[1]Dr. B. Srinivasan

Comprehensive and Comparative Analysis of Compliances and Data Privacy Techniques for protecting Data Leakages in Cloud Computing

Journal of Electrical Systems

Abstract: In the present era, much data and operations are being processed online. Today, an individual or a business organization has to utilize the data or a process regardless of where they are working with the help of computers or smart devices. Most notably, the Internet of Things (IoT) allows devices like sensors to transfer data with the help of the Internet. Consequently, accessing data and processes from anywhere is becoming an essential requirement in the modern day. In the beginning ages of computing and even the Internet, data were stored in a centralized database and processed either on a local machine or on the server side. During those periods, users had to shell out vast amounts of money to store data and install software products. However, after the genesis of Cloud Computing (CC), the problem of investing amounts for storage and software installation became optimal. Nowadays, Cloud Companies offer various services to their users with essential security and authentication methods. Nevertheless, they cannot offer any elucidation regarding data privacy and do not provide any promise about data leakage on the cloud. Currently, a plethora of data privacy practices, like data anonymization, pseudonymization, scrambling and masking, etc., are implemented in CC to protect data privacy. This paper discusses a deep analysis of the merits and demerits of those methods, along with case studies and applications.

Keywords: Cloud Computing, Data Privacy, Data Leakage, Pseudonymization, Anonymization, De-identification, Masking, Scrambling, Homomorphic Encryption, Authentication

I. INTRODUCTION

Today, the Internet is essential for a wide range of data processing and operations in the computing era. Individuals or businesses might operate data processing and programming execution as an instant necessity without considering their place. On top of everything else, the Internet of Things (IoT) environment enables the sensors, like devices, to send and receive data through the Internet without the intervention of human help. The proliferation of IoT-enabled devices for gathering data from their environment is good evidence of Internet utilization. The International Data Corporation (IDC) estimates that there will be 41.6 billion IoT devices worldwide by 2025 [1].

From the early dawn of computing and even after the advent of the Internet, data were stored in a centralized place called, a database and processed either locally or on the server side. During those days, clients had to shell out significant amounts of money for storage and the installation of software to process data. Such a type of environment is referred to as local computing or on-premise computing. In that environment, the authorities must install all the in-demand software and stack larger storage by contemplating future uses. However, less software and too much storage are utilized for current needs, tending to be a waste of cost.

After the emergence of Cloud Computing (CC), the provision of software and data needs is facilitated through three services: Software as a Service(SaaS), Infrastructure as a Service (IaaS), and Platform as a Service (PaaS). By taking advantage of any of these services, individual users or business organizations can start accessing software, hardware, and other types of services remotely while scaling up or down as needed. The needy clients only have to pay for the services that they demand, and thus drastically reduce the initial investments and continuous operational costs. CC offers resources, including CPU, system and application software, and storage through the Internet and has become the foremost business model for current business activities. [2].

The present world is experiencing a newer dimension of computing model known as Cloud Computing, which offers on-demand services to clients through the Internet [3]. According to the National Institute of Standards and

[1]Associate Professor of Computer Science
School of Arts and Science, VMRF DU Chennai Campus
Paiyanoor, Vinayaka Mission's Research Foundation
Deemed to be University, Salem, Tamil Nadu, India
srinivasan.avca014a@avsas.ac.in

Technology (NIST), Cloud Computing is defined as a common pool consisting of easily manipulatable programmable grids, servers, software, and storage [4].

Although there are benefits to adopting various types of cloud services, such as private, public, and hybrid, the majority of cloud service users are concerned about security risks. Gartner found that 70% of users are wary about using cloud services due to privacy and security concerns [3]. They aren't yet prepared to ditch their current systems and transfer all of their data to the cloud.

Leading cloud computing environments, including Google, Microsoft Azure, Microsoft Office 365, and Amazon Web Services (AWS), offer a range of services within the three categories of SaaS, PaaS, and IaaS. Additionally, they implement security solutions at host, network, and computation levels to safeguard customer data from potential threats and vulnerabilities.

The primary objective of all such cloud providers with their data security solutions is to prevent the consumer's data from loss, abuse, or any other disclosure. Currently, data privacy considerations in cloud computing extend beyond data security issues.

Data privacy primarily concerns sensitive information about an individual or group that can be either kept secret or shared with authorized parties as and when necessary. Data security checks if a client is an authorized user of a cloud service, but it doesn't guarantee that the client trying to log in is the right person. This leaves the door open for an attacker to get access to a user's account and potentially misuse their credentials. The ultimate consequence of data privacy issues from the side of the cloud service provider is data leakage so that attackers can easily use any data.

Privacy is paramount for preserving the consumer's data. As the world grows more interconnected, more data is kept digitally in a centralized place like a cloud. The following table 1 depicts the difference between data security and data privacy.

TABLE I. DATA SECURITY VS DATA PRIVACY

Data Security	***Data Privacy***
Safeguards the data	Concerns about how information is handled and user
Protects the data from unauthorized usage.	Controls the use of data among users
Enforcing the steps to guarantee data integrity, confidentiality, and availability.	Adhering to the rules regarding the handling of personal data
Encryption and intrusion detection are the key components	Privacy policies and procedures are the primary aspects.
Dealing with issues related to safeguarding information across threats.	The challenges include legal frameworks, shifting rules, and user consent and preferences.

The two most serious issues in the Cloud Computing environment are data breaches and leaks. A data breach occurs when someone from the outside gets or accesses data without permission. This usually happens because of a hack or security breach. In data security, a data breach is considered a major issue and requires many authentication techniques to prevent the data from being accessed unauthorizedly.

Data leaks happen when there is a lapse in data protection through privacy and when unauthorized individuals are permitted to access personal data due to reasons like flawed authentication or data sharing without authorization. One of the major issues in the cloud is data loss due to the occurrence of data leaks [5]. Moreover, people inside and outside of the cloud service can view the data due to the less reliability.

Some of the general strategies to avoid data leakage are listed below

- Organizations must have all the rules to keep the data safe by evaluating and auditing the securities.
- Data access should be limited so that anyone can only view the data, they need to do their job well.
- The old methods of storing data must be updated so that systems tend to have fewer vulnerabilities.
- The Information Technology infrastructure of a company should not accept any unauthorized device easily.
- Most importantly, using multi-factor authentication techniques, reduces the level of accessing the data.
- Offboarding principles should be followed among the employees of a business organization.

Aside from the local measures taken by individual business organizations, various global regularity initiatives have been sought to preserve the privacy of user data. One of the initiatives was taken by the European Union's General Data Protection Regulation

(GDPR), passed on May 25th,2018, to protect the online privacy of European users. By emulating the GDPR, several countries framed their compliances to protect users' online privacy.

This paper comprehensively analyzes the compliances used to protect users' online privacy and the various methods used to protect data leakage in the Cloud Computing environment.

