Journal of Information Security, 2023, 14, 343-365
https://www.scirp.org/journal/jis
ISSN Online: 2153-1242
ISSN Print: 2153-1234

An Extensive Study and Review of Privacy Preservation Models for the Multi-Institutional Data

Sagarkumar Patel[1], Rachna Patel[2], Ashok Akbari[3], Srinivasa Reddy Mukkala[4]

[1]Department of Biometrics, LabCorp Drug Development Inc., Somerset, USA
[2]Department of Biometrics, Catalyst Clinical Research LLC, Wilmington, USA
[3]Department of Pharmacy, Shree Naranjibhai Lalbhai Patel College of Pharmacy, Umrakh, Surat, India
[4]Department of Biostatistics, EpisData, Sterling Heights, USA
Email: Sagarkumar.patel1@fortrea.com, Rachna.patel@catalystcr.com, Ashokakabari@gmail.com, Smukkala@episdata.com

How to cite this paper: Patel, S., Patel, R., Akbari, A. and Mukkala, S.R. (2023) An Extensive Study and Review of Privacy Preservation Models for the Multi-Institutional Data. *Journal of Information Security*, **14**, 343-365.
https://doi.org/10.4236/jis.2023.144020

Received: September 15, 2023
Accepted: October 7, 2023
Published: October 10, 2023

Abstract

The deep learning models hold considerable potential for clinical applications, but there are many challenges to successfully training deep learning models. Large-scale data collection is required, which is frequently only possible through multi-institutional cooperation. Building large central repositories is one strategy for multi-institution studies. However, this is hampered by issues regarding data sharing, including patient privacy, data de-identification, regulation, intellectual property, and data storage. These difficulties have lessened the impracticality of central data storage. In this survey, we will look at 24 research publications that concentrate on machine learning approaches linked to privacy preservation techniques for multi-institutional data, highlighting the multiple shortcomings of the existing methodologies. Researching different approaches will be made simpler in this case based on a number of factors, such as performance measures, year of publication and journals, achievements of the strategies in numerical assessments, and other factors. A technique analysis that considers the benefits and drawbacks of the strategies is additionally provided. The article also looks at some potential areas for future research as well as the challenges associated with increasing the accuracy of privacy protection techniques. The comparative evaluation of the approaches offers a thorough justification for the research's purpose.

Keywords

Privacy Preservation Models, Multi Institutional Data, Bio Technologies, Clinical Trial and Pharmaceutical Industry

1. Introduction

Currently, the world is increasingly witnessing technological innovations such as big data, nanotechnology, cloud computing, biotechnologies, artificial intelligence, and the Internet of Things (IoT), which collectively form part of the fourth industrial revolution [1]. Furthermore, the fourth industrial revolution has brought in fast-paced development in commercial and governmental organizations alike, affecting everyday pursuits worldwide [2]. Chapter two discusses the works carried out by various researchers to assure secure data preservation in health care applications. Existing literature suggests that Anonymization-based models, Blockchain-based models, and optimization approaches can be used synergistically to address the issues concerning the secure provision of e-health data.

IoT is one such present-day innovation that has percolated routine life through various applications, ensuring a safe, innovative, and more accessible environment [3] [4]. IoT enables the interaction between humans as it integrates millions of people, intelligent nodes, substantial objects, services, and digital sensors; it also consists of millions of digital sensors [5]. The significant feature of the IoT is that all the objects in the environments are interlinked to each other as it transfers the data in any part of the world at any time [6] [7] [8]. Modern innovations such as big data, cloud computing, fog computing, allocated computing, and wireless communication aid the IoT to attain its intention of facilitating the interaction between intelligent [9] [10] [11] [12]. IoTs are used in various applications and domains, such as coordination, transportation, medical care, well-being, insight, and many more applications.

The IoT devices are able to produce enormous amounts of data known as big data due to a wide range of applications including transportation, smart homes, the health care industry, and electricity conservation [13]. The following section briefly elucidates IoT big data. IoT gadgets are known to involve copious amounts of private and sensitive information. For example, Cisco expects that 500 billion devices will be connected to the internet by 2025 [14] [15] [16]. As a result, the amount of structured and unstructured data has increased by 2.5 Exa bytes daily [17]. On the other hand, the worldwide server farm's IP traffic would only arrive at 10.4 zetta bytes [18] [19] [20]. This rapid expansion in information volume is attributed to web administrations, versatile information, and medical care information [21].

With the rapid adoption of global IoT-connected devices, enormous amounts of data are transferred between cloud-based and physical network environments. It has also brought in many technologies that create a vision of interconnecting the world through devices. A plethora of privacy and security challenges are witnessed in IoT architectures [22] [23]. Generally, IoT-based healthcare applications hold sensitive medical details of the patients, for which confidentiality is necessary to ensure the privacy of the patients. Due to challenges associated with digital data, conventional encryption strategies over structural and textual one-dimensional data are not used for e-health data directly. In addition, when sensi-

tive information is forwarded through open channels, patients may suffer from the loss of information contents. Hence, a secure key frame extraction strategy is needed to ensure appropriate privacy-preserving e-health services. The multi-institutional clinical data yield enough information to identify the small differences and improve general ability.

Between 2017 and 2019, data breach incidents rose from 15% to 26%, as reported by professionals involved in risk oversight activities [24]. The healthcare industry is turning to big data technology to improve and manage medial systems. For this purpose, healthcare companies and organizations are leveraging big data in health informatics [25]. The security gap where the question arises is whether the third-party policies and safeguards regarding IoT security are sufficient for preventing data breaches which is one of the main reasons for the rising IoT threats. Recently, sophisticated hacks on IoT devices have aggravated the problem, and therefore, the privacy and security problems in the IoT paradigm are discussed in detail. In this chapter, the existing pool of literature that is relevant to the research topic will be reviewed by surveying academic journals, technical reports and data, books, scholarly articles, and other relevant publications. The secondary literary sources will be reviewed and analyzed thoroughly that are relevant to the privacy and data security management in healthcare and clinical research, the role of IoT and big data analytics (BDA), as well as the risks and challenges concerning data privacy and security management in IoT and big data analytics.

Deep Learning Models in Clinical Applications

In recent years, the integration of deep learning models into clinical applications has ushered in a new era of healthcare innovation. These advanced computational tools have demonstrated immense potential in revolutionizing disease diagnosis, treatment planning, and patient care [26]. The ability of deep learning algorithms to analyze complex medical data, such as medical images, genomic sequences, and electronic health records, has opened up a plethora of opportunities for improving clinical outcomes and healthcare delivery [27]. Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown remarkable capabilities in tasks like medical image classification, natural language processing, and predictive modeling. From the early detection of tumors in radiological scans to the personalized treatment recommendations based on genetic profiles, the impact of these models on the medical field is profound [28] [29]. They offer the potential to augment the expertise of healthcare professionals, expedite diagnosis, reduce errors, and enhance patient outcomes. However, as the healthcare industry embraces the transformative potential of deep learning, it also faces a critical challenge: the protection of patient rights and data security. While deep learning algorithms excel at extracting insights from medical data, the sensitivity and confidentiality of patient information must not be compromised. Ensuring that the healthcare ecosystem remains

a trusted guardian of individual privacy is paramount.

