacy-preserving data mining methods from the literature and the criteria by which they are measured. It is mostly focused on application areas of privacy-preserving data mining and current difficulties and unresolved problems, some of which are related to healthcare data mining.

The author of [51] discusses the privacy and security of healthcare data when it is used for data mining. They make recommendations for best practises when dealing with privacy and security in healthcare data mining. A literature analysis on technical difficulties when dealing with privacy guarantees, as well as a case study highlighting possible risks when data mining personally identifiable information is also presented in the paper.

A thorough literature review of state-of-the-art proposals to maintain security and privacy in Healthcare 4.0 was published in 2020. It explores a blockchain-based solution to give researchers and practitioners insights into the area, and existing challenges of security and privacy in Healthcare 4.0 are discussed. The authors focus is not on data mining, but they address several security and privacy-preserving methods that can also be applied to data mining in healthcare and their advantages and limitations [52].

The authors of [53] focus on privacy and security concerns in big data. This paper aims to provide a major review of the privacy preservation mechanisms in big data and present the challenges for existing mechanisms. They discuss that big data privacy is a major concern due to the data sets' complexity and size and discuss some recent privacy-preserving techniques in big data. These techniques are also relevant to healthcare data.

1.3 Problem formulation

Previous research brings up various methods to protect and preserve the privacy of the data extracted with data mining in healthcare. However, no previous recent study was found that summarizes the different techniques that can be used to protect and preserve healthcare data privacy when using data mining in this field and also looks into how they are utilized.

The research questions this thesis will try to answer are the following.

RQ1. What security techniques can be used to protect data when using data mining in the healthcare domain?

RQ2. What privacy-preserving techniques can be used when using data mining in the healthcare domain?

RQ3. In what scenarios related to data mining in healthcare are these techniques used?

1.4 Motivation,

The security and privacy aspects of data mining in healthcare are highly prioritized, given the sensitivity of the data handled in this field. A leakage of healthcare data could be catastrophic for the individual with their medical data in the wrong hands and used in ways they did not agree to [1]. The proposed literature review will contribute to the knowledge of data mining in the healthcare domain by presenting a summary of the current state of security and privacy methods when dealing with data mining in this field. This thesis aims to act as an information basis for future research in the field by providing an overview of the methods and their use cases.

1.5 Scope/Limitation

The subject of data mining has a broad application domain, and one of those domains is healthcare. Data mining in this domain is a relevant and growing topic for research, and there is a lot of previous research in this field. Due to constraints on time and resources, this literature review will focus only on the various methods used to secure data and the privacy of data when performing data mining in the healthcare field and the ways those methods are being used. Search words, a date range, and language restrictions are set to find the most relevant publications in the field of study. Moreover, the selection of papers to include is performed by only one researcher, which increases the chance of bias in the paper selection process and, in turn, the chance that relevant publications are missed.

1.6 Target group

The results of this literature review are interesting to researchers, practitioners, and policymakers in the field of data mining in the healthcare domain, including practitioners working with the security of the involved data. The findings of this study can work as a baseline for further research into healthcare data mining and the security and privacy methods currently in use in the field.

1.7 Outline

The remainder of this thesis has the following structure. In Chapter 2, the chosen research methodology is described and discussed, as well as ethical considerations and concerns regarding validity and reliability. Chapter 3 introduces the theoretical background of general and healthcare data mining to lay a foundation for the following chapters. Chapter 4 presents the literature review results performed as a part of this research. In Chapter 5, an analysis of the results from Chapter 4 is performed and presented as answers to the research questions. Chapter 6 discusses areas of interest and observations made while performing the research. Chapter 7 consists of a concluding summary of the conducted research and recommendations for future research.

2 Method

To find an answer to what methods are being used to protect and preserve privacy when it comes to data mining in the healthcare field, a systematic literature review (SLR) was used as a method. SLR as a method was chosen since it allows for using previous research to find the answers to the research questions. The guidelines defined by [6] were followed and done in a Planning, Conducting, and Reporting phase to perform the SLR. Each phase has stages that are defined in Figure 2.1. The Introduction chapter describes the Planning phase, while the following sections in this chapter describe the Conducting phase. The final reporting phase will be addressed in Chapter 4.

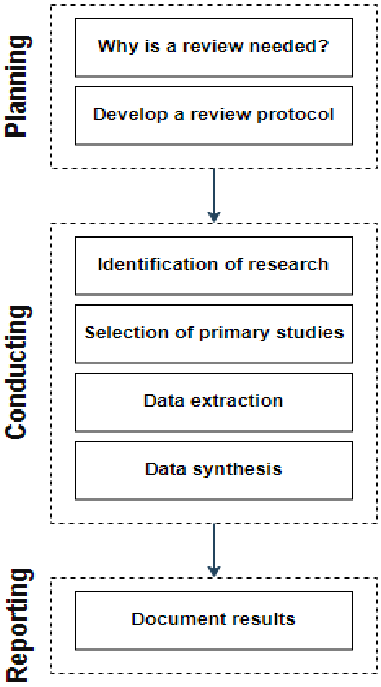

Figure 2.1: Systematic literature review phases, based on the phases in [55, Fig. 2]

2.1 Search strategy:

A search strategy was defined to find as many relevant studies on data mining security and privacy in healthcare as possible. The time and resource constraints limited the research to include sources that could be accessed free of charge or through a Linnaeus University login. Additionally, a date limit was set to exclude publications before 2013 to avoid getting results with outdated methods.

ACM Digital Library [7], IEEE Xplore [8], and Web of Science [9] were used as the databases to search in. Preliminary searches were done to extract the most suitable search words to find the relevant sources. Search words were initially used with abstracts and titles as filters, but the results were too broad, so the search strategy was narrowed, and a search string was used to search by pasting it into the basic search field in all the databases.

Regarding the search phrases, synonyms or alternative formulations were connected using the Boolean operator "OR," while other search terms were merged using the Boolean operator "AND."

The terms “data mining” OR “big data” OR KDD OR “knowledge discovery” was used to provide different synonyms for data mining, and the terms “health care” OR healthcare OR “health data” and health* were added and used in different combinations to find papers relevant to data mining in healthcare. Finally, security OR privacy* OR confidentiality OR protect* was added to find the papers related to the research topic.

The final search string that was used was as follows: ("data mining" OR "KDD") AND (healthcare or "health care" OR "health data") AND (security* OR privacy* OR protect* OR confidentiality).

The searches in the previously mentioned databases were performed on the 2nd of April 2023 and resulted in 1598 results on ACM Digital Library, 186 results on IEEE Xplore, and 133 results on Web of Science. All databases were filtered only to display publications between 2013 and 2023. The publications found during the research of related work in Chapter 1.2 were also included.

2.2 Study selection

The selection of studies to use was performed in several steps, and studies were excluded in each step if they had no relevance in answering the research question. The stages that were performed were the following:

1. In the three databases used, the results of a search were set by default to be sorted by relevance, and in the case of ACM Digital Library, where the results were the highest, a limit was set at 200 results. After looking at the keywords and titles of papers after the limit of 200 results, it was found that the number of publications relevant to the research questions was greatly diminished, and all publications after the limit of 200 were therefore excluded due to time and resource limitations.
   All search results were selected for the next step in the other two databases.
2. All papers were examined by looking at the title and keywords and excluded if they did not relate to the research topic.
3. If the publication was still deemed relevant, the introduction, abstract, and conclusion were read to see how they brought up security and privacy methods to evaluate their relevance to the research topic.
4. The remaining publications were fully examined and evaluated against the inclusion criteria.