The paper is organized by providing basic information about the CC and the security and privacy issues in the cloud environment under introduction. Users' privacy on the cloud and online is addressed in the second section, along with the data privacy compliances that are adhered to prevent data leaks. Next, a relevant literature survey is reviewed regarding data privacy on the cloud. Then in the fourth section, various anonymization techniques used for protecting the data leakage on the cloud environment are discussed with examples. Following that, a comparative analysis is carried out, and in the paper's final section, the conclusion remarks are presented.

II. DATA LEAKAGE AND PRIVACY COMPLIANCE

Data leakage in cloud computing describes the unlawful release of sensitive information to a third party, whether the person is either inside or outside of a firm. It might have been an accident or someone's deliberate move. Large monetary and non-monetary losses are incurred by an organization as a result of data leakage. Repeated data leakage instances cause increasing anxiety, even though data is an essential asset for any firm[7].

A new breed of cloud computing system called multi-tenant is currently paving the way for a shared virtual pool of computing resources. The ultimate result is that ecosystems are more vulnerable to data privacy issues such as data loss or the theft of personally identifiable information [8]. Nevertheless, storing users' data along with the intricacy of multi-tenant systems raises substantial governance and compliance apprehensions by the European General Data Protection Regulation (EU GDPR) [9].

The concern of data privacy is not a novel one and persisted with the Semayne lawsuit in 1604. As far as data privacy is concerned, everyone's house serves as a fortress. The concept of privacy evolved further and was later emphasized in an article titled "The Right to Privacy," written by Justice Louis Brandeis and Attorney Samuel Warren. This paper recognized the need of protecting individuals' right to privacy as the cornerstone of modern individual liberty [10].

Data privacy regulations were first drafted in 1970 in Germany, among other countries. On May 25, 2018, the EU GDPR was enacted, drastically changing the regulations pertaining to the security and privacy of personal information. Based on the EU GDPR, several countries. Following that, several countries have implemented data privacy regulations in response to EU GDPR, which is highly beneficial for preventing data leaks in cloud environments. The following sections,

The following sections detail the various data privacy compliance frameworks developed by different countries concerning data privacy, which would be an important aspect of securing personal information in the cloud.

A. General Data Protection Regulation (GDPR)

The General Data Protection Regulation (GDPR) is a large piece of data security legislation enacted on May 25, 2018, in the European Union (EU). The GDPR covers all European Union (EU) organizations that collect, process or store personal data about EU citizens. It doesn't matter where these organizations are located. The GDPR says that "personal data" is “any details on a particular or discernible person”[12]. This encompasses not explicit identities like names, IDs, and digital identifiers like IPs, cookies, and device identifiers.

The GDPR lays out seven essential principles for handling personal data [13]. The principles are listed below.

1. Legality, fairness, openness
2. Restrictions on usage
3. Minimization of Data
4. Precision
5. Constraints on storage space
6. Honesty and privacy
7. Responsibility

An individual is deemed a data subject under GDPR, who can be recognized using Personally Identifiable Information (PII), such as name, ID, location, or tangible attributes. The GDPR provides specific entitlements to data subjects, including the right to get the information, utilization, correction, deletion, processing limitation, data transfer, and the right to oppose, etc.[14].

The following significant consequences arise from the viewpoint of cloud computing as an effect of the GDPR.

- Both cloud service providers and their clients may be considered data controllers or processors. This means that they have many duties regarding data subject rights, data security, data retention, reporting breaches, and other things [15].
- As the data controller, the cloud customer must verify legality and GDPR compliance when sending personal data to the cloud. Effective vendor due diligence and contractual safeguards are necessary[11].
- Cross-border data transfer restrictions apply to EU-originating personal data transferred to countries lacking adequate data protection. exclusions, Standard Contractual Clauses, and Binding Corporate Rules may still allow transfers in certain situations[16].
- Cloud providers must promptly notify clients of any data breaches and report breaches to regulatory agencies, and data subjects must comply with strict time limits[17].
- Data controllers must perform Data Protection Impact Assessments (DPIAs) whenever they handle data that could substantially endanger the rights and freedoms of individuals, such as when they transfer sensitive data to the cloud[18].

B. Data Compliances in different countries.

In compliance with the EU GDPR, several nations have developed strategies to safeguard data stored in the cloud. The California Consumer Privacy Act (CCPA), established on January 1, 2020, is widely regarded as one of the most extensive privacy laws at the state level in the United States

The CCPA empowers California residents with rights similar to those in the GDPR, such as access to personal information, deletion requests, opt-outs, and non-discrimination [19].

The Digital Personal Data Protection Act (DPDP Act) was passed in August 2023 in India and aligns people's entitlement to protecting their personal data with the necessity to process such data for legal purposes. The Act offers information about the rights and responsibilities of Data Principals, the people whose data it is, and instructions for Data Fiduciaries, the people who handle data. It also establishes penalties for disobeying the

law. The DPDP act is a watershed point in the Indian privacy law. The new paradigm thoughtfully balances the benefits of technological advancements with the imperative of protecting individuals' privacy [20].

Investing in data privacy has multiple benefits, such as cultivating user confidence and bolstering one's reputation. Given the growing popularity of online sharing of personal information, it is imperative to prioritize its security and prevent unauthorized use or abuse. A number of data compliance procedures must be put into practice to prevent data leakage on clouds.

The present study explores various methodologies for preserving data privacy, including data anonymization, pseudonymization, and scrambling. It also outlines potential avenues for future research into cloud data leak protection and proposes a comparative analysis of the discussed methodologies with their benefits and drawbacks.

III. LITERATURE REVIEW

Data leakages, also popularly called data breaches, occur when someone exploits a system's weaknesses to extract information [21]. In addition to technology solutions at the system level, data leakage is acknowledged as an administrative and organisational issue[22].

Data breaches pose a significant problem in the corporate realm and numerous other institutions. Data distributors deliberately incorporate deceptive entities into the data they transmit to agents during discussions. By incorporating deceptive elements into the scattered data, data distributors can improve the effectiveness of identifying guilty agents. The users of these systems are in danger due to the vulnerability of algorithms used for leak detection [23]. Many researchers have extensively researched the subject of data leaking.

Mehrtak, Mohammad, et al. reviewed the security challenges in cloud computing and suggested solutions. Furthermore, they suggested utilizing data encryption for cloud storage and retrieval to ensure data access [24]. Murthy, Ch VNU Bharathi, et al emphasized that integrating blockchain technology with a scaled cloud environment improves trust, server performance, data protection, and user data administration [25]. Hathaliya et al. gave valuable insights into the data security and privacy concerns in Healthcare 4.0[26]. Gupta et al. provided a holistic view of data protection in communication and sharing contexts, applicable to any sort of organization [27]. Saleem et al. outlined the most dependable sequence of steps for ten widely acknowledged instances of data breaches[28]. The early discoveries and a series of obstacles in data leakage were identified by Neto, Nelson Novaes, et al [29]. Silva et al. thoroughly evaluated privacy principles, risks, and policies and assessed their relevance to cloud computing [30].