2. Literature Review

Categorization of the Privacy Preservation Models:

Figure 1 shows the schematic block diagram representing the categorization of the privacy preservation models in e-health data.

2.1. Schemes Concerning Anonymization N-Based Techniques

A presented a revolutionary system called Spark that makes use of Apache Spark to efficiently manage large amounts of health care data as well as K-anonymization and L-diversity to protect sensitive personal data. Furthermore, the developed strategy ensures that the shared e-health data does not reveal or isolate the original data before moving to the Hadoop distributed file system (HDFS) [5].

In a study, two solutions are offered that protect user privacy in parking recommender systems while analyzing the past parking history utilizing k-anonymity (anonymization) and differential privacy (perturbation) techniques [6]. The k-anonymity mechanism specifically creates an anonymized database from an original parking database that contains users' parking information. The users are indistinguishable in both methods due to differential privacy, which also perturbs the Laplace mechanism's query response. It is now possible for customers to receive parking place recommendations while maintaining their privacy due to experimental findings on a data set built from actual parking measurements [6].

In order to use Fully Homomorphic Encryption (FHE) schemes, a presented proxy re-ciphering as a service employing traditional methods such as threshold secret sharing, distributed semi-trusted proxy servers, and chameleon hash function [7]. The effectiveness of the developed strategy is analyzed using real-world data. Furthermore, the strategy's security characteristics are also analyzed over the general cyber threats, ensuring that the developed method is a sensible, scalable, and easy-to-use strategy for the long-term prevention of sensible data [7].

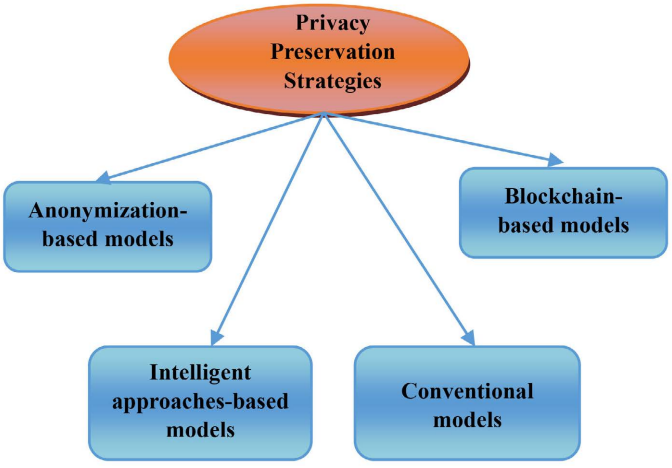

Figure 1. Categorization of the privacy preservation models.

A method to develop a productive anonymous algorithm to maintain the privacy of private data by the data owners [8], privacy frequently occupies a chief role in data mining strategies, and a number of anonymous algorithms introduce privacy in digging data. However, there are limits to privacy protection, and hence a novel strategy of Efficient Anonymous Algorithm (NEAA) is introduced. At first, the entity process's raw data and the data are preserved in the database. Then, the sensitive data is evaluated using the PCA-based Attribute selection algorithm. The hiding process introduces a novel algorithm based on Anonymous (NBA). Finally, anonymous data can be produced as a result [8]. The performance of the developed system is analyzed and found to provide enhanced performance compared to conventional systems [8].

A presented a privacy-preserving protocol based on keyword search with security for EHR system suggested [9]. Through the use of a keyword search in the cipher text, which is then once more re-encrypted by the cloud using the re-encryption key created by the patient, this method can quickly identify the history of health records that are related. It guarantees that confidential information cannot be disclosed by unauthorized users. An entity-based access control system ensures that only the intended data requester has access to the patients' medical records [9].

A model to recognize how the connection of records can assist in developing the entire patient profiles and thus adjoin importance to the conventional healthcare systems was presented by [11]. The usage of data anonymity shows how privacy is fulfilled by specifying the knowledge about the background and further limiting access to factual data. A semantic strategy counting policy formalization, compliance examination, and knowledge discovery was carried out to prevent the risks related to privacy with arbitrary linkages [11].

The tuple partitioning approach is used in an effective quasi-identifier index-based architecture for privacy preservation over incremental databases on the cloud design [12] here the merging of columns generates anonymous information for the user. In addition, a packetization strategy is applied to enhance the effectiveness of privacy-preserved information further. The outcomes of the method over a real-world database have proved that the developed model effectively preserves the incremental database of large volumes compared to conventional strategies [12].

2.2. Schemes Concerning Blockchain-Based Models

A blockchain structure for managing electronic health records (EHR) could provide the patient with rights and control over the EHRs. The Ancile structure also exposes a blockchain system that achieves a higher degree of decentralization while admitting some nodes as having greater power. According to the evaluation, it is very unlikely that all the data will be covered while maintaining a usable and interoperable model. Ancile, however, still offers a significant amount of data integrity and privacy protection due to the use of smart contracts to partition the data [13].

A blockchain-based medical data preservation scheme (DPS) ensures the primitiveness and verifiability of EHRs while preserving privacy for the data owner [9]. After receiving authorization from the data owner, the data requestor uses the expected keyword from the data provider to analyze the expected EHRs through the EHR consortium blockchain and obtains the re-encryption cipher text from the cloud server. To comprehend data security, privacy preservation, and access control, the technique largely uses searchable encryption and interim proxy re-encryption. Additionally, proof of agreement is developed as a compromise method for consortium blockchain to guarantee the availability of the system [9]. In order to demonstrate the effectiveness of the proposed scheme in terms of computing efficiency, the cryptographic primitives were also emulated and the planned scheme was implemented on the Ethereum platform [9].

They suggested the blockchain-based interoperability problem in EHR and determined the number of cross-blockchain-based EHR storage strategies [10]. Initially, an EHR privacy-preserving cross-blockchain strategy was presented based on Polka Dot chain technology for EHR circulation among several private blockchains of various hospitals. The issue of removing EHR data from each hospital's private blockchain is solved by RaaS. Using the developed mechanism constructs it beneficial and effective for the doctors to access the data and for patients to remove the unwanted EHR data when they visit various hospitals, which acts a significant function in the EHR privacy-preserving area and in contravening the remote islands of sharing medical data [10].

The blockchain-based model was developed for the management of EHR on a distribution network to ensure the eventual privacy of patients' health records, providing control to the patients to monitor the access of data by others through the developed method [15]. Gathering and organization of data into Big Data provide a huge possibility of varying the healthcare viewpoint, like in personalization care of patients, the discovery of the drug, the efficiency of the treatment, enhancement in clinical results, and the safety management of the patients. Furthermore, the blockchain offers a proposal for which the EHR of patients is preserved without tempering or any attacks. Then to ensure ultimate isolation and control over access to an EHR sheet on the blockchain, a channeling strategy ensures that the patients accept the entities only within a distributed network to completely access the data [15].