Inclusion criteria:

- Addresses and explains at least one technique of protecting data that can be used in healthcare data mining.
- Addresses and explains at least one privacy-preserving technique that can be used when using data mining in healthcare.
- Was published between 2013 and 2023.

Exclusion criteria:

- Written in another language than English.
- Payment or additional login to access is needed.

2.3 Data extraction and synthesis

All the papers evaluated as relevant to the study were structured in a data extraction form suggested in [6]. The relevant information needed to answer the research questions was gathered in this form and was structured in the following way:

- Title of publication
- Publication year
- Author(s)
- Type of publication
- Research type performed
- Methods that are brought up
- In what way are they used

2.4 Reliability and Validity

With the previously described search strategy and selection method, it should be possible to reproduce the results that are presented in this review. However, since the databases are set to sort by relevance, there is a possibility that papers can be sorted in some other order for future searches. The order of the results can change due to papers being calculated as more relevant due to, for example, an increased number of citations which might lead to a different search result than presented in this paper.

Given the time and resource restrictions, the literature review follows some guidelines proposed in [6], but others are left out. For example, they suggest that at least two researchers perform the data extraction individually to compare their work for disagreements or issues. Since the study selection in this review was performed by a single researcher, this can affect the validity of the outcome due to bias being introduced by that researcher. Additionally, with the time and personnel limitations previously mentioned, there is a risk of not finding relevant studies due to a lack of a thorough search strategy and selection process.

2.5 Ethical Considerations

Regarding ethical considerations for this thesis, the sensitive nature of data collected in the healthcare field was contemplated in the selection of the papers, and none of the papers were found to contain any sensitive information. Since this work focuses on providing an overview of the security and privacy methods used for data mining of healthcare data, no focus is put on any individual's data, and it does not contain any information that breaks any ethical violations.

3 Theoretical Background

This chapter will bring up the theory needed to understand the topic of data mining in general and what special considerations health data impose on the security and privacy aspects of data mining.

3.1 Data Mining

As mentioned in the introduction chapter, data mining is the process of finding patterns in data that can be used to make decisions, increase the efficiency of services, and predict future trends that benefit organizations and their customers [1].

The two main categories are descriptive (unsupervised learning) and predictive (supervised learning) data mining techniques. Descriptive data mining seeks to find similarities and identify patterns and relationships. Clustering and association rule mining are the key methods in descriptive data mining. On the other hand, predictive data mining aims to create prediction rules as a model to categorize the records based on a particular target (or label). The most popular method in predictive data mining is classification [4].

The knowledge is mined from the datasets using algorithms tailored to finding different patterns. However, before any algorithm can be applied, the data must be collected, cleaned, and transformed into suitable formats. If these steps are not properly performed, the algorithm will not perform well and have a bad accuracy of finding patterns in the data.

The data mining techniques commonly used for data mining are association rule mining, classification, and clustering [2].

3.1.1 Association rule mining

This is a technique that looks for patterns of associations of the attributes in the dataset that has a higher chance of occurring. When it is necessary to determine the link between attributes in a dataset, such as the association between the products in a customer's basket that were purchased, this technique is utilized. For example, this technique could be used in healthcare to find connections between different diseases [2].

3.1.2 Classification

Records that have already been classified or categorized are included in many datasets. Classification is a type of supervised learning where algorithms are taught patterns between the values of attributes that belong to a class and those that do not as a set of rules or a model. This model can forecast or assume the class value of new, similar records that do not yet have a class value and explain the relationships within the data. Common classification techniques are decision trees, neural networks, K-nearest neighbors, support vector machines, and Bayesian methods [4]. Examples of how classification can be used in healthcare are to predict whether a patient will have a specific health condition and the cost of different healthcare services [2].

3.1.3 Clustering

Not all datasets provide records with specified class values. In this case, learning about the data must occur without the supervision of class values, which is referred to as unsupervised learning. Clustering examines the data to find groups of records with a similar structure but different from other records and clusters [2]. This technique can be used when the information about the different types of data objects involved in a population is limited [4]. In healthcare, it can, for example, be used to find similar variants of viruses or infectious diseases.

3.2 Data Mining in Healthcare

When dealing with data mining in healthcare, some challenges must be considered. One of those challenges, the balance between safety and data quality, is highly related to the security and protection measures enforced on healthcare data for mining purposes.

The data handled in healthcare comes from various sources such as personal medical records, clinical trial data, sensor readings, and 3D imaging and varies in complexity and quality. Biases, noise, and irregularities are frequently present in healthcare data, which might hinder performing appropriate data mining, and in succession, decision-making and patient care are negatively affected. High-quality data ensures better accuracy in the mined information while reducing the analysis cost [5]. Therefore, data cleaning is highly recommended before performing any analysis that would be used in making decisions that, for the patient, could potentially be life-threatening. When cleaning the data, it is also common to take steps to safeguard the privacy of the data, which could lower the dataset's quality. For this reason, finding the right balance between keeping patient information private and preserving the accuracy and usability of the data presents one of the most significant challenges related to security and privacy issues [5], [39]. Using security and privacy techniques to solve this problem can make data mining in hospitals valuable for gaining accurate health information [3].

Sharing available data across several hospitals or other parties is another significant challenge due to privacy and legal concerns. Hospital compliance with privacy regulations often restricts sharing of health information about patients with other parties. Similarly, when data is distributed over patients' personal devices like mobile phones or wearable devices, it is not always feasible to assume that all sources holding part of the data can share their information with a trusted third party because the privacy of the data may not be protected from that party [3].

4 Results

The search for relevant literature was performed according to the search strategy that was detailed in Chapter 2.1 and the results are illustrated in Table 4.1. The steps of the study selection described in Chapter 2.2 and the number of publications they resulted in are depicted in Table 4.2.

Table 4.1 Results per searched database

Database	Number of results	Results used
ACM Digital Library	1598 results	200 first results
IEEE Xplore	186 results	186 (all results)
Web of Science	133 results	133 (all results)
		Total: 519 results

During the selection process the steps in Chapter 2.2 were followed, and Table 4.2 shows the number of papers that were selected in each step including the papers that were found when researching related work for Chapter 1.2.

Table 4.2 Number of papers selected in each step of the study selection

Selection step	Number of papers
Screening of title and keywords	70 papers
Screening of abstract, introduction and conclusion	64 papers
Final screening of full paper	37 papers
Related work	7 papers
	Total:44 results

The search strategy and selection process resulted in 37 studies that were found to have relevance to the problem that was defined in Chapter 1.3. The 7 studies found in the research for related work were also included resulting in a final selection of 44 papers. The included studies were published between 2013 and 2023, with an increase in relevant papers published in the last five years, indicating that security and privacy aspects in data mining in healthcare are a growing field of interest. The distribution of publications per year is illustrated in Figure 4.1.