Kangwa Mukuka et al. developed an algorithm to produce random Pseudo IDs to protect online shoppers' anonymity [31]. Abd Razak et al. introduced a novel technique for generating pseudonyms on anonymized data with the aim of ensuring the preservation of information security[32]. Sahana Lokesh and Ranganatha suggested a framework to preserve medical record privacy using the African Vultures Optimization Algorithm (AVOA) and genetic algorithms with simulated annealing [33]. Ahmadi Sina emphasized the security and privacy hurdles in cloud-based data warehousing utilized by private and government entities [34].

Varshney, Shipra, et al. provided a comprehensive overview of data breach prevention methods at various big data life cycle stages. Anonymization was implemented to prevent data misuse[35]. Tachepun Chitanut and Sotarat Thammaboosadee provided tokenization, obfuscation, and anonymization to protect personal data through data classification and risk assessment [36]. Abrera Joseph stated GDPR, HIPAA, and CCPA regulate cloud data privacy, requiring strict handling and protection [37]. Gao Lei et al. conducted a content analysis, and the results can benefit regulators and corporations by reducing GDPR and GDPR-covered individuals [38].

IV. PERSONAL DATA PROTECTION TECHNIQUES

There is an abundance of data being collected, saved, and used all around the world these days. Among the many pieces of information included are occupations, locations of employment, hobbies, purchasing habits, vacation spots, pets, and much more. Most of this information is shared through social media, and it seems that these platforms are making money by selling user data to interested buyers with the help of clouds to some extent, which could lead to legal consequences.

Significant security and privacy concerns impede the widespread use of public clouds. Most businesses understandably hesitate to entrust such unregulated spaces with sensitive data, such as customers' private information. Data privacy legislation such as the EU GDPR and CCPA enable clients to exercise their entitlement to anonymity. When an organization starts using people's data, it becomes progressively difficult to make a clear connection between the data and the individuals it represents.

Various tactics have been used in cloud computing to improve the security of personal data and protect individuals' privacy. Common methods include data anonymization, pseudonymization, masking, and hybrid encryption. The following sections will explain these strategies with illustrative examples. After discussing all the methods in detail, a comprehensive comparative analysis will be conducted using the results obtained from each technique covered.

A. Data Anonymization

Data anonymization is a non-cryptographic technique that entails completely eliminating or annihilating any trace of PII from a dataset to ensure an individual's anonymity. Data anonymity involves modifying personal data to prevent direct or indirect identification of the individual. It also ensures that the cloud service provider (CSP) cannot access any personal information [39]. The process of anonymizing data is depicted in Fig. 1.

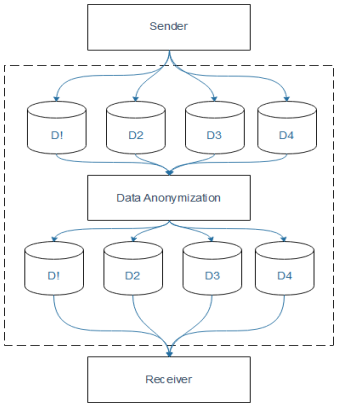

Fig. 1 The Data Anonymization Process

Data anonymization is a prevalent practice for externalizing confidential data. For instance, banks analyze client transactions to identify cases of fraudulent activity. Doctors require comprehensive health data to diagnose and treat patients accurately, while testers are tasked with evaluating software and ensuring that the genuine data remains unaffected. Additionally, businesses utilize client feedback to enhance their products and services.

An empirical study was conducted by utilizing patient healthcare data to exemplify a core principle of data anonymization called de-identification. The study employed a dataset including 55,501 patient records obtained from the Kaggle [40]. The primary goal of this study is to demonstrate the application of data anonymization in preserving PII in the provided dataset and allowing for further processing.

The case study first finds all the PII in the given dataset and then gets rid of them. Only the data needed for healthcare-related data analysis and modeling applications is kept in the dataset, and the entire process is done with the help of Python code. Fig. 2 shows the whole structure of the dataset before de-identification. The dataset consists of 15 columns, including Name, Age, Gender, Blood Type, Room Number, and Hospital. These columns contain PII that can be used to identify patients who underwent treatment in a particular hospital. According to EU GDPR Act 4(1) [12], any data that can identify an individual is treated as PII, and that data is to be hidden from the public point of view to preserve privacy.

According to NIST, PII refers to data that can be used to distinguish or trace an individual's identity, either on its own or when combined with other information linked to or able to be linked to a specific person [41]. Such information is to be eliminated to protect an individual's privacy. In Fig. 2, all the PII in the given dataset are highlighted with the box.

#	Column	Non-Null Count	Dtype
0	Name	55500 non-null	object
1	Age	55500 non-null	int64
2	Gender	55500 non-null	object
3	Blood Type	55500 non-null	object
4	Medical Condition	55500 non-null	object
5	Date of Admission	55500 non-null	object
6	Doctor	55500 non-null	object
7	Hospital	55500 non-null	object
8	Insurance Provider	55500 non-null	object
9	Billing Amount	55500 non-null	float64
10	Room Number	55500 non-null	int64
11	Admission Type	55500 non-null	object
12	Discharge Date	55500 non-null	object
13	Medication	55500 non-null	object
14	Test Results	55500 non-null	object

Fig 2 Patients Details with PII

The highlighted columns in Fig. 2 are eliminated to preserve the privacy of patients who took treatment, and after the removal, the dataset with non-PII is shown in Fig.3

#	Column	Non-Null Count	Dtype
0	Gender	10 non-null	object
1	Doctor	10 non-null	object
2	Hospital	10 non-null	object
3	Insurance Provider	10 non-null	object
4	Billing Amount	10 non-null	float64
5	Discharge Date	10 non-null	object
6	Medication	10 non-null	object
7	Test Results	10 non-null	object

Fig. 3 Patients Details without PII

By removing PII from the dataset, an individual's confidentiality is preserved, allowing for the execution of essential processes such as healthcare data analysis. The removal of PII will not affect the accuracy of non-PII data subjected to further processing. The performance analysis of the dataset with mean and median over the first 500 data, before the de-identification of PII is shown in Fig. 4

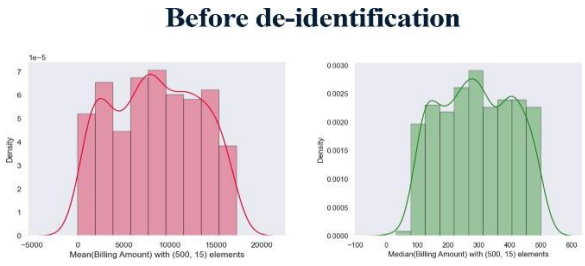

Fig. 4 Performance Analysis before de-identification

The statistical measures being used are the mean and median of the Billing amount column. This field is of float type and is also non-PII. The mean value represents the central tendency of the data. At the same time, the median is more suitable for skewed data or data with outliers, as it is less affected by extreme values. After removing PII from the dataset, the performance analysis can be seen in Fig. 5.