2.3. Schemes Concerning Intelligent Techniques

In order to protect patient privacy, the Privacy-Preserving Optimization of Clinical Pathway Query (PPO-CPQ) system uses a safe clinical pathway query on cloud servers for e-healthcare., and the sensitive data corresponding to the hospitals, like treatment, expense, and medication. Under this strategy, a secure and privacy-preserving sub-protocol is initially designed, which constitutes a comparison of privacy-preservation, selection of privacy-preserving stage, and so on, to assure the privacy of the e-Healthcare system. The query is then safely executed

using the greedy algorithm, and the system's efficiency is increased by using the Min-Heap technique [16].

Big data privacy has been successfully preserved using an algorithm known as Grey Wolf Optimizer-Cat Swarm Optimization (GWO-CSO). The generated model, which is used to create the k-anonymization database in which k numbers of duplicate records are created within the real database [17], is obtained by changing the update rules of GWO in the presence of the CSO algorithm. In order to provide safe data transmission over to end users, the newly created technique is used in the k-anonymized database to conceal the sensitive information relating to the data owners. The created technique ensures the k-anonymization phenomena by producing the parameters required to build the k-anonymized database ideally [17]. The privacy and utility metrics consider evaluating the fitness of the solutions obtained using the developed algorithm [17].

Mandala and Rao presented the privacy preservation model for private healthcare data [18]. The designed model mainly concentrated on introducing an effective sanitizing model to conceal the sensitive information of users. A key was framed and selected optimally using the Adaptive Awareness Probability-based Crow search algorithm (AAP-CSA) approach to conceal the sensitive medical data. Moreover, the designed strategy was analyzed in terms of various attacks with different algorithms to show the effectiveness of the designed method [18].

2.4. Schemes Concerning Conventional Techniques

A privacy-preserving chaos-based privacy-preserving encryption model was designed by to protect patients' privacy [19]. The developed model could protect the images of the patients from a cooperative broker. To secure the essential frames of the data gathered from the wireless capsule endoscopy strategy specifically, a quick probabilistic model was devised and a prioritization method was used. The created model produces encrypted images that are random in nature, which improves computational efficiency. The created methodology also entails processing medical data without any leaks, and protecting patient privacy by allowing only authorized users to access the data [19].

Yang presented a practical and privacy-preserving prediction scheme to predict disease risk using e-healthcare, named EPDP [20]. In contrast to the current approaches, the created EPDP successfully completes the two steps of disease risk prediction, such as disease model training and disease prediction, while ensuring improved privacy preservation. Notably, a cryptographic approach is used with a super-increasing order to efficiently collect each disease's symptoms throughout the disease model training phase. Results from the disease risk prediction phase were evaluated using the bloom filter approach.

Chamikara modeled a solution to maintain data privacy in significant data distribution and evaluation strategies, a challenge in smart cyber-physical models [21]. The developed algorithm called SEAL is used to preserve data privacy. The linear time complexity associated with SEAL assists in working with the

continuously increasing big data and data streams effectively. The method attains increased accuracy in classification, scalability, and efficiency while maintaining enhanced privacy and higher attack resistance compared to other methods. The findings indicate that the approach is appropriate for smart cyber-physical environments, including grids, cars, healthcare systems, and homes because it can control the continuous data streams produced by sensors monitoring a single person or a group of people and processing them before sending them to cloud systems for additional analysis [21].

The remote data integrity checking strategy used the fine-grained update for big data storage [22]. The developed strategy attains general processes of modification, insertion, and deletion at any online location level in a file with a mapping relation among the block-level update and the line-level update. The analysis of the method depicts that the developed strategy assists in privacy preservation and public verification. On the other hand, it carries out data integrity with a low reduced cost for communication and computation [22].

To protect private information, a big data probabilistic approach based on clustering was deployed in such a way as to obtain maximum privacy and minimum perturbation. In this model, sensitive data is preserved after recognizing the confidential information from the data clusters to adjust or generalize them. The resultant database is examined to evaluate the level of accuracy of the model concerning hidden data and loose data due to reconstruction. Results demonstrate that the developed clustering-based Privacy preservation strategy in big data led to successful reconstruction [23].

2.5. Risk Outcomes towards Data Privacy and Security Management

In accordance with [24], if safe encryption is used in devices or if the internet service provider or network observers analyze the internet traffic from the smart homes connected to the IoT devices, they can gather sensitive information about the activities that take place at home. To prevent the gadgets from becoming inoperable, the user must not block outgoing traffic from their residence, an example of privacy risk outcomes associated with IoT applications in smart homes. IoT's privacy risks and threats can be user identification, user tracking, profiling, utility controlling, and monitoring. From the privacy perspective, the threat associated with user identification is the ability of the device to distinguish or reveal the identity of the person based on acquired data like name, address, or any such personal information. Such a threat aggravates other associated threats like tracking and profiling individuals' behavior.

The high volume of healthcare data creates a big challenge, the desire for scalable storage and support for distributed queries across multiple data sources. Specifically, the challenge is being able to locate and mine specific pieces of data in an enormous, partially structured dataset [25]. The study results showed that privacy, familiarity, and security levels affected users' trust in using IoT devices.

Due to the privacy and security concerns associated with IoT, the amount of trust in devices also had an impact on people's perceptions of risk and their attitudes towards using it.

According to a recent study conducted IoT devices are increasingly becoming pervasive in our everyday life, so it is important to understand the underpinning privacy and security risks associated with them [26]. These risk factors result in attacks by cyber attackers faced by consumers using IoT devices.

IoT has tremendous benefits but presents a reminder that IoT has security and privacy implications. The health data is susceptible, and its granularity poses significant challenges to the anonymization of personal information and thereby exposes consumers to data security and privacy risks. Unauthorized people can intrude into IoT data and use them in authorized ways. Moreover, the increased reliance upon big data and IoT-based devices heightens the risk of security threats and a vulnerability point for intruders to access users' personal information [27]. In addition to security threats, IoT devices are vulnerable due to many factors. Firstly, the IT manufacturers of IoT devices are inexperienced regarding the risks related to data security relative to hardware or software items. Secondly, the security measures like encryption also need to be fully considered, and thirdly, it is hard to update these devices periodically with security fixes.

3. Analysis and Discussion

This section examines the method for multi-institution data that preserves privacy. The parts that follow analyze the study based on the parameter measurements.

3.1. Search Strategy

To conduct a comprehensive survey on the topic of privacy preservation for multi-institutional data, a systematic search strategy was employed across a diverse set of journals and sources. The search primarily focuses on the following key databases and journals: IEEE, Wiley, Elsevier, Arxiv, Future Generation Computer Systems, International Journal of Advanced Computer Science, Springer, and Berkeley Technology Law Journal. Additionally, relevant proceedings from the 2018 workshop and publications in the International Journal and Elementary Education are included.