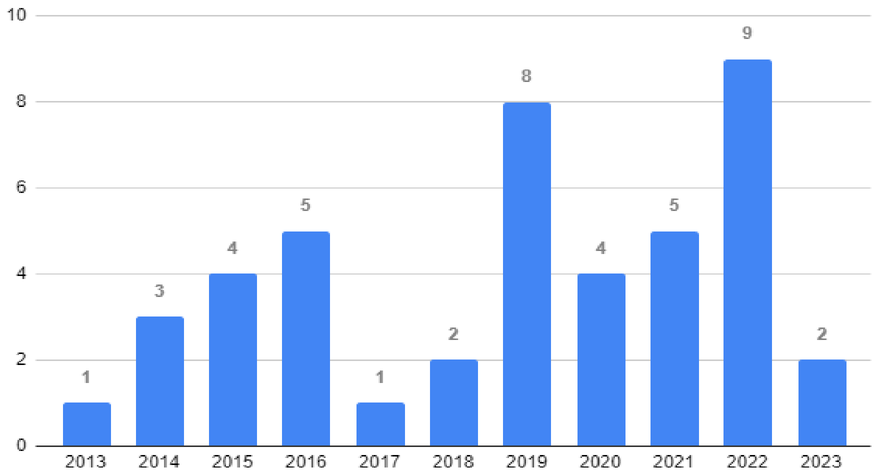

Figure 4.1 Distribution of publications per year

The final selection of relevant papers is conference papers and journal articles. The most common research type for the papers is design science, where a new method of privacy-preserving or security for healthcare data is proposed. The papers' research types are depicted in Figure 4.2, and the publication types are in Table 4.3.

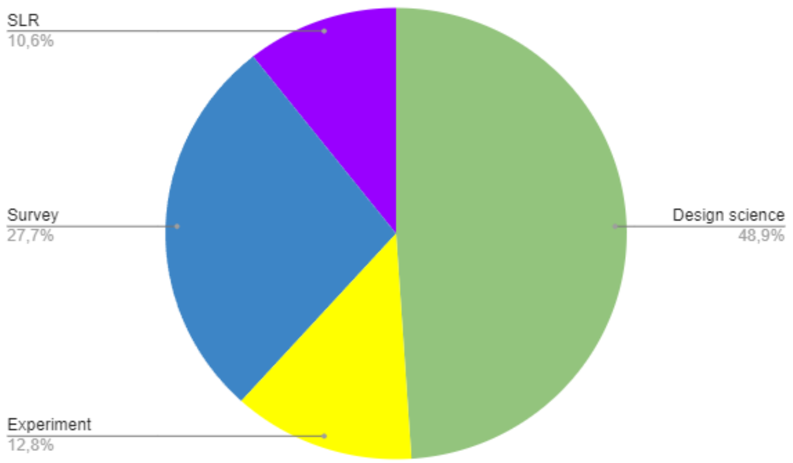

Figure 4.2 Distribution of research types

Table 4.3 Distribution of publication types

Type of publication	**Number of papers**
Conference paper	18
Journal article	26

The following methods and use cases were found to be the ones most commonly used in the current efforts in data mining of healthcare data to protect and privacy-preserve the data.

Anonymization is used to remove identifiers of the data so it is kept private. This needs to be done while also keeping the data quality high, which was mentioned in Chapter 1.1 as one of the biggest challenges when mining health data. This technique is used to preserve the privacy of medical data and to protect data privacy when sharing and storing data.

Cryptography is the technique of encrypting and decrypting data and is used to ensure access control, secure transactional data, protection of data privacy, secure cloud computing, and storage.

Blockchain is mentioned as the up-and-coming method of security and privacy for sharing medical records, safe collecting of mobile health data, access control, secure storage of data, and security in cloud and fog computing.

Differential privacy works by adding noise to the data to protect its privacy and is used when sharing and storing data. As with anonymization, there is a trade-off between keeping the data safe and its utility when using differential privacy.

Randomization is a method where the data is modified by injecting noise with a known statistical distribution so that data mining methods can recreate the original data distribution but not the actual individual values. This method is used to protect data privacy when sharing and storing it.

Federated learning is the process of creating machine learning models across datasets spread across several data centers, including hospitals, clinical research labs, and mobile devices, while preventing data leakage. This approach protects data privacy by only providing mathematical parameters and information, keeping the actual data as secure as possible, and avoiding attacks and tracebacks. The number of papers that mentioned each method is shown in Table 4.4, and the methods will be discussed in more detail in Chapter 5.

Table 4.4 Privacy and security methods in healthcare data mining

Method	**Number of papers**
Anonymization	23
Cryptography	23
Blockchain	11
Differential privacy	8
Randomization	7
Federated learning	2

5 Analysis

In this chapter, an analysis of the SLR findings will be presented as answers to the research questions.

5.1 Methods to secure and privacy-preserve data used for healthcare data mining (RQ1, RQ2)

In the following subsections an analysis of the methods mentioned in the previous chapter is presented. The analysis follows the same order as they are presented in the results.

5.1.1 Anonymization

Data anonymization is a means of protecting data subjects' privacy by providing altered versions of data that prevent re-identification [17], [18]. The most current and frequent anonymization techniques that were identified during the research of privacy techniques were different variations of k-anonymity, l-diversity, perturbation, and t-closeness [54]. Two frequently mentioned methods to achieve anonymization are suppression and generalization [11], [17], [18], [22], [39], [41], [47], [48], [49], [50], [52] where suppression means the substitution or concealment of quasi-identifiers [18] while with generalization, 'quasi-identifiers are replaced by more general values from higher levels of the hierarchy' [18, p. 2]. Quasi-identifiers are attributes that, on their own, can usually not be used to identify an individual's information but might do so if combined with other attributes [18]. There is always a trade-off between data privacy and the utility of the data when using these techniques [17], [22], [26], [35], [37], [39], [40, [31], [43] and anonymized data can still be vulnerable to correlation attacks, even when obvious personal identification information such as IP addresses and usernames are masked [29]. Increasing privacy generally involves distorting the data, which damages its representativeness of real-world phenomena [17], [22], [26], [35], [37], [39], [40], [41], [43].

The k-anonymity model is commonly used to ensure data anonymization [11]. It has been shown to protect data from disclosure and prevent individuals from being connected with their health records [11], [47]. For achieving k-anonymity, a dataset must contain at least k-1 more entries that match any tuple with provided attributes [18], [19]. K-anonymity modifies data before submitting it for data analytics to prevent deidentification, resulting in K indistinguishable records [18], [19], [35]. Still, it has several areas for improvement, including inadequate protection for attribute disclosure, assuming that each record represents a unique individual, and lack of diversity needed for sensitive attributes [11], [50]. Variations of k-anonymity, such as complete k-anonymity, privacy-constrained anonymity, (l-k)-anonymity, (k-1)-anonymity, and (k-k)-anonymity, have been proposed by researchers to handle the limitations[11], [26], [43]. These models focus on preserving the smallest size of K groups, the sensitive attributes of each group, and the sensitivity level of each attribute value [39]. Each variation has its strengths and drawbacks, and data owners must choose an appropriate privacy-preserving level to ensure maximum privacy while minimizing information loss [11], [17], [37], [39], [40].

L-diversity is considered an extension of k-anonymity [11], [35], [40], [43], [47], [49], [50], and the method seeks to diversify sensitive data attributes by ensuring that each quasi-identifier equivalence class has at least L distinct sensitive attribute values [18], [19], [35], [39], [40], [50]. The L-diversity method is a form of group-based anonymization that is utilized for safeguarding privacy in data sets by diminishing the granularity of data representation [47], and it may require injecting fictitious data, which increases security but can present problems during analysis [18], [47], [52]. Even though this method handles some of the weaknesses of k-anonymity, it has weaknesses of its own and is not enough to ensure protection against attribute disclosure [47].