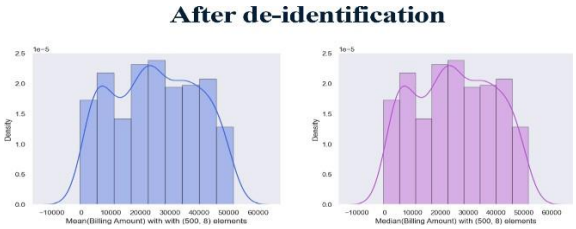

Fig. 5 Performance Analysis after de-identification

From Fig. 4 and Fig. 5, it is clear that the mean and median values of the billing amount remain unchanged, and the density distribution curves are the same before and after the PII are removed from the dataset. The main problem with de-identifying PII is that it cannot be reverted once it is de-identified. In addition, there should be no confusion about whether specific data should be classified as PII or non-PII, leading to a significant decrease in result accuracy.

B. Data Masking

Data masking, also known as obfuscation, hides sensitive information by applying a unique pattern or sequence of numbers to remove any identifying connections between individuals and stored data effectively. The technique aims to thwart illicit access to or exploitation of sensitive information like PII, financial data, and ownership rights [42]. Data masking involves more than just the use of algorithms; it also focuses on analyzing and interpreting public datasets[43]. A simple data masking is illustrated in Fig. 6

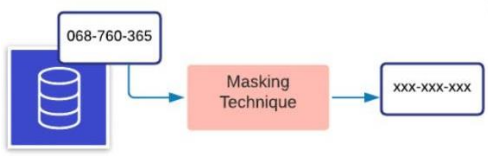

Fig. 6 A simple data masking technique

Data masking can be categorized into two primary types: dynamic and static. Dynamic data masking involves hiding sensitive information in real-time while programs or users access it. This approach is very helpful for protecting data in operational situations without affecting the functionality of the applications.

Static data masking involves creating a disguised copy of the original data, which can be used for testing, development, and analytics. This approach is often used when sharing data with others or using data in a different environment from the usual work setting. Data masking is commonly used in non-production environments such as software development, testing, staff training, and credit or debit card processing.

In order to deepen the understanding of data masking, an experimental case study has been conducted using the same healthcare dataset[40], used in the data anonymization section's de-identification process. Now, a string of character patterns is used to mask the PII during their usage rather than remove them from the dataset. Dynamic data masking is applied over the selected dataset by replacing all numerical values with '*' and non-numerical values with 'X'.

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication
0	Bobby JacksOn	30	Male	B-	Cancer	1/31/2024	Matthew Smith	Sons and Miller	Blue Cross	18856.28131	328	Urgent	2/2/2024	Paracetamol
1	LesLie TErRy	62	Male	A+	Obesity	8/20/2019	Samantha Davies	Kim Inc	Medicare	33643.32729	265	Emergency	8/26/2019	Ibuprofen
2	DaNnY sMitH	76	Female	A-	Obesity	9/22/2022	Tiffany Mitchell	Cook PLC	Aetna	27955.09608	205	Emergency	10/7/2022	Aspirin
3	andrEw waTtS	28	Female	O+	Diabetes	11/18/2020	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.78241	450	Elective	12/18/2020	Ibuprofen
4	adrIENNE bEll	43	Female	AB+	Cancer	9/19/2022	Kathleen Hanna	White-White	Aetna	14238.31781	458	Urgent	10/9/2022	Penicillin
5	EMILY JOHNSOn	36	Male	A+	Asthma	12/20/2023	Taylor Newton	Nunez-Humphrey	UnitedHealthcare	48145.11095	389	Urgent	12/24/2023	Ibuprofen
6	edwArD EDWaRDs	21	Female	AB-	Diabetes	11/3/2020	Kelly Olson	Group Middleton	Medicare	19580.87234	389	Emergency	11/15/2020	Paracetamol
7	CHrisTinA MARtinez	20	Female	A+	Cancer	12/28/2021	Suzanne Thomas	Powell Robinson and Valdez,	Cigna	45820.46272	277	Emergency	1/7/2022	Paracetamol
8	JASmINe aGuIlaR	82	Male	AB+	Asthma	7/1/2020	Daniel Ferguson	Sons Rich and	Cigna	50119.22279	316	Elective	7/14/2020	Aspirin
9	ChRISTopher BerG	58	Female	AB-	Cancer	5/23/2021	Heather Day	Padilla-Walker	UnitedHealthcare	19784.63106	249	Elective	6/22/2021	Paracetamol

Fig. 7 Healthcare dataset before masking

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date
0	XXXXXXXXXXXX	**	Male	XX	XXXXXX	*/**/****	Matthew Smith	Sons and Miller	Blue Cross	18856.28131	***	XXXXXX	2/2/2024
1	XXXXXXXXXXXX	**	Male	XX	XXXXXXX	*/**/****	Samantha Davies	Kim Inc	Medicare	33643.32729	***	XXXXXXXX	8/26/2019
2	XXXXXXXXXX	**	Female	XX	XXXXXXX	*/**/****	Tiffany Mitchell	Cook PLC	Aetna	27955.09608	***	XXXXXXXX	10/7/2022
3	XXXXXXXXXXX	**	Female	XX	XXXXXXXX	**/**/****	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.78241	***	XXXXXXX	12/18/2020
4	XXXXXXXXXXX	**	Female	XXX	XXXXXX	*/**/****	Kathleen Hanna	White-White	Aetna	14238.31781	***	XXXXXX	10/9/2022
5	XXXXXXXXXXX	**	Male	XX	XXXXXX	**/**/****	Taylor Newton	Nunez-Humphrey	UnitedHealthcare	48145.11095	***	XXXXXX	12/24/2023
6	XXXXXXXXXXXX	**	Female	XXX	XXXXXXXX	**/*/****	Kelly Olson	Group Middleton	Medicare	19580.87234	***	XXXXXXXX	11/15/2020
7	XXXXXXXXXXXXXXXX	**	Female	XX	XXXXXX	**/**/****	Suzanne Thomas	Powell Robinson and Valdez,	Cigna	45820.46272	***	XXXXXXXX	1/7/2022
8	XXXXXXXXXXXXX	**	Male	XXX	XXXXXX	*/**/****	Daniel Ferguson	Sons Rich and	Cigna	50119.22279	***	XXXXXXX	7/14/2020
9	XXXXXXXXXXXXX	**	Female	XXX	XXXXXX	*/**/****	Heather Day	Padilla-Walker	UnitedHealthcare	19784.63106	***	XXXXXXX	6/22/2021

Fig. 8 Healthcare dataset after masking

Data masking is an effective strategy that preserves PII with other non-PII inside the dataset. After all, PII is masked with certain patterns, and the process's performance utilizing the dataset remains unaffected. However, data masking becomes ineffective when PII, such as social security numbers, credit card numbers or an individual's age, is required for analysis.