3.1.1. Search String

The search string involves using a combination of relevant keywords and phrases, such as privacy preservation, multi-institutional privacy, data security, multi-institutional data, healthcare, confidentiality, and institutional privacy safeguards. A backward and forward citation tracking approach also is employed to identify seminal papers and relevant references. This search approach aims to encompass a broad range of sources and perspectives to provide a comprehensive overview of privacy preservation practices in multi-institutional contexts. By

specifying the publication year range from 2016 to 2023, the search retrieves relevant and up-to-date literature on this important topic.

3.1.2. Inclusion Criteria

The Selected papers primarily focus on methods, techniques, or frameworks related to the preservation of privacy within the context of multi-institutional data, ensuring that the research is directly aligned with the survey's subject matter. Articles published between 2016 and 2023 included allowing for an up-to-date understanding of the evolving landscape of privacy preservation in multi-institutional data. The survey considers research articles, peer-reviewed conference papers, books, and relevant reports with rigorous academic or professional analysis.

3.1.3. Exclusion Criteria

Studies published before the year 2016 will be excluded from the survey. This criterion ensures that focus on the most recent research and developments in privacy preservation, considering that older publications may not reflect current practices and technologies. Exclude articles that do not have direct involvement or expertise in multi-institutional data management, as their input may not contribute to meaningful insights. Exclude papers with incomplete or missing answers to key questions, as these can introduce bias and reduce the reliability of the findings.

3.2. Analysis Based on the Performance Metrics

The analysis is conducted using parameter measurements that have been employed by numerous academics to demonstrate the model's efficacy. A number of metrics are observed, including MAE, RMSE, information loss, time, computation cost, communication cost, search time in seconds, search time per identifier in seconds, communication overhead, information loss, classification accuracy, attack range, and cost function NPCR-number of pixel change rate, UACI-unified average changing in tensity tests, minimum std, average overhead bandwidth. Observations reveal that the metrics time, computation cost, and range of attacks as shown in **Table 1** are frequently used by researchers. The measures that the reviewers used are explained in **Figure 2**.

3.3. Analysis Based on Publication Year

Depending on the journal's publication year, the analysis is done in this section. **Table 2** presents a review of the output from 2013 to 2021. **Figure 3** gives detailed information and represents the majority of literary works from the 2020s that are pertinent to the analysis of privacy preservation techniques.

3.4. Analysis Based on Journals

This section evaluates the articles that have been published in the aforementioned publications and are relevant to privacy protection techniques. **Table 3**, shows analysis concerning of journals and their affiliated papers. Journals from

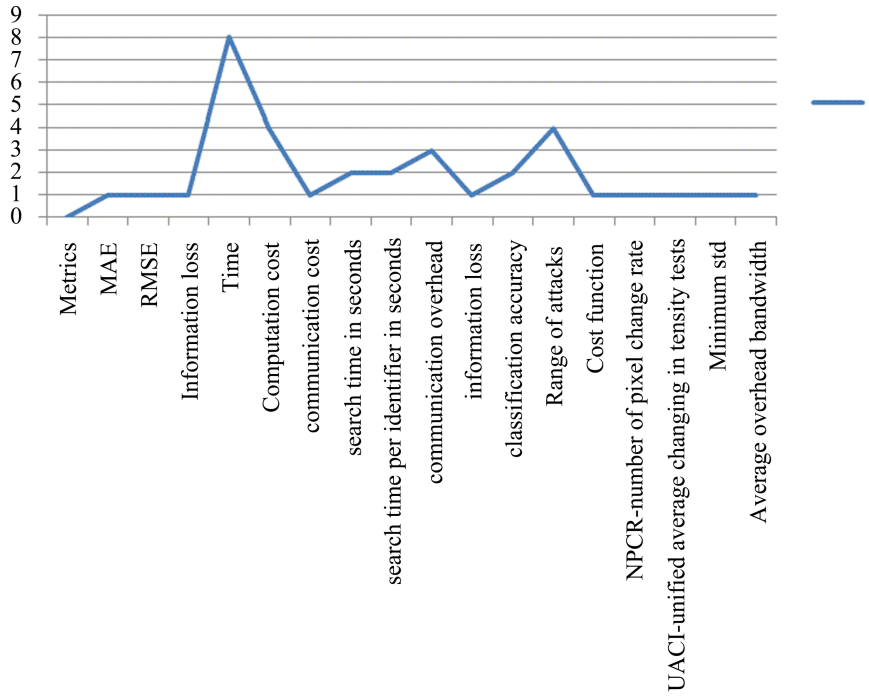

Figure 2. Analysis concerning metrics.

Table 1. Analysis concerning metrics.

Metrics	Papers
MAE	[6]
RMSE	[6]
Information loss	[8]
Time	[5] [7] [8] [11] [12] [15] [21] [22]
Computation cost	[9] [14] [16] [20]
Communication cost	[9]
Search time in seconds	[10] [23]
Search time per identifiers in seconds	[10]
Communication overhead	[13] [16] [20]
Information loss	[17]
Classification accuracy	[17] [21]
Range of attacks	[1] [2] [3] [18]
Cost function	[18]
NPCR-number of pixel change rate	[19]
UACI-unified average changing in tensity tests	[19]
Minimum std	[21]
Average overhead bandwidth	[24]

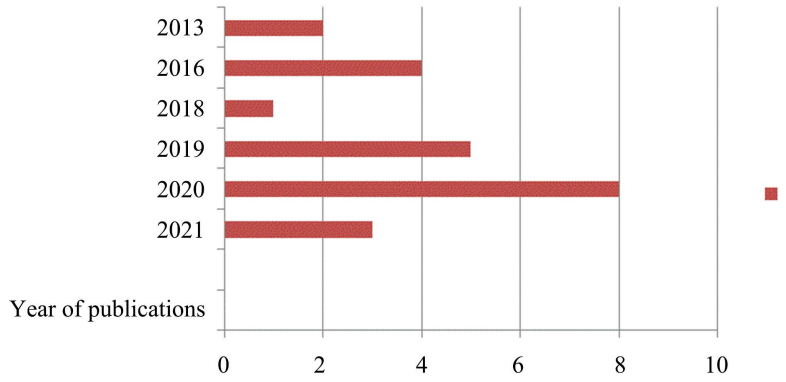

Figure 3. Analysis concerning publication year.

Table 2. Analysis concerning publication year.

Year of Publications	Papers
2021	[6] [8] [17]
2020	[2] [7] [14] [15] [16] [19] [21] [23]
2019	[1] [9] [10] [13] [18]
2018	[12] [20] [22] [24]
2016	[3] [4]
2013	[11]

Table 3. Analysis concerning journals.