T-closeness is a technique that extends and improves l-diversity [11], [35], [47], [52]. It ensures that the distribution of given sensitive attributes does not deviate from the true sensitive attribute by more than t distance [11], [18], [19], [31], [35]. This technique is particularly beneficial in intercepting attribute disclosures, but as data volume increases and the varieties of information increase, so do the chances for re-identification [18], [47].

Perturbing data before sharing it could be a solution to address privacy concerns [17], but perturbation-based solutions have limitations in satisfying data privacy and data utility requirements [17], [33]. Data utility can decrease if the perturbation is not precisely controlled, and privacy will not be preserved if the perturbation is not sufficient [17], [43]. The random perturbation technique randomly modifies original data to protect privacy [32], [43]. It distorts sensitive attribute values while preserving underlying distribution information [32]. While effective at preserving privacy, perturbation techniques can also cause a loss of implicit information available in multi-dimensional records due to the treatment of each attribute independently [17], [33]. One form of random perturbation is the data-swapping algorithm [32]. It retains information while randomly swapping data values to protect sensitive information without disturbing non-sensitive attributes [32].Slicing is another technique that perturbs data and effectively preserves privacy while maintaining data utility [33]. These techniques modify the original data by adding noise to preserve the sensitive attributes' privacy while maintaining the data's statistical properties [39], [43], [48], [50]. Researchers have found that perturbation-based techniques maintain the accuracy of the dataset while preserving privacy [43]. However, since perturbation only alters the distributions of the data and not the actual values, new distribution-based mining algorithms must be developed for each data problem, such as clustering, classification, or association rule mining. While some distribution-based mining algorithms exist, using distributions instead of actual records limits the range of algorithmic processes that can be used on the data [48]. Other privacy-preserving techniques, such as cryptographic methods, are being developed to overcome the limitations of perturbation-based mining algorithms [48].

Some other techniques of anonymization are mentioned in the research. One is swapping, where an attribute is swapped with another data point, which has been shown to protect individual privacy while minimizing data loss effectively [22], [35], [41], [47] but can still lead to excessive data distortion [22]. Masking is another technique where sensitive information is replaced with an unidentifiable value [47], [49]. It involves de-identifying data sets by masking personal identifiers like names and social security numbers and suppressing or generalizing quasi-identifiers like date-of-birth and zip codes [29], [47]. A significant benefit of data masking is that it reduces the cost of securing big data deployments [47]. These techniques are not broadly discussed in the research but are more briefly touched upon regarding the other techniques. Figure 5.1 shows the distribution of how many studies mentioned each anonymization method.

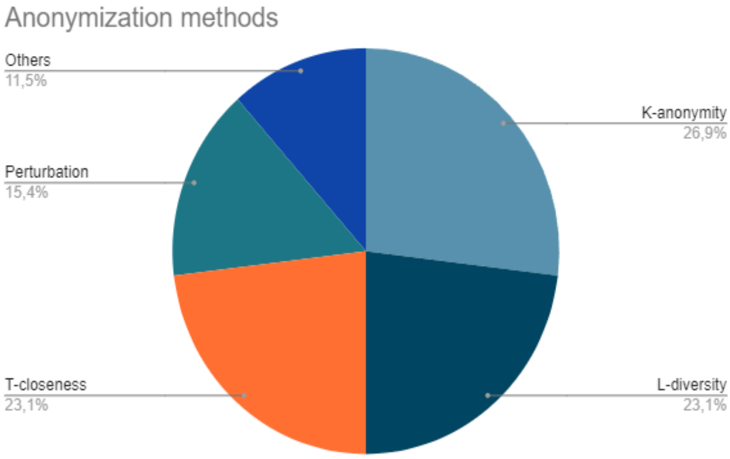

Figure 5.1 Distribution of anonymization methods

5.1.2 Cryptography

Cryptography is a powerful tool used to safeguard the privacy of data mining [16], [23], [26], [32], [47], [53], particularly in situations where sensitive information needs to be shared between multiple parties who do not necessarily trust each other [12], [32], [39], [48]. Data encryption transforms original patient data into an encoded pattern that requires decryption to access [13], [34].

To ensure privacy and security when data mining large amounts of healthcare data, various cryptographic techniques can be employed [16], [26] at different levels, including packet, cluster, and field levels [13]. Homomorphic encryption and Secret sharing are two commonly used techniques [14], [15], [16], [23], [24], [37], [38]. Homomorphic encryption enables computations on encrypted data without having to decrypt it first [16], [24], [37], [38], [44], [50], while Secret sharing involves transmitting secrets to different parties and merging individual shares to reconstruct the secret [38]. One form of homomorphic cryptosystem used to encrypt medical data is the Paillier cryptosystem [14], [24], [44]. The Paillier Cryptosystem is provably secure and consists of three algorithms: key generation, encryption, and decryption [14]. The homomorphic encryption algorithm used is an additive one designed by Pascal Paillier, allowing only addition operations on ciphertext [44]. Operations over encrypted data are required to protect individual privacy in big data analytics [53], but these operations are often complex, time-consuming, and inefficient for high-volume big data mining [17], [53] due to the significant communication and computation costs involved [17].

To perform data mining over distributed data in multiple parties while preserving privacy, secure multi-party computation (SMC) can be used [17], [38], [50]. SMC allows computing intermediate results without revealing the secret values required for the computation [17]. A basic building block of most SMC techniques is the 1-out-of-2 oblivious-transfer protocol, which allows a receiver to learn one out of two inputs/messages given by a sender without the sender learning anything [50]. The protocol relies on secure data transfer and privacy-preserving distributed computation [50]. The trade-off with SMC techniques is that they can result in increased communication and computation overheads, leading to efficiency issues [17], [38].

5.1.3 Blockchain

Blockchain is a technology that can create a more secure, decentralized, and safer environment for exchanging electronic health record (EHR) data [10], [25], [29], [34], [46], [52]. It provides a unique method for fully distributed bookkeeping, effectively used to overcome centralized issues [10]. EHRs are generally shared among healthcare stakeholders, making them susceptible to data misuse, a lack of privacy, security, and an audit trail [25]. However, blockchain provides a distributed and decentralized environment where nodes in a list of networks can connect without the need for a central authority [10], [25], [52]. Transactions are validated by a network of peer nodes using algorithms and, if verified, are added to an immutable distributed ledger [10], [34], [46], [52]. This technology improves the efficiency, speed, and traceability of transactions, making it a secure and safe option for executing all kinds of transactions [34], [36], [46], [52].

Through blockchain, healthcare organizations can record and secure patient data more accurately and efficiently than with other technologies currently available [18], [46]. Medical data mining vulnerabilities can be mitigated by incorporating blockchain technology into the fog paradigm, which increases data access methods, accountability, and authentication [46]. Furthermore, the security and confidentiality of health information can be guaranteed through blockchain technology [18], [29], [34], [36], [46], [52], which allows healthcare organizations to benefit from the decentralized data security without facing the associated risks [46], [52].