Fig. 9 presents the line plot of the first 500 data points from the dataset before applying any data masking. Since these are numeric variables, three lines have been drawn for Billing Amount, Age, and Room Number.

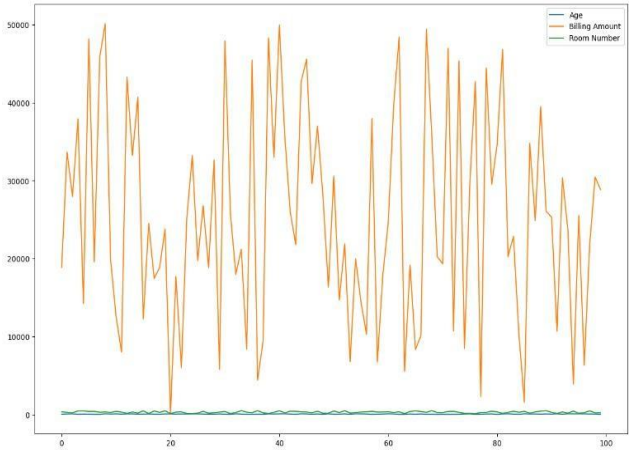

Fig. 9 Line plot before data masking

A larger spread represents the billing amount, while the lines at the bottom represent the Age and Room Number.

Fig. 10 displays the same dataset's line plot after masking of PII columns, Age, and Room Number.

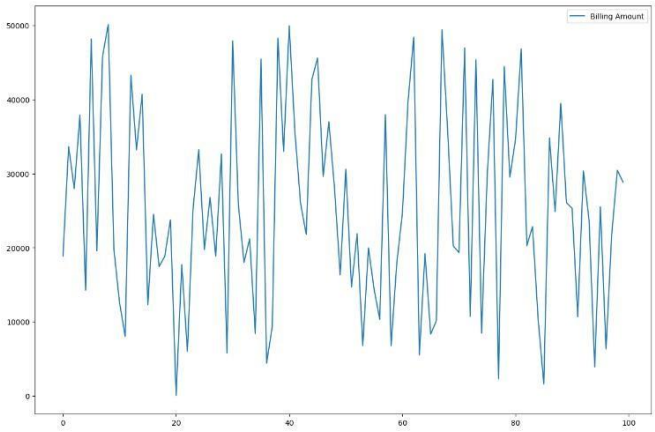

Fig. 10 Line plot after data masking

A single line is presented for the non-PII column Billing Amount in Fig. 10, as the lines for PII elements, Age and Room Number, were omitted due to their masking with '*'. From Fig. 9 and Fig.10, it is observed that there could be no change in the used non-PII column Billing Amount because, on both line plots, the shape of the column remains the same. Data masking is applied to the PII of the given dataset and utilized by those who accessed the entire dataset.

C. Pseudonomyzation

Pseudonymization is a data privacy technique employed to maintain the anonymity of individuals in the provided data. According to the EU GDPR, pseudonymization refers to the processing of personal data in a manner that can no longer be pointed out to an individual without any supplementary information [12]. Pseudonymization is the process of substituting one data attribute for another to make the data less identifiable [44]. According to Article 32(1) of the EU GDPR, pseudonymization is the first and foremost measure to be taken to ensure the safety of users' personal information[45]. The entire pseudonymization for the healthcare process is represented in Fig. 11.

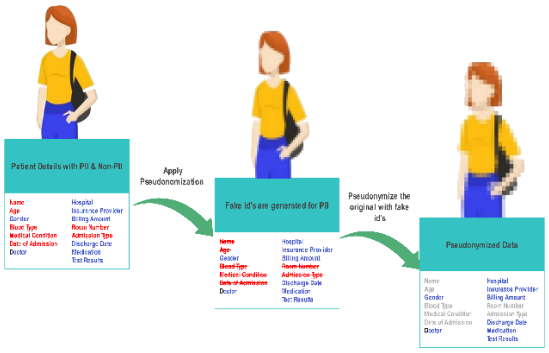

Fig.11 the Pseudonymization process

In the aforementioned Fig. 11, the PII is initially detected, and corresponding fake IDs are produced. The fabricated IDs are substituted with the PII, resulting in a dataset with counterfeit IDs corresponding to the PII and authentic non-PII data. The pseudonymization method is elucidated through a case study utilizing the healthcare dataset referenced in earlier de-identification and data masking sections.

After applying fake IDs, the pseudonymized version of the dataset in Fig. 7 is illustrated in Fig. 12.

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medicati
0	Lisa Thomas	3753	S8-1437388L	D0-6988050m	K2-5599517C	R6-8152053l	Matthew Smith	Sons and Miller	Blue Cross	18856.281310	L1-362902G	J2-0764757n	2/2/2024	Paracetan
1	Sheila Gibson	776	u5-9169827G	N7-5027261t	I3-0041215F	T4-1357905E	Samantha Davies	Kim Inc	Medicare	33643.327290	z2-5281776w	j8-3764924K	8/26/2019	Ibuprof
2	Michelle Cohen	9244	n1-482449Z	v5-1652130d	R5-1088387m	y0-5548189E	Tiffany Mitchell	Cook PLC	Aetna	27955.096080	U2-7532134u	b5-8519383u	10/7/2022	Aspi
3	Aaron Wilcox	7264	F0-1578210x	t8-90575R	B8-3414075S	X6-0691084U	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	Y5-0784347M	V2-7879407t	12/18/2020	Ibuprof
4	Zachary Skinner	1783	P0-4973211R	x0-8546000	P7-5807804A	p6-4021161g	Kathleen Hanna	White-White	Aetna	14238.317810	G7-6412365K	c9-3653667p	10/9/2022	Penici

Fig.12 Healthcare dataset after pseudonymization

The pseudonymized data from Fig. 12 will not affect the dataset's application in subsequent healthcare analytical procedures until any pseudonymized PII is needed for further processing.

The visual plotting of the numerical data columns, namely, Age, Billing Amount, and Room Number before pseudonymization, is illustrated in Fig. 13

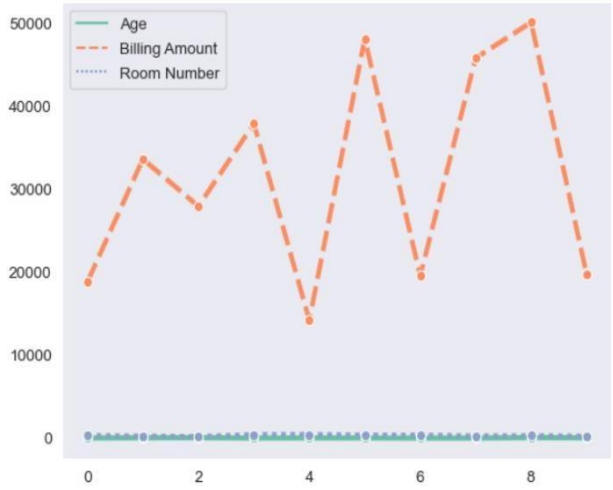

Fig.13 Visual plotting of Age, Billing Amount, and Room Number before Pseudonymization.