Journal Names	Papers
IEEE	[2] [4] [5] [9] [10] [13] [14] [15] [16] [20]
Wiley	[6] [7]
Elementary Education	[8]
Elsevier	[11] [12] [18] [19]
International Journal	[17]
Arxiv	[21]
Future Generation Computer Systems	[22]
International Journal of Advanced Computer Science	[23]
In Proceedings of the 2018 Workshop	[24]
Springer	[1]
Berkeley Technology Law Journal	[4]

the IEEE, Wiley, Springer, elementary education, Elsevier, international journals, arxiv, and next generation computer system can all be used to get publications about privacy preservation approaches. The vast majority of publications from Elsevier and IEEE have extensive analyses and reviews, which are interpreted in **Figure 4**.

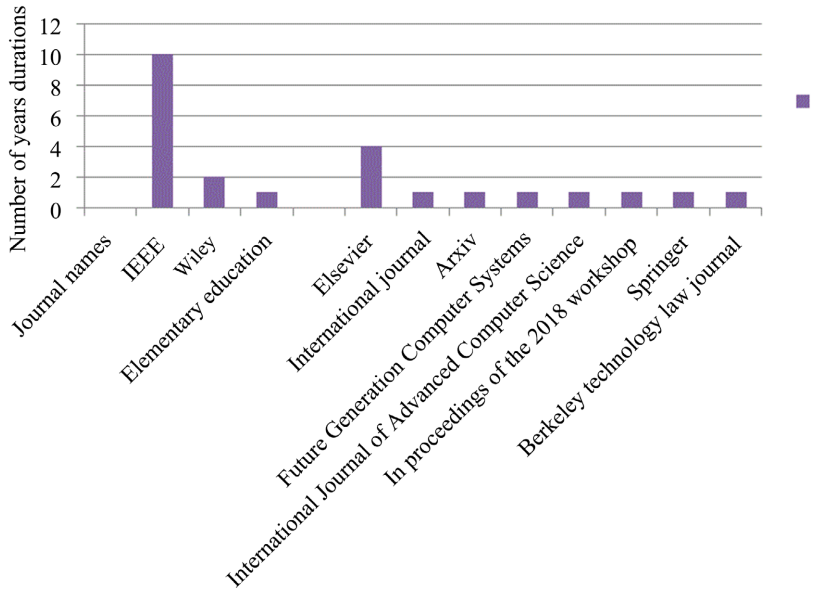

Figure 4. Analysis concerning journals.

3.5. Analysis Based on Dataset

Based on the datasets used by various researchers, Table 4 offers an interpretation of the analysis. Several datasets, including Apache Spark, Real Word, Mockaroo, EHR, Eron Email, Medical Record, Standard Adult, and others, are assessed. Figure 5 provides a description of the datasets used by the reviewers.

4. Research Gaps and Challenges

- Privacy protection is of significant consideration in big data, and hence demanding resilient strategies is necessary to safeguard the customers' privacy.
- Apart from privacy concerns, efficiency issues, such as the communication overhead and computational cost among the providers and the servers, must also be considered.
- Achieving a better level of security while maintaining reasonable computational complexity remains an onerous task with the transmission of medical images in real-time applications.
- The expanded dimensions of e-health data are one of the major issues with demanding privacy. The complexity of traditional data encryption techniques prevents them from being used in modern applications or for real-time image transmission.
- Satisfying the demands of an image encryption model in terms of greater security to reduce processing complexity is a difficult task.
- Sensitive data in the healthcare industry must be kept private in order to avoid interfering with patient privacy or the activities of healthcare associations.
- The preservation of privacy over incremental data sets is still tricky in the cloud framework, as most data sets are large in volume and scattered across multiple storage nodes.

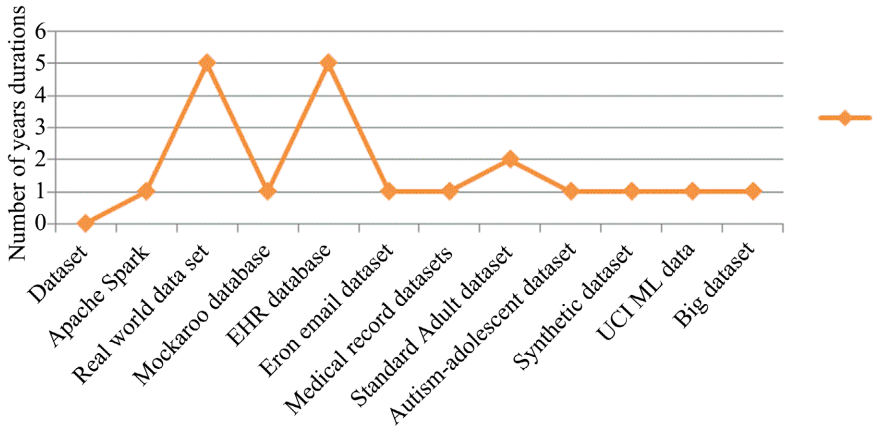

Figure 5. Analysis concerning dataset.

Table 4. Analysis concerning dataset.

Dataset	Research Papers
Apache Spark	[5]
Real world Data Set	[3] [6] [11] [20] [24]
Mockaroo Database	[8]
EHR Database	[9] [12] [13] [14] [15]
Eron Email Dataset	[10]
Medical Record Datasets	[1] [2] [4] [16] [22]
Standard Adult Dataset	[17] [19]
Autism-Adolescent Dataset	[18]
Synthetic Dataset	[20]
UCI ML Data	[21]
Big Dataset	[23]

- The privacy preservation using query set consumes more time and more resources to carry out the computation and hence cannot be used in real-time applications.
- In distributed data mining, certain sum, scalar product, secure set union, and set intersection are a few of the operations regarded as the fundamental operations. Due to the increasing computational complexity, they are unable to offer enough data utility and are unreasonable for privacy-preserving data mining (PPDM).
- Most conventional authentication techniques) could not provide enhanced security and performance characteristics to prevent potential attacks.
- An authentication protocol based on ECC developed showed formal and informal security studies to confirm security efficiency. However, the developed strategy cannot avoid the attacks, such as stolen-verifier and offline password-guessing attack.

Table 5. Analysis of privacy preservation in multi-institutional data.

