Privacy-Preserving Deep Learning Models for Analysis of Patient Data in Cloud Environment

Sandhya Avasthi and Ritu Chauhan

Abstract A substantial amount of patient data is being generated every second by the healthcare sector. The medical data especially patient data could be used for analysis through advanced deep learning models, but the private nature of patient data limits the use. Massive volumes of diverse data must be collected, which is often only possible through multi-institutional collaborations. One way to create large central repositories is through multi-institutional studies. This method is limited to privacy issues, intellectual property, data identification, standards, and data storage when data sharing is done. As a result of these challenges, cloud data storage has become increasingly viable. The various models for exchanging medical records on the cloud while protecting privacy are discussed in this chapter. Furthermore, vertical partitioning of medical datasets that exploits attribute categories in health records is explained and analyzed in order to examine distinct areas of medical data with varying privacy issues. These methods can ease the strain on communication costs while minimizing the need to communicate sensitive patient information.

1 Introduction

Medical institutions and research agencies have been collecting medical records in the forms of Electronic Health Records (EHR) and Personal Health Records (PHR) of patients due to the use and availability of all kinds of health applications by patients and healthcare providers. Such medical data in the forms of patient records, clinical trials, and other medical reports are massive that is stored on third-party storage solution providers because such data and its maintenance are very expensive and time-consuming. In addition to the primary use of medical data like diagnosis,

S. Avasthi
Department of Computer Science and Engineering, ABES Engineering College, Ghaziabad, India
e-mail: sandhya_avasthi@yahoo.com

R. Chauhan (✉)
Center for Computational Biology and Bioinformatics, Amity University, Noida, India
e-mail: rituchauha@gmail.com

D. P. Acharjya and K. Ma (eds.), *Computational Intelligence in Healthcare Informatics*, Studies in Computational Intelligence 1132,
https://doi.org/10.1007/978-981-99-8853-2_20

treatment, and prediction of patient disease, secondary use of such data has become quite common due to data analytics by medical institutions. Medical institutions have collected massive genetics data, medical, clinical, and biological. Smartphones and many other wearable devices have enabled third-party corporations to mHealth services through these services they can collect health data of patients. Secondary use of health data is defined by the American Medical Informatics Association as "any use of Personal Health Information (PHI) for purposes other than direct care, including but not limited to analysis, research, quality measurement, public health, payment, provider certification or accreditation, marketing, and other business activities, including strictly commercial activities [1, 2]". Secondary use of data analytics in conjunction with advanced deep learning models aids in clinical decision-making, the extraction of useful patterns and information about medicine and diseases, and the improvement of patient care while lowering healthcare costs and benefiting public health policy [3–5].

Health Insurance Portability and Accountability Act (HIPAA) defines 18 categories of protected health information to preserve the privacy of patient information [6], introduced in the United States. Privacy issues constrain the use of health big data for secondary purposes. To make a balance between privacy protection and the secondary use of all kinds of patient information, organizations are incorporating technical and legal solutions [7]. COVID-19, which was just announced, may highlight the challenge of preserving health information while ensuring its accessibility to address the challenges posed by a significant worldwide epidemic. China, South Korea, and other countries are following the guidelines of the suit [8] to have a mandate for the usage of data from contact surveillance devices. To be able to get useful insights and results, deep learning technologies and cloud infrastructure as well as collaborative model training can be suitable. Because user devices have limited resources, transferring resource-intensive operations to external infrastructure, such as the cloud, which has high-power computing and huge storage, is the solution. Collaborative learning, on the other hand, improves learning accuracy by large-scale diverse datasets originating from disparate sources like patient devices, hospitals, and medical institutions. However, there have been privacy concerns expressed for private medical data usage for model training and inference in deep learning.

Since the advent of the latest information technology tools, healthcare applications have become very adaptable and scalable, to achieve universal scalability and accessibility cloud-based healthcare solutions are in demand. Private cloud infrastructures enhance information flow between healthcare centers and end-users. Heterogeneous communication standards that support various applications and services promote accessibility and information transmission. The cloud platform service provides centralized and decentralized storage as well as ubiquitous access to information across several terminals. Access and retrieval are key storage design capabilities that enable the flexible distribution of information via networked systems [9, 10]. The end-users and intelligent processing systems have access to the same information store but with distinct control and constraint mechanisms. Many intelligent systems or applications process raw patient data to identify patterns or even provide diagnostic information. Further, many end-user applications provide statistical reports and analysis for which

they depend on health data [11]. Effective indexing and storage retrieval qualities are required for information storage and retrieval, notably in healthcare applications. Among many other requirements, flexibility, concurrency, and efficient data retrieval are required from a cloud-based healthcare system [12].