In Fig. 13 above, the major deviated line is plotted for the Billing Amount and is the non-PII. There is no significant deviation for the PII columns Age and Room Number, so their lines have been plotted as straight. The equivalent visualization of the dataset after pseudonymization is represented in Fig. 14 with the same PII and non-PII columns as in Fig.13.

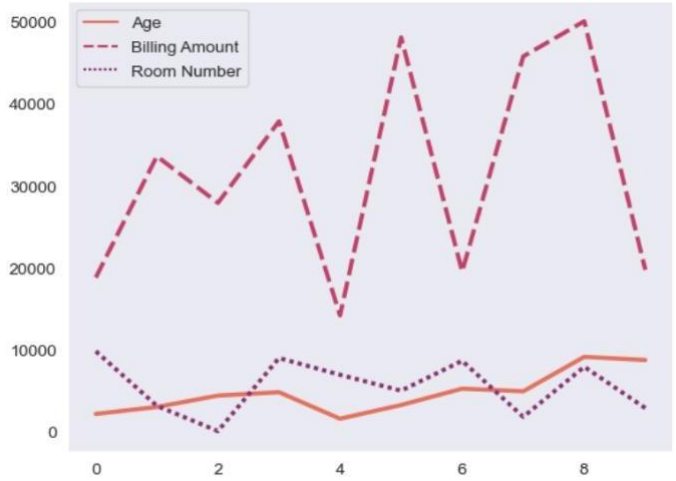

Fig.14 Visualization of PIIs and non-PIIs after Pseudonymization.

The non-PII data column, Billing Amount, represented in Fig. 14, remains the same as in Fig. 13. However, the PIIs' Age and Room Number are pseudonymized with fake IDs, and their plotting has significant changes. The primary advantage of the pseudonymization process is its reversibility, allowing for the restoration of original data after of pseudonymizing data. The Pseudonymization approach is mostly utilized in the financial and healthcare industries to safeguard individual privacy. Many cloud computing environments have begun implementing this approach to preserve client privacy.

Another kind of pseudonymization is data scrambling, in which the characters in the provided value are randomly shuffled. Data scrambling is a data masking technique employed to rearrange characters to conceal the actual information. The syntax of the alphabetic and numeric characters is rearranged in the actual dataset[46]. The status of applying scrambling over the same healthcare dataset is depicted in Fig. 15.

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medicatio
0	bocB kJasnbOy	30	elMa	-B	Ccrena	/321/1024	Matthew Smith	Sons and Miller	Blue Cross	18856.281310	283	rtengU	2/2/2024	Paracetamc
1	eTLryRiseL E	26	aleM	+A	btsieOy	220/0/918	Samantha Davies	Kim Inc	Medicare	33643.327290	652	ecngmryEe	8/26/2019	Ibuprofe
2	Matni DsYHN	67	eemaFl	A-	sbieOyt	2029/22/2	Tiffany Mitchell	Cook PLC	Aetna	27955.096080	502	ygmrenEce	10/7/2022	Aspiri
3	ETdaSrwnat w	82	eFmlea	O+	beatsieD	0211810/2/	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	tveEclie	12/18/2020	Ibuprofe
4	NNIlrbElEd aE	43	Femlae	AB+	craCen	91/29022/	Kathleen Hanna	White-White	Aetna	14238.317810	485	enUgtr	10/9/2022	Penicilli
...	...	...	...	...	...	...	...	...	...	...	...	...	...	

Fig.15 Scrambling over the healthcare dataset

Fig. 15 shows that both PII and non-PII are retained in the same dataset after scrambling and produce the same results as in the case of pseudonymization with fake IDs. The scrambling technique is weaker because, for some general PII, like age and blood type, the original data can easily be guessed by anyone after several trial-and-error methods.

D. Homomorphic Encryption(HE)

Homomorphic Encryption (HE) was introduced in 1978 by Ron Rivest, Adi Shamir, and Leonard Adleman to combat the authority challenges in the cloud environment. HE converts data into encrypted text, allowing a system to perform actions on encrypted data without needing access to the confidential decryption key; the data owner maintains exclusive control of the secret key. Arithmetic operations on encrypted data produce results congruent to those derived from unencrypted data [47]. The stages involved in the HE are illustrated in Fig. 16.

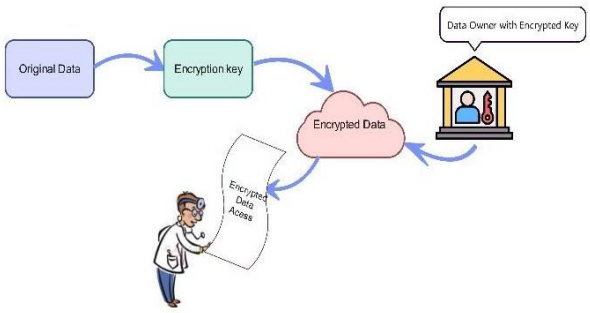

Fig.16 The Homomorphic Encryption Process

First, the original data is turned into ciphertext using a standard encryption method and a key, as shown in Fig. 16. The ciphertext is kept in the cloud, and only the data owner can access the data and provide the key for decryption. Other users can manipulate the encrypted data without decryption, maintaining data privacy. Homomorphic encryption has three categories: Partial Homomorphic Encryption (PHE), Some What Homomorphic Encryption (SWHE), and Fully Homomorphic Encryption. PHE enables addition or multiplication, representing a single action on encrypted data. SWHE supports a restricted number of operations, namely assessing the circuit to a certain threshold or depth. FHE enables the execution of both operations on encrypted data at an infinite number of times[48].

A simple case study for homomorphic encryption has been carried out using the healthcare dataset used in the previous sections. Fig. 17 illustrates the result of applying HE to the dataset shown in Fig. 7.

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date
0	Lyli TkmucYx	206000	Wkvo	L-	Mkxmob	1/31/2024	Matthew Smith	Sons and Miller	Blue Cross	18856.281310	265600	Ebqaxd	2/2/2024
1	VocVso DObBi	212400	Wkvo	K+	Yiocsdi	8/20/2019	Samantha Davies	Kim Inc	Medicare	33643.327280	253000	Owcbqoxmi	8/26/2019
2	NkXxl cWsdR	215200	Powkvo	K-	Yiocsdi	9/22/2022	Tiffany Mitchell	Cook PLC	Aetna	27955.096080	241000	Owcbqoxmi	10/7/2022
3	lxnbOg gkDdC	205600	Powkvo	Y+	Nskkodoc	11/18/2020	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	290000	Ovomdsfo	12/18/2020
4	lxnbSOXXO lOvv	208600	Powkvo	KL+	Mkxmob	9/19/2022	Kathleen Hanna	White-White	Aetna	14238.317810	291600	Ebqaxd	10/9/2022

Fig . 17 Healthcare dataset after Homomorphic Encryption

The numerical PII, such as Age and Room Number, is encrypted with addition and multiplication circuits, while the non-numerical PII, such as Name, Gender, Blood Type, etc., is encrypted with the Caeser cryptography technique. The same techniques are used to decrypt the encrypted text for further processing. Fig. 18 displays the visual representation of numerical PII and non-PII in the healthcare dataset before homomorphic encryption.