Paper No	Data Acquisition	Applications	Parameter	Advantage	Disadvantage	Solution
1	EHR	IOT		Blockchain-based EHR sharing gives patients greater control over their health records, allowing them to manage access and permissions securely.	Implementing blockchain-based EHR systems requires a good understanding of blockchain technology and may be complex for healthcare organizations with limited expertise.	Creating user-friendly interfaces and tools for patients and healthcare providers can simplify the use of blockchain-based EHR systems.
2	Real Time	IOT Based Health Care		helps in understanding user attitudes towards IoT-based healthcare and the impact of privacy and security on trust.	Privacy and security concerns associated with IoT can affect users' trust in using IoT devices.	Manufacturers should implement transparent privacy measures in IoT devices and educate users about them to build trust.
3	Real-Time	Consumer IoT		The research highlights IoT device privacy and security risks, encouraging a proactive response.	As IoT devices become more pervasive, the risk of privacy and security breaches may rise.	Continuous monitoring and regular updates of IoT devices can help mitigate evolving security threats.
4	Real Time	IOT		The challenges of health data anonymization and the potential for unauthorized access to IoT data.	IoT devices are vulnerable due to factors such as inexperienced manufacturers and inadequate security measures.	Requires data security training and must embed strong encryption and security features into IoT devices.
5	Eron Email Dataset	Healthcare Big Data		The Spark system efficiently manages large healthcare datasets, improving data processing speed and scalability.	Implementing Apache Spark and advanced anonymization techniques may be technically challenging for some organizations.	Invest in optimizing the computational resources required for Spark and anonymization processes.
6	Real Parking Dataset	Smart Parking System	Time stamp	K-anonymity and differential privacy techniques protect user privacy in parking recommender systems.	Anonymization techniques can impact the quality and utility of data.	Research and develop methods to balance privacy protection with data utility.

Continued

7	TELE ECG Database	Cloud-IoT	Minimum number of proxy servers in a homomorphic encryption.	The tuple partitioning approach effectively preserves privacy over incremental databases on the cloud.	Implementing tuple partitioning and packetization may be complex.	Explore simplified implementations of privacy-preserving strategies for broader use.
8	Mockaroo Database	Data Mining		The Efficient Anonymous Algorithm (NEAA) provides enhanced privacy for private data.	Implementing privacy algorithms may require technical expertise.	Continuously refine NEAA and similar algorithms to improve privacy protection.
9	Electronic Health Record (EHR)	Healthcare	System parameters	A prototype demonstrates the practical applicability of the established protocol.	Implementing mutual anonymous authentication protocols may require technical expertise.	Offer training and resources to facilitate the implementation of authentication protocols.
10	Eron Email Dataset	Blockchain	Security parameter	The proposed bloom filter-enabled multi-keyword search protocol emphasizes privacy preservation. By reducing the exposure of intermediate results, it minimizes the risk of service peers or other entities accessing sensitive information associated with the encrypted keywords.	While the protocol shows promise in a simulated environment, its real-world applicability and robustness may need further validation and testing in actual blockchain systems, which could present unforeseen challenges.	To address the complexity issue, clear and comprehensive documentation, along with training resources, should be made available to facilitate the implementation and operation of the protocol.
11	Adult dataset from UCI Machine Learning Repository	Cloud Computing		Scalability is a key advantage, particularly in cloud environments where data can be distributed across multiple storage nodes.	While the approach offers scalability and efficiency, it may also introduce complexity due to the need for indexing and specialized algorithms.	user-friendly tools and interfaces that allow healthcare organizations to easily implement the proposed approach without requiring extensive technical expertise.
12	EHR	Blockchain		Proxy re-ciphering and traditional methods enhance data security for sensitive information.	Some encryption methods may demand significant computational resources.	Invest in optimizing resource-intensive encryption techniques.
13	EHR	Blockchain	System parameters	The protocol based on keyword search enhances data security and privacy in the EHR system.	Certain encryption techniques may require significant computational resources.	Explore resource-efficient encryption methods.

Continued

14	EHR	Blockchain		The paper addresses the issue of cross-blockchain-based EHR storage strategies, which can improve interoperability between healthcare institutions. This ensures that patients can access their EHR data efficiently, even when visiting different hospitals.	Implementing cross-blockchain solutions based on Polka Dot chain technology and RaaS may involve technical complexities. Healthcare organizations may require specialized expertise for deployment.	Implementing cross-blockchain solutions based on Polka Dot chain technology and RaaS may involve technical complexities. Healthcare organizations may require specialized expertise for deployment.
15	EHR	Blockchain		The blockchain-based model proposed in this paper provides patients with control over their health records. Patients can monitor data access and ensure data integrity, enhancing patient empowerment.	Integrating existing healthcare systems with blockchain technology can be challenging, particularly in large, established healthcare organizations.	Designing user-friendly interfaces and tools for patients and healthcare providers can simplify the use of blockchain-based EHR systems.
16	International Classification of Diseases (ICD)	E-Health Care System	Minimum information loss, total number of clusters, cluster centroid	The development of privacy-preserving sub-protocols demonstrates a systematic approach to privacy protection, making it easier to adapt the scheme to various clinical scenarios.	Privacy-preserving protocols, while effective, can be complex to implement and require a thorough understanding of cryptographic techniques. This complexity may hinder adoption by healthcare professionals.	Continued research into optimizing k-anonymization techniques can reduce resource requirements while maintaining strong privacy guarantees.
17	Adult Dataset		Awareness probability	The proposed framework prioritizes patient data privacy by utilizing local differential privacy, ensuring that sensitive patient information remains secure during the collaborative training process.	Implementing a fog-based federated framework with differential privacy can be technically challenging, requiring expertise in both healthcare and privacy-preserving machine learning.	Continuous research and development efforts can focus on optimizing the communication protocols and algorithms used in federated learning to reduce overhead and improve efficiency.

Continued

18	Autism-Adolescent Dataset	Health Care	Awareness probability	The blockchain-oriented privacy-preserving EHR sharing protocol ensures that only authorized data requestors can access sensitive EHRs, protecting patient privacy effectively.	The cryptographic operations involved in the protocol may introduce computational overhead, potentially affecting system performance. This complexity can be a limitation in resource-constrained environments.	Researchers can focus on optimizing the cryptographic primitives used in the protocol to reduce computational overhead and enhance efficiency. This can make the solution more practical for real-world implementation.
19	WCE Dataset	IoT-E Health Care	Trajectory	The model employs chaos-based privacy-preserving encryption to protect patients' privacy effectively. It ensures that sensitive patient images remain confidential and secure.	Implementing advanced encryption techniques can be computationally intensive. This may pose challenges in resource-constrained healthcare environments.	To address computational intensity, research can focus on optimizing encryption algorithms for efficiency, making them more suitable for healthcare applications.
20	Acute Inflammations Dataset, Real Time Dataset	E-Healthcare	Bilinear pairing parameters	The usage of data anonymity enhances privacy by limiting access to factual data.	Implementing data anonymity and semantic strategies may be complex.	Offer training and guidance on the implementation of data anonymity and semantic strategies.
21	PABIDOT Perturbs	Health Care	Perturbation parameters	SEAL algorithm ensures data privacy in significant data distribution and evaluation strategies while maintaining efficiency, scalability, and higher attack resistance.	Implementing advanced privacy-preserving algorithms may require expertise and careful configuration.	Developing user-friendly interfaces for complex algorithms can facilitate adoption in healthcare settings.
22	Data Access Control System	Health Care	Security parameter	The fine-grained update strategy ensures data integrity and privacy preservation while minimizing communication and computation costs.	Implementing fine-grained updates may introduce complexity in data management.	Developing clear implementation guidelines can assist healthcare organizations in effectively adopting such strategies.
23	Adult's Dataset and Bank Marketing Dataset			The clustering-based Privacy Preservation strategy in big data led to successful data reconstruction, ensuring privacy and utility.	Data clustering and generalization may lead to some loss of data granularity.	Research can focus on optimizing data generalization techniques to minimize data loss.