There are still challenges in adopting blockchain-based healthcare systems, including technological, organizational, cultural, ethical, and legal issues, as well as psychological and cultural challenges [29]. Processing big data with high volume, velocity, variety, variability, and veracity is also challenging [29], [46]. Overall, blockchain technology has the potential to provide a secure and efficient healthcare system by facilitating healthcare interoperability, patient ownership and control of electronic health records, and secure record transfers [18], [29], [34], [36], [46], [52]. The use of blockchain as a secure, decentralized form of communication is on the rise, and the technology can assist in ensuring security in healthcare settings [30][36].

5.1.4 Differential privacy

This method allows researchers and analysts to extract useful information from statistical databases containing personal information while still protecting the privacy of individuals [2],[40], [44]. This is accomplished by introducing a minimal amount of distortion to the data provided by the database [44]. This method aims to solve issues with other privacy-preserving techniques by introducing an intermediary software, or privacy guard, between the database and the analyst to ensure that the privacy of individuals is protected while still allowing for useful information to be extracted from the database [44].

Methods that provide differential privacy add noise to the dataset to preserve privacy while sharing the data [2], [8], [26], [40], [44]. Differential algorithms add noise to quasi-identifying information like zip code, gender, and birthday to prevent possible identification of an individual's data [2],[10]. Laplace and Gaussian statistical distribution methods of differential privacy are used to obscure identifying information with random noise, ensuring privacy [10]. When combined with k-anonymity, the Laplace method proves to be more effective in preserving privacy [2].

Differential privacy is often used to guarantee the desired privacy level for a given purpose and is increasingly used for privacy-preserving data mining [26]. Noise addition adds a significant amount of uncertainty to the data points within a cluster, ensuring differential privacy for each added data point while preserving privacy for the resulting clusters [26]. The method has significant potential for protecting privacy while still keeping the information useful for data mining [10], [17],[44].

5.1.5 Randomization

This technique involves using data randomization by jumbling up the data so that receivers cannot determine the actual probabilities beyond a certain threshold [17], [30], [32], [35], [39], [48]. This can be done using methods such as adding noise with a known statistical distribution or applying multiplicative noise [50]. Randomization is considered a subset of perturbation operations, and other methods besides additive and multiplicative noise can also be used at different phases of data modification [50].

The randomization method was first proposed for solving survey problems [32], [39], where data collected from individuals is aggregated to obtain more accurate results. This approach is cost-effective and is commonly used in surveys that involve sensitive data [39], and is effective in combination with reconstruction for categorical attributes [49].

While this technique effectively maintains privacy, it introduces ambiguity in the data, making it difficult for receivers to determine whether it is accurate or false [39]. The randomization technique also has other weaknesses; for example, it treats all records equally, regardless of their density, making exception records more vulnerable to attacks [48]. It is also possible to estimate the original values using noise removal techniques, thereby weakening the effectiveness of such methods in providing strong privacy guarantees [17]. It is also potentially unsuitable for multiple sensitive attribute databases [39]. Despite its weaknesses, randomization is still considered a simple yet effective method for preserving individual privacy [49].

5.1.6 Federated learning

This security method is a machine learning approach that enables the development of models using distributed datasets from various locations without the risk of data leakage [17], [38]. In the healthcare sector, where Electronic Health Records (EHR) are distributed across hospitals and clinical research labs, federated learning has emerged as a promising technique for training models without compromising data privacy [38].

With federated learning, machine learning models can be trained on localized data with only model parameters temporarily kept on centralized servers, which facilitates not having to share the data over different devices [38]. This has, however, been proven not to be enough protection against some attacks, and researchers have experimented with using differential privacy and encryption in conjunction with federated learning to overcome some of the weaknesses [17] but adding other techniques to increase security can lower the performance of the machine learning models [17]. Therefore, more research is needed to find safe enough techniques to maintain the quality of the data while providing high security.

5.1.7 Suitability on classification, clustering, and association rule mining

The methods discussed above have different suitability for the data mining techniques described in Chapter 3.1. The suitability can depend on the techniques used in conjunction with each other and the degree of decrease in utility they impose on the data. Table 5.1 shows how well-suited each discussed method is to the data mining techniques mentioned in Chapter 3.1. Blockchain has been left out since no discussion regarding these techniques was found in the reviewed papers that covered blockchain.

Table 5.1 Privacy and security methods suitability to each data mining technique

Method	Classification	Association rule mining	Clustering
Anonymization	It is well-suited, but the utility can decrease.	It is well-suited, but the utility can decrease.	It is well-suited, but the utility can decrease.
Cryptography	It can be problematic due to computation and communication overhead but usually works well with homomorphic encryption.	It is well-suited with homomorphic encryption but can be impractical with other encryption techniques.	It is well-suited with homomorphic encryption but can be impractical with other encryption techniques.
Differential privacy	It is well-suited when used together with federated learning. It can be problematic due to the trade-off between privacy and data utility.	Not mentioned in the reviewed papers.	It is well-suited together with other noise injection and data abstraction techniques.
Randomization	It is well-suited, but the utility can decrease.	It can be used, but the effect is not discussed in the reviewed papers.	It is well-suited, but the utility can decrease.
Federated learning	It is well-suited.	Not mentioned in the reviewed papers.	Not mentioned in the reviewed papers.

5.2 Use cases of the identified methods (RQ3)

In the following subsections an analysis of use cases of the identified methods is presented in the same order that they are presented in previous subsections.

5.2.1 Use cases of anonymization techniques

Anonymization restricts attackers' capability to derive individuals' records from a dataset [11]. In healthcare data privacy, various anonymization techniques have been proposed to mitigate linkage attacks [19], [40]. A linkage attack is when the re-identification of data can be done from an anonymized dataset by combining quasi-identifiers with background information of data and, through that, discovering identifying connections [32]. Several k-anonymity, L-diversity, and perturbation models have been suggested to be used on healthcare datasets to address linkage issues [19].

Another scenario where anonymization is used is when healthcare providers, hospitals, or drug stores share their data [22], [43], [50]. They can benefit greatly from data mining techniques to analyze their data, but since data mining can be challenging for them to perform on their own, they may choose to send their data to a third party for analysis. This can lead to privacy breaches, as sensitive information about patients' diseases or medicinal use can be inferred by the data miner. Here anonymization can protect individual privacy while maintaining data utility before sending the data to an external party [22].

Finally, data anonymization techniques are used to protect the data that is distributed and stored in cloud settings [37].

5.2.2 Use cases of cryptography techniques

The rise of IoT devices collecting private information and disseminating data among multiple parties in today's data-driven era poses a security threat. Neither the communication network nor the third party can be considered completely secure or reliable [3]. To ensure the authenticity and integrity of sensitive data, a mechanism is required for data exchange between senders and recipients, and this is where cryptography can be used for securing multipart transactions in healthcare data mining [12], [23], [24], [32], [38], [39], [47], [48]. Crypto algorithms are also used to effectively hide large amounts of patient data and medical images by transforming the original patient data into encoded patterns that require decryption to access [13], [34]. Moreover, with the help of homomorphic cryptosystems and SMC, medical data can be encrypted and used for data mining on databases, distributed computing environments, and cloud servers and shared between multiple parties without compromising the privacy of the data [14], [16], [21], [32], [237], [38], [39], [47], [48], [50], [53].