This chapter reviews various methods and strategies to store massive health data and the secondary use of such data in deep learning to gain insights. Since medical data is huge, the cloud platform has become important to store medical data. Privacy is safeguarded purely by technical measures in the cloud where medical data is stored, and users have to trust the cloud platform service providers. Authentication, anonymous communication, anonymous yet approved transactions, data deidentification, and database pseudonymization are some technical measures for privacy-preserving systems. The rest of the chapter is organized as follows. Section 2 discusses foundations of medical data, deep learning, and cloud computing. Electronic health records and its categories are discussed in Sect. 3 followed by protected health information and regulations in Sect. 4. Section 5 discusses deep learning approaches for privacy-preservation. Further, in Sect. 6, cloud environment and privacy-preserving framework are discussed. The vertical partitioning approach is discussed in Sect. 7 followed by conclusion in Sect. 8.

2 Medical Data, Deep Learning, and Cloud Computing

Consistency of clinical data from different time frames, organizations, and research sites is crucial for secondary data analytics utilization. To use current state-of-the-art technologies, deep learning models, and other implementations, standard terminology and ontologies need to be applied to biomedical data.

2.1 Medical Data and Secondary Usage

Any use of patient data beyond its regular intended purposes such as diagnosis, identifying symptoms, and generating reports is called secondary use. Analysis, research, measuring safety and quality, payments, and marketing purpose are some activities that organizations do which are categorized as secondary use. This entails a taxonomy to identify to clarify all kinds of technical and legal issues that might arise due to secondary usage. These medical data mainly contain a database of administrative information, claims, and patient details. The purpose of secondary use is research and applications to improve the quality of treatment and personal care with the help of the latest technologies [13, 14]. The difficulties associated with secondary usage is refactoring, management, aggregation of variables, and maintaining the quality of data (missing data). The other very critical concern associated with secondary use is the security and privacy of patient information. The typical flow of medical data and secondary usage to improve clinical care are shown in Fig. 1.

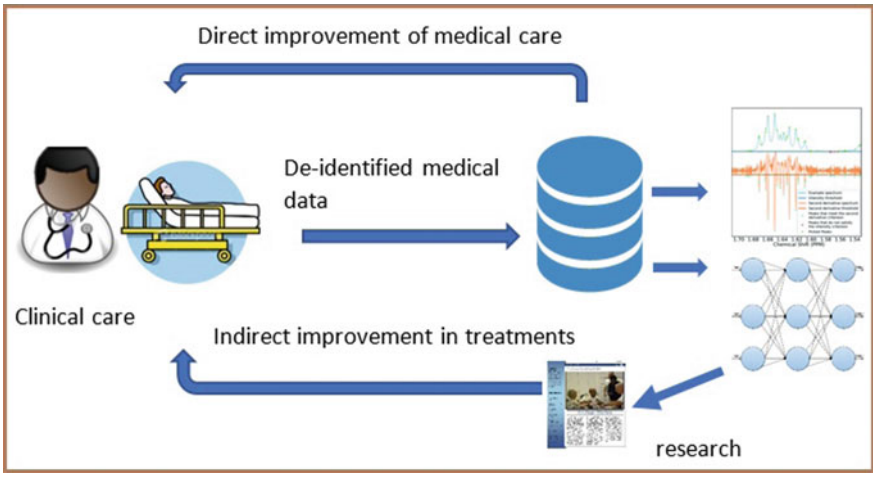

Fig. 1 Clinical and medical data flow in research and analysis

2.2 Deep Learning

A multi-layer computational approach for learning data representations at various levels of abstraction is called deep learning. The model begins with raw data, and each level can employ non-linear transformations to translate the previous level's representation into a higher level representation. Compiling a significant number of these changes enables the learning of complex functions. Deep learning frameworks are applied in a varied range of applications such as AI, image identification [15], object identification, speech recognition, biometric systems, face detection [16], and expert systems [17]. Deep learning is typically divided into two phases: a training phase used to enhance the model's accuracy and an inference phase used to classify the system.