Continued

24	Real Time	Smart Home IoT	Independent link padding (ILP) or dependent link padding (DLP), control padding and fragmentation parameters.	privacy risks associated with IoT applications in smart homes, shedding light on potential threats and concerns.	Users may not be able to block outgoing traffic from IoT devices, posing privacy risk	IoT devices should provide users with robust privacy settings to control outgoing traffic effectively.
25	Block Chain	Educational Institution	Local parameters	The system prioritizes privacy by allowing users to have control over their credentials. This is crucial for protecting sensitive student information.	Integrating the blockchain solution with existing education systems and institutions may pose challenges, as standardization and compatibility issues can arise.	Establish interoperability standards to facilitate the integration of blockchain solutions with existing education systems seamlessly.
26	COVID-19 CT Datasets: COVID-19-CTSeg, Mos MedData Dataset	IoMT	Model weights	Collaborative training using data from multiple institutions enhances the robustness and generalizability of deep learning models in medical imaging, making them more effective in diverse clinical settings.	Collaborative learning involves data exchange among institutions, which can result in increased communication overhead and potential latency issues.	Developing efficient communication protocols and data compression techniques can reduce communication overhead and address latency concerns.
27	National Lung Screening Trial, Medical Imaging Data Resource Center, BRATS, and Alzheimer Disease Neuroimaging Initiative,	Medical Diagnosis	Uantitative parameters,	Federated Learning (FL) allows the development of deep learning models across multiple centers without direct data sharing, addressing privacy concerns and legal/ethical issues associated with centralized datasets.	Implementing FL can be complex, requiring coordination among multiple centers and setting up secure communication channels. It may also demand substantial computational resources.	Establishing common data acquisition and reconstruction protocols across centers can mitigate data heterogeneity and enhance model generalization.
28	Real Time	Health Care	Local parameters	FL models can be trained on diverse datasets from different clinical centers, leading to more generalizable models that perform well across a variety of imaging protocols and patient populations.	FL involves iterative communication between centers and a central server, which can lead to increased communication overhead and potentially slower convergence compared to centralized training.	Developing efficient communication protocols and strategies for FL can reduce communication overhead and speed up convergence.

Continued

29	Real Time	Health Care	Synthetic datasets aim to mimic real data, ensuring that researchers can still derive valuable insights and conduct meaningful analyses without access to the original, sensitive data.	The synthetic data may not capture fine-grained details present in the original data, potentially limiting certain types of analyses.	Fine-tune synthetic data generation models to specific research objectives and datasets to achieve better mimicry.
30	Real Time	IoHT	Local deep-learning models are trained collaboratively, reducing the need for centralized data collection, which can be time-consuming and resource-intensive.	Implementing and managing a fog-based federated framework can be complex, requiring specialized expertise and infrastructure.	Continuous research and development in privacy-preserving techniques can help bolster security in federated learning, ensuring patient data remains confidential.

- An anonymization model, known as optimal balancing scheduling, based on Map Reduce strategy introduced to overcome scalability. The method assists in enhancing re-anonymization and better handling the problems associated with data locality. However, the model failed to deal with the security issue during privacy preservation.

The analysis based on Data Acquisition, Applications, Parameters, Advantages, Disadvantages, and Solution for the privacy preservation of multi-institutional data are tabulated in Table 5.

Discussion on Laws, Regulations and Policies

In the realm of healthcare data management and privacy protection, adherence to regulations and standards is paramount to safeguard patient information. HIPAA establishes a comprehensive framework for protecting patient health information through security measures, privacy policies, and data breach protocols. GDPR imposes strict privacy requirements on handling personal health data, emphasizing transparency and individual rights [27]. Challenges and Limitations of Cross-Border Data Sharing. Varying global privacy laws complicate harmonizing standards for international data sharing. Stringent data sovereignty laws mandate local storage, hindering seamless cross-border data transfer. Collaboration among healthcare entities and policymakers fosters standardized cross-border data-sharing protocols [26] [29].

5. Conclusion

In this research, various kinds of literature to clarify the issues associated with the privacy preservation of e-Health data have been investigated. Multiple techniques such as encryption, sanitation, and anonymization are evaluated to enhance the privacy of the data transmitted through the IoT platform. A brief

study of the anonymization strategies, optimization algorithms, and blockchain-based strategies used to solve the problems in the privacy preservation of the sensitive medical data of patients has been done. The challenges associated with existing methods are also analyzed in detail to solve the problems with the development of the proposed method. Briefly, the research gaps and issues related to the existing techniques are analyzed and presented in this section. Hence, protecting patient data privacy is a prerequisite for data management and sharing to safeguard the patients whose medical data is shared in clinical research while making information available for future research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Sébastien, Z., Crettaz, C., Kim, E., Skarmeta, A., Bernabe, J.B., Trapero, R. and Bianchi, S. (2019) Privacy and Security Threats on the Internet of Things. In: Ziegler, S., Ed., *Internet of Things Security and Data Protection*, Springer, Berlin, 9-43. https://doi.org/10.1007/978-3-030-04984-3_2

[2] Naser, A.M., Farooque, M.M.J. and Khashab, B.L. (2019) The Effect of Security, Privacy, Familiarity, and Trust on Users' Attitudes toward the Use of the IoT-Based Healthcare: The Mediation Role of Risk Perception. *IEEE Access*, **7**, 111341-111354. https://doi.org/10.1109/ACCESS.2019.2904006

[3] Tejasvi, A., Chamola, V., Sikdar, B. and Choo, K.-K.R. (2020) Consumer IoT: Security Vulnerability Case Studies and Solutions. *IEEE Consumer Electronics Magazine*, **9**, 17-25. https://doi.org/10.1109/MCE.2019.2953740

[4] Swaroop, P. (2016) Internet of Things: Underlying Technologies, Interoperability, and Threats to Privacy and Security. *Berkeley Technology Law Journal*, **31**, 997-1022.

[5] Suneetha, V., Suresh, S. and Jhananie, V. (2020) A Novel Framework Using Apache Spark for Privacy Preservation of Healthcare Big Data. 2020 2*nd International Conference on Innovative Mechanisms for Industry Applications* (*ICIMIA*), Bangalore, 5-7 March 2020, 743-749. https://doi.org/10.1109/ICIMIA48430.2020.9074867

[6] Saleem, Y., Rehmani, M.H., Crespi, N. and Minerva, R. (2021) Parking Recommender System Privacy Preservation through Anonymization and Differential Privacy. *Engineering Reports*, **3**, e12297. https://doi.org/10.1002/eng2.12297

[7] Shruthi, R. and Govindarasu, M. (2020) An Efficient Framework for Privacy-Preserving Computations on Encrypted IoT Data. *IEEE Internet of Things Journal*, **7**, 8700-8708. https://doi.org/10.1109/JIOT.2020.2998109

[8] Dhaval, J. and Panchal, R. (2021) A Novel Anonymity Algorithm for Privacy Preservation. *Elementary Education Online*, **20**, 2402-2402.