5.2.3 Use cases of blockchain techniques

Blockchain (BC) can be used to overcome the security and privacy risks of storing medical records in healthcare cloud and application (HCA) servers and for effective and transparent sharing of EHRs [10], [18], [25]. BC addresses the security and trust-based concerns of shared EHR and personalized health records (PHR) through the integration of BC at HCA nodes, which maintains a trusted and chronological ledger among different stakeholders [10], [29]. Using BC, a distributed ledger is created among all stakeholders. This ledger is chronological, auditable, and timestamped, ensuring the integrity of the recorded transactions. One of the key benefits of BC is the ability to hold fake or incorrect transactional updates accountable. Any such updates can be traced back to the specific owner node responsible for making the unauthorized change

[8]. BC can also secure other types of storing and private transactions of medical data [18], [29], [34], [36], [45], and secure data management for mobile healthcare and IoT [28], [52].

5.2.4 Use cases of differential privacy techniques

Differential privacy aims to protect private statistical databases and increase the accuracy of queries from statistical databases while reducing the risk of identifying individual records [11], [49]. Several differential privacy models have been proposed to protect patient information in databases and achieve good utility when executing queries [11], [53]. Differential privacy can also be used to preserve individual privacy to be able to safely share data [17], and some differential privacy models are aimed at improving the security and privacy of healthcare data vulnerable to attacks such as the similarity attack and have been proven effective in protecting the data from such attacks [19]. Another area where differential privacy has proven useful is in securing collaborative filtering, a technique recommender systems use [26]. Furthermore, it can be used to determine how sensitive a dataset is given small changes in its composition, and it can be used to cluster sensitive attributes into separate "buckets" to disallow linkage attacks between members of different buckets [35].

5.2.5 Use cases of randomization

Randomization can be used to provide security and privacy of healthcare data in scenarios where sharing of the data is needed [17], [48] and to ensure privacy while providing utility for association rule mining on categorical and binary datasets [39], [49]. Random projection is a randomization technique that is a powerful method for dimensionality reduction in data mining. This technique projects the original data into a random low-dimensional subspace to generate results comparable to conventional dimensionality reduction techniques [32]. Randomization remains useful as it can be used as a standalone method or as an augmentation to other strategies [35].

5.2.6 Use cases of federated learning techniques

Federated learning has emerged as a collaborative approach for training machine learning models, wherein one party orchestrates the process while maintaining decentralized training data. In conjunction with differential privacy, Federated learning can be useful for preserving privacy in distributed data mining [17]. It can also preserve privacy when data mining is used on distributed client devices without a central server [38].

6 Discussion

This thesis aims to examine what methods are the most prevalent ones when it comes to securing and privacy-preserving data when performing data mining actions in the healthcare field. The findings of this research confirm the challenges discussed in Chapter 3.2 and extend upon the information gathered in related work. The evidence shows that security in this field presents a big challenge for healthcare organizations. It can often be a bottleneck in getting useful information from healthcare data with the help of data mining.

The different variations of anonymization are the methods that seem to be most frequently used, and that is most likely due to ease of use and the fact that to be able to adhere to the regulations of healthcare data, a first step is to anonymize the data to protect it from disclosure. However, even when data mining is done on data that is only shared between trusted parties, anonymization of data is not enough to ensure privacy. Because no single method can assure sufficient protection against attacks or deidentification of data, a combination of methods has to be used to live up to the regulations in place to ensure proper protection of healthcare data. This is supported by the authors of [35], who argue that combining different privacy-preserving data mining (PPDM) methods can effectively combat attack vulnerabilities and enhance privacy preservation. The combination of multiple security designs can potentially counter attackers from learning the original data, and some combinations of PPDM methods have been proven to work very well together, such as differential privacy and federated learning or anonymization.

Cryptography is the second method most often used, with homomorphic variations seemingly the most popular given that they provide the ability to use the data for data mining without decrypting it first. As with anonymization, cryptography is not enough to use independently and is often combined with other methods.

Blockchain and federated learning are the two newest methods. The research found on these techniques is not enough to draw any general conclusions. However, research is constantly ongoing to find better variations of these techniques that provide higher security and privacy while still being efficient and keeping the data accurate enough for valuable data mining. Much research is also ongoing in cloud storage and cloud computing, and constant efforts are being made to find better security and privacy methods to make these environments secure to use with health data. With the rising use of different kinds of wearable devices and IoT devices, there is a need to securely store and share the health data collected from these devices, and blockchain and cloud storage show promising signs of being a viable solution.

The methods discussed in this work have their strengths and weaknesses. It is up to the data owners to properly secure their data and find the most appropriate privacy-preserving methods to ensure a high amount of privacy while still keeping the data useful for data mining.

7 Conclusion and Future Work

The first objective of this thesis is to investigate the methods currently being used to secure and privacy-preserve data used for data mining efforts in healthcare. The second objective is to look into for what purposes the methods are being used to secure and privacy-preserve the data. A systematic literature review is used to find previous research on this subject and potential gaps within the current research. This research shows that the field of privacy-preserving techniques in data mining and healthcare data has seen significant recent advancements with Anonymization, Cryptography, Randomization, Blockchain, Differential privacy, and Federated learning found to be the most currently used methods for this purpose, and the methods, their variations, and use cases are discussed. The results also show that sharing healthcare data between different stakeholders and protecting data privacy while keeping a high utility of it are major challenges.

Privacy-preserving techniques in data mining and healthcare data offer various methods to secure individuals' privacy while enabling data analysis and information extraction. Each technique has its strengths and limitations, and the choice of technique depends on the specific privacy requirements, data characteristics, and trade-offs between privacy and utility. Ongoing research and advancements are necessary to address challenges and develop more effective and efficient privacy-preserving methods.

A more detailed review of the methods discussed in this thesis and their variations could be done for future research. Especially the fields of blockchain and federated learning have much room for further research with the need for more efficient ways to secure and utilize healthcare data.

It could also be interesting to investigate how a hospital or other healthcare provider is utilizing these methods in practice or to test any of them with real or fictitious healthcare data to evaluate their efficiency in protecting the data while still keeping it useful.

Another area that is not in the scope of this research but highly connected to it and needs more research is storing the huge amounts of healthcare data collected every day and the scalability and security of the available options.

References

[1] L. Coventry, and D. Branley, "Cybersecurity in healthcare: A narrative review of trends, threats and ways forward," Maturitas, vol. 113, no. 20, pp. 48-52, Jul. 2018. [Online]. Available: https://doi.org/10.1016/j.maturitas.2018.04.008

[2] M. H. Tekieh and B. Raahemi, “Importance of data mining in healthcare: A survey,” in 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, Aug 2015, pp. 1057-1062. [Online]. Available: 10.1145/2808797.2809367

[3] T. Sarwar *et al*., “The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges.” ACM Comput. Surv. 55, 2, Article 33, Feb. 2023. [Online]. Available: https://doi-org.proxy.lnu.se/10.1145/3490234

[4] D. Yates, and Md. Z. Islam. “Data Mining on Smartphones: An Introduction and Survey,” ACM Comput. Surv. 55, 5, Article 101, 2022. [Online]. Available: https://doi-org.proxy.lnu.se/10.1145/3529753

[5] R. Fang, S. Pouyanfar, Y. Yang, S-C. Chen, and S. S. Iyengar, “Computational Health Informatics in the Big Data Age: A Survey,” ACM Comput. Surv. 49, 1, Article 12, 2016 [Online]. Available: https://doi-org.proxy.lnu.se/10.1145/2932707

[6] B. Kitchenham, "Procedures for Performing Systematic Reviews," Keele University, UK, Technical Report TR/SE-0401, ISSN:1353-7776, July 2004. [Online]. Available: https://www.researchgate.net/publication/228756057_Procedures_for_Performing_Systematic_Reviews

[7] ACM Digital Library. [Online]. Available: https://dl.acm.org

[8] IEEE Xplore. [Online]. Available: https://ieeexplore.ieee.org

[9] Web of Science. [Online]. Available: https://www.webofscience.com

[10] H. Ghayvat *et al.*, "CP-BDHCA: Blockchain-Based Confidentiality-Privacy Preserving Big Data Scheme for Healthcare Clouds and Applications," in *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 1937-1948, May 2022, doi: 10.1109/JBHI.2021.3097237.