Deep learning is used in medicine to seek patterns in patient histories and spot anomalies in medical imaging that assist in illness diagnosis and prognosis. The application of machine learning in health care can threaten patient privacy by identifying genetic markers [18]. Deep learning is also commonly utilized in finance for a variety of purposes, including price prediction and portfolio creation. In these instances, an entity normally trains its model, and the model parameters are kept private. It is deemed an invasion of privacy to be able to locate or infer them [19]. The aforementioned advancements have been made possible by the ease of access to big datasets and high processing capacity (GPUs and TPUs). These datasets are frequently gathered from the public and may contain sensitive information. Because neural networks are employed in several facets of our life [20–22], they raise severe privacy problems.

2.3 Cloud as a Platform to Store Health Data

Cloud computing platform combines distributed, grid, and utility computing to provide infrastructure, platforms, and applications over the Internet. Cloud platforms and services intend to provide the hosting of patient data, installation, or software services and also provide the necessary infrastructure to manage healthcare operations. It can also help with health record transfer, availability, and retrieval, as well as huge data volume sharing and Electronic Medical Record (EMR) exchange between hospitals and healthcare institutions. Furthermore, it allows medical providers to monitor patients remotely. On the other hand, there are limitations to cloud-based health data exchange. Because it includes personal information, medical history, treatment, linked diseases, symptoms, and even family health history, patient health information is extremely sensitive. Moving sensitive health data or health records to a cloud controlled by third parties exposes them to illegal access and, as a result, poses serious privacy concerns. For example, a trustworthy and approved doctor in a hospital has complete access to all health information, including personal information. Patients, on the other hand, do not want their personal and sensitive information to be shared with someone who is not trustworthy or approved.

3 Electronic Health Records and Categories

Individuals can access and manage their health information via the Internet, which is particularly relevant for our continued consideration of PHRs. Individuals can also share this information with others and have access to the information of others to whom they have permission. Apart from the main use of medical data, Health Records (HRs) also have a feature known as secondary data use. It is concerned with the use of PHI for uses other than direct treatment, such as analysis, research, quality and safety measurement, public health, payment, provider certification or accreditation, marketing, and other commercial operations. In contrast to the benefits listed above, HRs are not without risk. This is simply because the central availability of health-related data increases the risk of these data being misused. Data protection officials and representatives from medical institutions frequently address this issue, with the so-called transparent patient and physician being a prominent concern. The former implies that a person's health status be completely transparent to anybody who has access to their HRs.

Any information on a user's health could be deemed health information in general. Clinical data, particularly Electronic Medical Records (EMR), is the most important sort of health data. It is created by various levels of hospitals. Many other types of health data are being recorded like diet or exercise or heart rate data, and IoT devices generate huge amounts of such data. Technological advancement has made people dependent on wearable devices and other mobile applications making this situation even more data-centric. In general, health-related data can be categorized into four

categories [23]. This research focuses on the first two types of data, which are directly related to users' health and privacy.

Health data are generated by the healthcare system referred to as Category 1. When a patient receives healthcare services in a hospital or clinic, clinical personnel collect clinical data. Clinical data sources include the EMR, prescriptions, test results, pathology images, radiography, and payor claim data. Patients' prior and current conditions are documented to determine therapy needs. It is vital to collect and exchange clinical data with several healthcare professionals over time to improve patient care. It was proposed that patients' clinical data from several institutions and across their lifetime be combined into a Personal Health Record (PHR). This sort of health data is created and gathered regularly as part of the healthcare process to analyze and enhance therapy. Clinical data has a high level of health-related privacy due to the nature of clinical treatment and the high amount of trust consumers have in healthcare experts and organizations. As a result, clinical data privacy is the focus of the health privacy legislation. Thus, a significant proportion of clinical data has been designated for internal use solely by medical institutions. Meanwhile, clinical data is especially valuable for secondary use because it is provided by professionals and provides an up-to-date picture of customers' health status. The delicate balance between utility and privacy associated with this type of health data has been one of the most perplexing issues in the age of big data.

The consumer health and wellness industry provides health data that comes under Category 2. This type of health information complements clinical data well. Consumer attitudes around health have evolved substantially away from passive treatment and toward active health as a result of the widespread adoption of next generation information technologies such as the Internet of Things, mobile health, smartphones, and wearable gadgets. Consumer health data can be generated through wearable fitness trackers, medical wearables such as insulin pumps and pacemakers, health monitoring apps, and online health services. Examples of health data include breathing, heart rate, blood pressure, blood glucose, walking, weight, diet preference, location, and online health consultation. These goods and services, as well as health data, are key components of consumers' daily health management, especially for those with chronic diseases. Industry and academia are increasingly concentrating their efforts on this subject. Consumer health informatics is a representative field [23, 24]. Although this type of unusual health-related data is frequently as revealing of health status as conventional data, it is typically less accessible to physicians, patients, and public health officials to improve individual and community health. To protect the privacy of patients, these massive amounts of health data are divided among organizations. Apart from the utility-privacy trade-off, integration and connectivity of this type of health data at the individual level present other challenges. Table 1 summarizes the two aspects of health data and their distinctions.