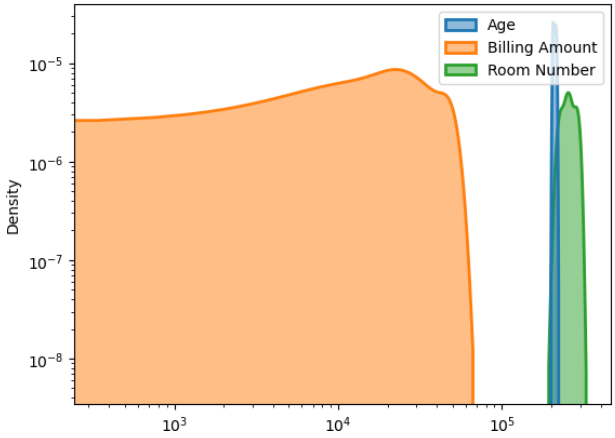

Fig . 18 Representation of numerical PII and non-PII before HE

The numerical PII Age and Room Number have a minimum range of values; hence, they are represented as slightly dense curves, and the non-PII Billing Amount has broader values, so it is represented as a widely spread curve.

Fig. 19 illustrates the visual representation of identical numerical PII and non-PII after applying HE.

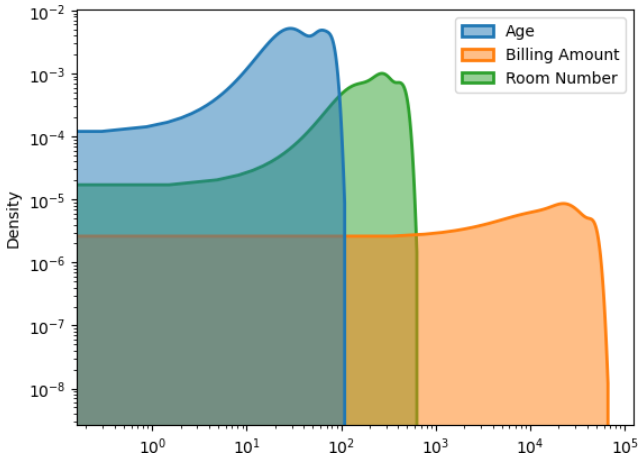

Fig . 19 Plotting of numerical PII and non-PII after HE

Fig. 19 shows significant PII Age and Room Number charting changes after applying HE to the healthcare dataset. This occurs because basic encoding is implemented on both datasets using addition and multiplication circuit operations. The non-PII Billing Amount saw no alterations; hence, its form remains unchanged.

Homomorphic encryption is mostly employed in cloud computing, data analytics, healthcare, and artificial intelligence. The principal challenges of HE encompass suboptimal performance resulting from its processing demands and latency, the requirement for enhanced proficiency in the encryption and decryption techniques deployed, and the compatibility of the employed schemes.

. COMPARATIVE ANALYSIS

The previous two sections discussed various data compliances and data protection techniques for preserving individuals' privacy. This section provides a detailed comparative analysis of the protection schemes discussed. The methods discussed have their properties, applications, advantages and disadvantages. All the concepts are summarized in Table. 2.

TABLE II. COMPARATIVE ANALYSIS OF DISCUSSED DATA PROTECTION TECHNIQUES

Name	***Advantages***	***Disadvantages***
De-identification (Anonymization)	PII is removed for privacy	Irreversible process
Data masking	Protects privacy using patterns	Performance slowed. Needs significant patterns.
Fake IDs (Pseudonymization)	Replaces PII with fake IDs for privacy, reversible.	High computing complexity and poor reversal.
Scrambling (Pseudonymization)	Protects privacy by interchanging PII characters	Weak approach, readily guessed
Homomorphic Encryption	Highly protect privacy	Pricey and easily guessed.

Cloud environments employ one of the discussed methods, contingent upon the data applications and their respective advantages and disadvantages. Still, the field of protecting privacy in such environments is open in data privacy research in cloud computing.

VI. CONCLUSION

The present paper presented a detailed, comprehensive, and comparative analysis of the compliance and data protection techniques used to preserve privacy in cloud computing environments. Standard compliances like EU GDPR, CCPA, and DPDP have been reviewed in terms of data privacy in cloud environments. Data privacy is now only recognized and given importance in online environments. The differences between data security and data privacy are discussed clearly. The current work discussed the available data protection techniques like data anonymization, masking, pseudonymization, and Homomorphic Encryption with case studies and compared their merits and demerits. The work provides basic knowledge about data privacy and data leakage problems in cloud computing. It opens a way for researchers to move towards new solutions for protecting the privacy of individuals in online environments like clouds.

REFERENCES

[1] Yang, Pan, Naixue Xiong, and Jingli Ren. "Data security and privacy protection for cloud storage: A survey." Ieee Access 8 (2020): 131723-131740.

[2] Sunyaev, Ali, and Ali Sunyaev. "Cloud computing." Internet computing: Principles of distributed systems and emerging internet-based technologies (2020): 195-236.

[3] Mohammed, Saja J., and Dujan B. Taha. "From cloud computing security towards homomorphic encryption: A comprehensive review." TELKOMNIKA (Telecommunication Computing Electronics and Control) 19.4 (2021): 1152-1161.

[4] Akbar, Hussain, Muhammad Zubair, and Muhammad Shairoze Malik. "The security issues and challenges in cloud computing." International Journal for Electronic Crime Investigation 7.1 (2023): 13-32.

[5] Alouffi, Bader, et al. "A systematic literature review on cloud computing security: threats and mitigation strategies." Ieee Access 9 (2021): 57792-57807.

[6] Lukic, Karlo, Klaus M. Miller, and Bernd Skiera. "The impact of the General Data Protection Regulation (GDPR) on online tracking." Available at SSRN 4399388 (2023).

[7] Nayak, Suvendu Kumar, and Ananta Charan Ojha. "Data leakage detection and prevention: Review and research directions." Machine Learning and Information Processing: Proceedings of ICMLIP 2019 (2020): 203-212.

[8] Barati, Masoud, et al. "Privacy-aware cloud auditing for GDPR compliance verification in online healthcare." IEEE Transactions on Industrial Informatics 18.7 (2021): 4808-4819.

[9] Russo, Barbara, et al. "Cloud computing and the new EU general data protection regulation." IEEE Cloud Computing 5.6 (2018): 58-68.