[11] L. A. Abuwardih, W. Shatnawi, and A. Aleroud, "Privacy preserving data mining on published data in healthcare: A survey," *2016 7th International Conference on Computer Science and Information Technology (CSIT)*, Amman, Jordan, 2016, pp. 1-6. doi: 10.1109/CSIT.2016.7549444.

[12] F. Zhu, X. Yi, A. Abuadbba, I. Khalil, S. Nepal, and X. Huang, "Cost-Effective Authenticated Data Redaction With Privacy Protection in IoT," in IEEE Internet of Things Journal, vol. 8, no. 14, pp. 11678-11689, July 2021, doi: 10.1109/JIOT.2021.3059570.

[13] N. R. K. L, M. Nithya, S. R, and S. Samundeswari, "NS's Secure Framework for Healthcare," *2022 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2022, pp. 1-7, doi: 10.1109/ICCCI54379.2022.9740978.

[14] G. Wang, R. Lu, and C. Huang, "PSLP: Privacy-preserving single-layer perceptron learning for e-Healthcare," *2015 10th International Conference on Information, Communications and Signal Processing (ICICS)*, Singapore, 2015, pp. 1-5. doi: 10.1109/ICICS.2015.7459925.

[15] X. Guo, H. Lin, C. Xu, and W. Lin, "A Data Clustering Strategy for Enhancing Mutual Privacy in Healthcare System of IoT," *2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking, and Services (SmartCNS)*, Shenyang, China, 2019, pp. 521-526, doi: 10.1109/IUCC/DSCI/SmartCNS.2019.00112.

[16] S. Sathya and T. Sethukarasi, "Efficient privacy preservation technique for healthcare records using big data," *2016 International Conference on Information Communication and Embedded Systems (ICICES)*, Chennai, India, 2016, pp. 1-6, doi: 10.1109/ICICES.2016.7518878.

[17] A. Aminifar, M. Shokri, F. Rabbi, V. K. I. Pun, and Y. Lamo, "Extremely Randomized Trees With Privacy Preservation for Distributed Structured Health Data," in *IEEE Access*, vol. 10, pp. 6010-6027, 2022, doi: 10.1109/ACCESS.2022.3141709.

[18] A. Vaghela and A. Suthar, "Comprehensive Analysis of Privacy and Data Mining Techniques," *2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA*, Pune, India, 2022, pp. 1-6, doi: 10.1109/ICCUBEA54992.2022.10010944.

[19] M. Patel, V. K. Prasad, P. Bhattacharya, M. Bhavsar and M. Zuhair, "Privacy Preservation for Big Data Healthcare Management," *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, London, United Kingdom, 2022, pp. 211-216, doi: 10.1109/ICIEM54221.2022.9853038.

[20] O. Ali and A. Ouda, "A classification module in data masking framework for Business Intelligence platform in healthcare," *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Vancouver, BC, Canada, 2016, pp. 1-8, doi: 10.1109/IEMCON.2016.7746327.

[21] N. Mohammed, S. Barouti, D. Alhadidi and, R. Chen, "Secure and Private Management of Healthcare Databases for Data Mining," *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, Sao Carlos, Brazil, 2015, pp. 191-196, doi: 10.1109/CBMS.2015.54.

[22] D. Gunawan, Y. S. Nugroho, Maryam, and F. Y. Al Irsyadi, "Anonymizing Prescription Data Against Individual Privacy Breach in Healthcare Database," *2021 9th International Conference on Information and Communication Technology (ICoICT)*, Yogyakarta, Indonesia, 2021, pp. 138-143, doi: 10.1109/ICoICT52021.2021.9527430.

[23] X. Liu, Y. Zheng, X. Yi, and S. Nepal, "Privacy-Preserving Collaborative Analytics on Medical Time Series Data," in *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 3, pp. 1687-1702, 1 May-June 2022, doi: 10.1109/TDSC.2020.3035592.

[24] F. Wang *et al.*, "Privacy-Preserving Collaborative Model Learning Scheme for E-Healthcare," in *IEEE Access*, vol. 7, pp. 166054-166065, 2019, doi: 10.1109/ACCESS.2019.2953495.

[25] A. Haddad, M. H. Habaebi, M. R. Islam, N. F. Hasbullah and S. A. Zabidi, "Systematic Review on AI-Blockchain Based E-Healthcare Records Management Systems," in *IEEE Access*, vol. 10, pp. 94583-94615, 2022, doi: 10.1109/ACCESS.2022.3201878.

[26] Q. Zhang, B. Lian, P. Cao, Y. Sang, W. Huang, and L. Qi, "Multi-Source Medical Data Integration and Mining for Healthcare Services," in *IEEE Access*, vol. 8, pp. 165010-165017, 2020, doi: 10.1109/ACCESS.2020.3023332.

[27] A. Hmood, B. C. M. Fung and F. Iqbal, "Privacy-Preserving Medical Reports Publishing for Cluster Analysis," *2014 6th International Conference on New Technologies, Mobility and Security (NTMS)*, Dubai, United Arab Emirates, 2014, pp. 1-8, doi: 10.1109/NTMS.2014.6814045.

[28] W. Ni, X. Huang, J. Zhang, and R. Yu, "HealChain: A Decentralized Data Management System for Mobile Healthcare Using Consortium Blockchain," *2019 Chinese Control Conference (CCC)*, Guangzhou, China, 2019, pp. 6333-6338, doi: 10.23919/ChiCC.2019.8865388.

[29] L. Wang and R. Jones, "Big Data, Cybersecurity, and Challenges in Healthcare," *2019 SoutheastCon*, Huntsville, AL, USA, 2019, pp. 1-6, doi: 10.1109/SoutheastCon42311.2019.9020632.

[30] M. Hanley and H. Tewari, "Managing Lifetime Healthcare Data on the Blockchain," *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced*

& Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Guangzhou, China, 2018, pp. 246-251, doi: 10.1109/SmartWorld.2018.00077.

[31] R. Somolinos *et al*., "Service for the Pseudonymization of Electronic Healthcare Records Based on ISO/EN 13606 for the Secondary Use of Information," in *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1937-1944, Nov. 2015, doi: 10.1109/JBHI.2014.2360546.