[9] Qi, F., He, D., Wang, H., Zhou, L. and Choo, K.-K.R. (2019) Lightweight Collaborative Authentication with Key Protection for Smart Electronic Health Record System. *IEEE Sensors Journal*, **20**, 2181-2196. https://doi.org/10.1109/JSEN.2019.2949717

[10] Shan, J., Cao, J., McCann, J.A., Yang, Y., Liu, Y., Wang, X. and Deng, Y. (2019) Privacy-Preserving and Efficient Multi-Keyword Search over Encrypted Data on Block-

chain. 2019 *IEEE International Conference on Blockchain*, Atlanta, 14-17 July 2019, 405-410.

[11] Xuyun, Z., Liu, C., Nepal, S. and Chen, J. (2013) An Efficient Quasi-Identifier Index Based Approach for Privacy Preservation over Incremental Data Sets on Cloud. *Journal of Computer and System Sciences*, **79**, 542-555. https://doi.org/10.1016/j.jcss.2012.11.008

[12] Dagher, G.G., Mohler, J., Milojkovic, M. and Marella, P.B. (2018) Ancile: Privacy-Preserving Framework for Access Control and Interoperability of Electronic Health Records Using Blockchain Technology. *Sustainable Cities and Society*, **39**, 283-297. https://doi.org/10.1016/j.scs.2018.02.014

[13] Yong, W., Zhang, A., Zhang, P. and Wang, H. (2019) Cloud-Assisted EHR Sharing with Security and Privacy Preservation via Consortium Blockchain. *IEEE Access*, **7**, 136704-136719. https://doi.org/10.1109/ACCESS.2019.2943153

[14] Sheng, C., Wang, J., Du, X., Zhang, X. and Qin, X. (2020) CEPS: A Cross-Blockchain Based Electronic Health Records Privacy-Preserving Scheme. *ICC* 2020-2020 *IEEE International Conference on Communications* (*ICC*), Dublin, 7-11 June 2020, 1-6.

[15] Nuetey, N.R., Yue, L., Agdedanu, P.R. and Adjeisah, M. (2019) Privacy Module for Distributed Electronic Health Records (EHRs) Using the Blockchain. 2019 *IEEE* 4*th International Conference on Big Data Analytics* (*ICBDA*), Suzhou, 15-18 March 2019, 369-374.

[16] Zhang, M.W., Chen, Y. and Susilo, W. (2020) PPO-CPQ: A Privacy-Preserving Opti-mization of Clinical Pathway Query for e-Healthcare Systems. *IEEE Internet of Things Journal*, **7**, 10660-10672. https://doi.org/10.1109/JIOT.2020.3007518

[17] Suman, M. and Goswami, P. (2021) A Technique for Securing Big Data Using k-Anonymization with a Hybrid Optimization Algorithm. *International Journal of Operations Research and Information Systems* (*IJORIS*), **12**, 1-21. https://doi.org/10.4018/IJORIS.20211001.oa3

[18] Jyothi, M. and Rao, C.S. (2019) Privacy Preservation of Data Using Crow Search with Adaptive Awareness Probability. *Journal of Information Security and Applications*, **44**, 157-169. https://doi.org/10.1016/j.jisa.2018.12.005

[19] Rafik, H., Yan, Z., Muhammad, K., Bellavista, P. and Titouna, F. (2020) A Privacy-Preserving Cryptosystem for IoT E-Healthcare. *Information Sciences*, **527**, 493-510. https://doi.org/10.1016/j.ins.2019.01.070

[20] Xue, Y., Lu, R., Shao, J., Tang, X. and Yang, H. (2018) An Efficient and Privacy-Preserving Disease Risk Prediction Scheme for e-Healthcare. *IEEE Internet of Things Journal*, **6**, 3284-3297. https://doi.org/10.1109/JIOT.2018.2882224

[21] Chamikara, M.A.P., Bertok, P., Liu, D., Camtepe, S. and Khalil, I. (2020) Efficient Privacy Preservation of Big Data for Accurate Data Mining. *Information Sciences*, **527**, 420-443. https://doi.org/10.1016/j.ins.2019.05.053

[22] Chen, Z., Zhang, F., Zhang, P., Liu, J.K., Huang, J., Zhao, H. and Shen, J. (2018) Verifiable Keyword Search for Secure Big Data-Based Mobile Healthcare Networks with Fine-Grained Authorization Control. *Future Generation Computer Systems*, **87**, 712-724. https://doi.org/10.1016/j.future.2017.10.022

[23] Saira, K., Iqbal, K., Faizullah, S., Fahad, M., Ali, J. and Ahmed, W. (2020) Clustering Based Privacy Preserving of Big Data Using Fuzzification and Anonymization Operation.

[24] Trisha, D., Apthorpe, N. and Feamster, N. (2018) A Developer-Friendly Library for Smart Home Iot Privacy-Preserving Traffic Obfuscation. *Proceedings of the* 2018 *Workshop on IoT Security and Privacy*, Budapest, 20 August 2018, 43-48.

[25] Fang, R., Pouyanfar, S., Yang, Y., Chen, S.C. and Iyengar, S.S. (2016). Computational Health Informatics in the Big Data Age: A Survey. *ACM Computing Surveys* (*CSUR*), **49**, 1-36.

[26] Abdel-Basset, M., Hawash, H. and Abouhawwash, M. (2022) Collaborative Screening of COVID-19-Like Disease from Multi-Institutional Radiographs: A Federated Learning Approach. *Mathematics*, **10**, 4766.

[27] Gupta, S., Kumar, S., Chang, K., Lu, C., Singh, P. and Kalpathy-Cramer, J. (2023) Collaborative Privacy-Preserving Approaches for Distributed Deep Learning Using Multi-Institutional Data. *RadioGraphics*, **43**, e220107. https://doi.org/10.1148/rg.220107

[28] Shiri, I., Vafaei Sadr, A., Akhavan, A., Salimi, Y., Sanaat, A., Amini, M. and Zaidi, H. (2023) Decentralized Collaborative Multi-Institutional PET Attenuation and Scatter Correction Using Federated Deep Learning. *European Journal of Nuclear Medicine and Molecular Imaging*, **50**, 1034-1050. https://doi.org/10.1007/s00259-022-06053-8

[29] Sun, H., Plawinski, J., Subramaniam, S., Jamaludin, A., Kadir, T., Readie, A. and Coroller, T. (2023) A Deep Learning Approach to Private Data Sharing of Medical Images Using Conditional Generative Adversarial Networks (GANs). *PLOS ONE*, **18**, e0280316. https://doi.org/10.1371/journal.pone.0280316