[10] Volini, Anthony G. "Right to Data Privacy: Revisiting Warren & Brandeis." Nw. J. Tech. & Intell. Prop. 21 (2023): 1.

[11] Kanungo, Satyanarayan. "Data Privacy and Compliance Issues in Cloud Computing: Legal and Regulatory Perspectives." International Journal of Intelligent Systems and Applications in Engineering (IJISAE) 12.21s (2024): 1721-1734.

[12] General Data Protection Regulation (GDPR), art. 4(1), 2016 O.J. (L 119) 1.

[13] General Data Protection Regulation (GDPR), art. 5, 2016 O.J. (L 119) 1.

[14] General Data Protection Regulation (GDPR), arts. 12-23, 2016 O.J. (L 119) 1.

[15] European Data Protection Board. (2021). Guidelines 07/2020 on the concepts of controller and processor in the GDPR.https://edpb.europa.eu/our-work-tools/ourdocuments/guidelines/guidelines-072020-concepts-controller-and-processor-gdpr_en

[16] General Data Protection Regulation (GDPR), arts. 44-50, 2016 O.J. (L 119) 1.

[17] General Data Protection Regulation (GDPR), arts. 33-34, 2016 O.J. (L 119) 1.

[18] General Data Protection Regulation (GDPR), art. 35, 2016 O.J. (L 119) 1.

[19] Cal. Civ. Code §§ 1798.100, 1798.105, 1798.110, 1798.115, 1798.120, 1798.125.

[20] Kashyap, Pradip Kumar. "DIGITAL PERSONAL DATA PROTECTION ACT, 2023: A NEW LIGHT INTO THE DATA PROTECTION AND PRIVACY LAW IN INDIA." ICREPJOURNAL OF INTERDISCIPLINARY STUDIES , Volume 2, Issue 1, 2023

[21] Schlackl, Frederic, Nico Link, and Hartmut Hoehle. "Antecedents and consequences of data breaches: A systematic review." Information & Management 59.4 (2022): 103638.

[22] Khan, Freeha, et al. "Data breach management: An integrated risk model." Information & Management 58.1 (2021): 103392.

[23] Singh, Rishabh, and V. Gokul Rajan. "Data Leakage and Security on Cloud Computing." Int J Sci Dev Res (IJSDR) 12.4 (2022): 967-973.

[24] Mehrtak, Mohammad, et al. "Security challenges and solutions using healthcare cloud computing." Journal of medicine and life 14.4 (2021): 448.

[25] Murthy, Ch VNU Bharathi, et al. "Blockchain based cloud computing: Architecture and research challenges." IEEE access 8 (2020): 205190-205205.

[26] Hathaliya, Jigna J., and Sudeep Tanwar. "An exhaustive survey on security and privacy issues in Healthcare 4.0." Computer Communications 153 (2020): 311-335.

[27] Gupta, Ishu, and Ashutosh Kumar Singh. "A holistic view on data protection for sharing, communicating, and computing environments: Taxonomy and future directions." arXiv preprint arXiv:2202.11965 (2022).

[28] Saleem, Hamza, and Muhammad Naveed. "Sok: Anatomy of data breaches." Proceedings on Privacy Enhancing Technologies (2020).

[29] Neto, Nelson Novaes, et al. "Developing a global data breach database and the challenges encountered." Journal of Data and Information Quality (JDIQ) 13.1 (2021): 1-33.

[30] Silva, Paulo, Edmundo Monteiro, and Paulo Simoes. "Privacy in the cloud: A survey of existing solutions and research challenges." IEEE access 9 (2021): 10473-10497.

[31] Kangwa, Mukuka. Prevention of personally identifiable information leakage in ecommerce using offline data minimization and online pseudonymisation. Diss. The University of Zambia, 2023.

[32] Abd Razak, Shukor, Nur Hafizah Mohd Nazari, and Arafat Al-Dhaqm. "Data anonymization using pseudonym system to preserve data privacy." Ieee Access 8 (2020): 43256-43264.

[33] Sahana, Lokesh R., and H. R. Ranganatha. "Efficient Data Anonymization Approach to Preserve Privacy of Sensitive Data In Cloud Storage." NeuroQuantology 20.17 (2022): 850.

[34] Ahmadi, Sina. "Security And Privacy Challenges in Cloud-Based Data Warehousing: A Comprehensive Review." International Journal of Computer Science Trends and Technology (IJCST)–Volume 11 (2023).

[35] Varshney and Shipra "Big data privacy breach prevention strategies." 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC). IEEE, 2020.

[36] Tachepun, Chitanut, and Sotarat Thammaboosadee. "A Data masking guideline for optimizing insights and privacy under GDPR compliance." Proceedings of the 11th international conference on advances in information technology. 2020.

[37] Abrera, Joseph. "Data Privacy and Security in Cloud Computing: A Comprehensive Review." Journal of Computer Science and Information Technology 1.1 (2024): 01-09.

[38] Gao, Lei, C. Kevin Eller, and Austin F. Eggers. "GDPR and the cloud: examining readability deficiencies in cloud computing providers' privacy policies." Policy Studies 44.6 (2023): 832-854.

[39] Hassan, Junaid, et al. "[Retracted] The Rise of Cloud Computing: Data Protection, Privacy, and Open Research Challenges—A Systematic Literature Review (SLR)." Computational intelligence and neuroscience 2022.1 (2022): 8303504.

[40] https://www.kaggle.com/datasets/prasad22/healthcare-dataset

[41] https://csrc.nist.gov/glossary/term/personally_identifiable_information

[42] Sachin Popat Patil. "Enhancing Cloud Security by Integrating Data Masking Techniques With AWS for Effective DDoS Prevention". International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 22s, July 2024, pp. 748 -, https://ijisae.org/index.php/IJISAE/article/view/6551.

[43] Sharmila, K., S. Borgia Anne Catherine, and V. S. Sreeja. "A comprehensive Study of Data Masking Techniques on cloud." International Journal of Pure and Applied Mathematics 119.15 (2018): 3719-3727.

[44] Ciampi, Mario, Mario Sicuranza, and Stefano Silvestri. "A privacy-preserving and standard-based architecture for secondary use of clinical data." Information 13.2 (2022): 87.

[45] General Data Protection Regulation (GDPR), art. 32(1), 2016 O.J. (L 119) 1.

[46] Arshad, Umair. "Revolutionizing Open Data Privacy Unveiling COBAD's Superiority over Traditional Methods." (2023).

[47] Awadallah, Ruba, and Azman Samsudin. "Homomorphic encryption for cloud computing and its challenges." 2020 IEEE 7th International Conference on Engineering Technologies and Applied Sciences (ICETAS). IEEE, 2020.

[48] Munjal, Kundan, and Rekha Bhatia. "A systematic review of homomorphic encryption and its contributions in healthcare industry." Complex & Intelligent Systems 9.4 (2023): 3759-3786