[32] M. Md. Siraj, N. A. Rahmat, and M. M. Din, " A Survey on Privacy Preserving Data Mining Approaches and Techniques," in *Proceedings of the 2019 8th International Conference on Software and Computer Applications (ICSCA '19)*. Association for Computing Machinery, New York, NY, USA, 2019, pp. 65–69. [Online]. Available: https://doi-org.proxy.lnu.se/10.1145/3316615.3316632

[33] R. Praveena Priyadarsini, M. L. Valarmathi, and S. Sivakumari, "Feature Creation based Slicing for Privacy Preserving Data Mining," in *Proceedings of the 3rd IKDD Conference on Data Science, 2016 (CODS '16)*. Association for Computing Machinery, New York, NY, USA, 2016, Article 15, 1–8. [Online]. Available: https://doi-org.proxy.lnu.se/10.1145/2888451.2888462

[34] A. K. Singh, A. Anand, Z. Lv, H. Ko, and A. Mohan, "A Survey on Healthcare Data: A Security Perspective," ACM Trans. Multimedia Comput. Commun. Appl. 17, 2s, Article 59, June 2021. [Online]. Available: https://doi-org.proxy.lnu.se/10.1145/3422816

[35] C. Desmet and D. J. Cook, "Recent Developments in Privacy-preserving Mining of Clinical Data," ACM/IMS Trans. Data Sci. 2, 4, Article 28, Nov 2021. [Online]. Available: https://doi-org.proxy.lnu.se/10.1145/3447774

[36] E. Júlio D. Aguiar, B. S. Faiçal, B, Krishnamachari, and J. Ueyama, A Survey of Blockchain-Based Strategies for Healthcare. ACM Comput. Surv. 53, 2, Article 27, Mar 2021. [Online]. Available: https://doi-org.proxy.lnu.se/10.1145/3376915

[37] T. Sinha, V. Srikanth, M. Sain, and H. J. Lee, "Trends and research directions for privacy preserving approaches on the cloud," in *Proceedings of the 6th ACM India Computing Convention (Compute '13)*. Association for Computing Machinery, New York, NY, USA, 2013, Article 21, 1–12. [Online]. Available: https://doi-org.proxy.lnu.se/10.1145/2522548.2523138

[38] M. Joshi, A. Pal, and M. Sankarasubbu, "Federated Learning for Healthcare Domain - Pipeline, Applications and Challenges," ACM Trans. Comput. Healthcare 3, 4, Article 40 2022, [Online]. Available: https://doi-org.proxy.lnu.se/10.1145/3533708

[39] S. S. Zainab, and T. Kechadi, “Sensitive and Private Data Analysis: A Systematic Review,” in *Proceedings of the 3rd International Conference on Future Networks and Distributed Systems (ICFNDS '19)*. Association for Computing Machinery, New York, NY, USA, 2019, Article 12, 1–11, [Online]. Available: https://doi-org.proxy.lnu.se/10.1145/3341325.3342002

[40] G. T. Kovács, and G. Kardkovács, “Survey on privacy preserving data mining techniques in health care databases,” Acta Universitatis Sapientiae, Informatica, 2014, 6(1), pp. 33-55, [Online]. Available: https://doi.org/10.2478/ausi-2014-0017

[41] A. Pika, M. T. Wynn, S. Budiono, A. H. M. ter Hofstede, W. M. P. van der Aalst, and H. A. Reijers, “Privacy-Preserving Process Mining in Healthcare,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 5, p. 1612, Mar. 2020, doi: 10.3390/ijerph17051612. [Online]. Available: http://dx.doi.org/10.3390/ijerph17051612

[42] K. Saranya and K. Premalatha, "Multi attribute case based privacy-preserving for healthcare transactional data using cryptography," *Intelligent Automation & Soft Computing*, vol. 35, no.2, pp. 2029–2042, 2023, [Online]. Available: https://doi.org/10.32604/iasc.2023.027949

[43] R. Ratra, P. Gulia, N. Singh Gill, and J. M. Chatterjee, "Big Data Privacy Preservation Using Principal Component Analysis and Random Projection in Healthcare", *Mathematical Problems in Engineering*, vol. 2022, Article ID 6402274, 12 pages, 2022. [Online]. Available: https://doi.org/10.1155/2022/6402274

[44] A. Alabdulkarim, M. Al-Rodhaan, T. Ma, and Y. Tian, “PPSDT: A Novel Privacy-Preserving Single Decision Tree Algorithm for Clinical Decision-Support Systems Using IoT Devices,” *Sensors*, vol. 19, no. 1, p. 142, Jan. 2019, doi: 10.3390/s19010142. [Online]. Available: http://dx.doi.org/10.3390/s19010142

[45] L. Wang *et al.*, "A User-Centered Medical Data Sharing Scheme for Privacy-Preserving Machine Learning", *Security and Communication Networks*, vol. 2022, Article ID 3670107, 16 pages, 2022. [Online]. Available: https://doi.org/10.1155/2022/3670107

[46] P. Jain *et al.*, "Blockchain-Enabled Smart Surveillance System with Artificial Intelligence," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 2792639, 9 pages, 2022. [Online]. Available: https://doi.org/10.1155/2022/2792639

[47] S38. K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, “Big healthcare data: preserving security and privacy,” *J Big Data* 5, 1, 2018, [Online]. Available: https://doi-org.proxy.lnu.se/10.1186/s40537-017-0110-7

[48] S39. S. J. Gabriel, Dr. P. Sengottuvelan, "A survey on privacy preserving data mining its related applications in health care domain," Journal of Algebraic Statistics, vol. 13, no. 3, p. 701 - 708. 2022, [Online]. Available: https://publishoa.com/index.php/journal/article/view/678/567

[49] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, “A comprehensive review on privacy preserving data mining,” *SpringerPlus* 4, 694, 2015, [Online]. Available: https://doi-org.proxy.lnu.se/10.1186/s40064-015-1481-x

[50] R. Mendes, and J. P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications," in *IEEE Access*, vol. 5, pp. 10562-10582, 2017, [Online]. Available: https://ieeexplore-ieee-org.proxy.lnu.se/abstract/document/7950921

[51] Dr. E. K. Reddy, “Security and privacy in Healthcare Data Mining,” *Academia.edu*, Oct. 2014. [Online]. Available: https://www.academia.edu/8970007/SECURITY_AND_PRIVACY_IN_HEALTHCARE_DATA_MINING

[52] J. J. Hathaliya, S. Tanwar, “An exhaustive survey on security and privacy issues in Healthcare 4.0,” Computer Communications, Volume 153,2020, Pages 311-335, ISSN 0140-3664, [Online]. Available: https://doi.org/10.1016/j.comcom.2020.02.018

[53] P. Jain, M. Gyanchandani, and N. Khare, “Big data privacy: a technological perspective and review,” *J Big Data* 3, 25, 2016, [Online]. Available: https://doi-org.proxy.lnu.se/10.1186/s40537-016-0059-y

[54] N. Venkataramanan, and A. Shriram, A, *Data Privacy: Principles and Practice*. Storbritannien: CRC Press, 2016. Accessed: May 22, 2023. [Online]. Available: https://www.google.se/books/edition/Data_Privacy/lpWKDQAAQBAJ?hl=sv&gbpv=1

[55] F. Sabir, F. Palma, G. Rasool, N. Moha, and Y-G. Guéhéneuc, “A systematic literature review on the detection of smells and their evolution in object-oriented and service-oriented systems,” Software: Practice and Experience. 49. 2018. doi: 10.1002/spe.2639.