Table 1 Clinical data and health data of patient summary

Categories	Category 1	Category 2
Generated/recorded by	Medical equipment, Clinical professional, Healthcare system	Wearable device in treatments, IoT devices
Data detail	Name, age, id, phone, medical history, family history, conditions, medicine use, therapy, narratives, prescriptions, test results	Name, id, type, phone, address, position, age, weight, heart rate, breathing pattern, test, blood pressure, blood glucose, exercise data, diet preference, online health consultation
Data features	Discrete but more professional, more clinical information and more privacy, stored in the healthcare system, passive	Less standardized, more health information, privacy tends to be ignored, stored by different providers, active, vast amounts

4 Protected Health Information and Regulations

Due to the sensitive nature of patient personal and health data, several privacy protection regulations have been established to govern the secondary use of clinical and personal health data. The Fair Information Practices Principles System (FIPPS) serves as the bedrock of contemporary data protection regulation [25]. One of these rules is the Health Insurance Portability and Accountability Act (HIPAA), which was enacted to oversee and regulate the use of medical information [26]. The act protects healthcare systems and insurance firms from all sorts of fraud, theft, and misuse. The HIPAA Safe Harbor (SH) rule requires the destruction of 18 types of expressly or potentially identifying characteristics, collectively referred to as protected health information, before the distribution of health data to a third party. HIPAA also applies to electronically protected health information. This includes, but is not limited to, medical imaging and EHRs. Table 2 is primarily composed of PHI components that pertain to identity data and do not contain any sensitive aspects. That is, HIPAA does not provide instructions on how to safeguard sensitive attribute data; rather, the primary goal of the HIPAA SH rule is to safeguard privacy by preventing identity exposure. On the other hand, other sensitive characteristics can be uniquely combined into a quasi-identifier, enabling data users to reidentify the individuals to whom the data pertains. As a result, strict adherence to the SH rule may not be adequate to guarantee data quality or privacy to whom the data relates [27].

Table 2 Patient information protected by HIPAA

Categories	Description	Categories	Description
1	Date, time	10	Certificate/license number
2	Location	11	Health plan beneficiary numbers
3	Patient Names	12	Vehicle identifier and serial number
4	Phone numbers	13	Device identifier and serial numbers
5	e-mail address	14	URLs of website, domain name
6	Social security numbers	15	Speech, fingerprints, and other biometric markers
7	Medical record numbers	16	Face images and other images
8	IP address details	17	Unique identifying number, characteristics or code
9	Account numbers		

5 Deep Learning Approaches for Privacy-Preservation

Deep Learning (DL) approaches based on artificial neural models can be applied in various applications, including image classification, autonomous driving, natural language processing, medical diagnosis, intrusion detection, and credit risk assessment. Reverse engineering and stealing of data through DL are very common, and it is possible to recreate images of a patient, and sensitive training data could be inferred. This section reviews the research on strategies and techniques of DL to protect the privacy of sensitive medical data belonging to a patient. The different privacy-preserving mechanism is summarized and classified in Fig. 2. Mainly three types of strategies can be found. The first method is data aggregation which collects data and aggregates data from various sources into one dataset keeping contributor anonymity [28, 29]. The second mechanism encompasses a substantial amount of research that focuses on finding strategies to keep the model training process private so that sensitive information about the training dataset's participants is not revealed. Finally, the third mechanism focuses on the inference phase of deep learning.

Data Aggregation: Here are some of the most well-known data privacy-preserving techniques. Although not all of these tactics apply to deep learning, this chapter will go through them quickly for completeness. Context-free and context-aware privacy approaches are the two types of strategies available. Differential privacy and other context-free privacy solutions have no way of knowing what context or purpose the data will be used for. Context-aware privacy solutions, such as information-theoretic privacy, consider how the data will be used and can improve the privacy-utility trade-off.

Naive Data Anonymization: This is the process of removing identifiers from data, such as participants' names, addresses, and full postcodes, to protect privacy. This

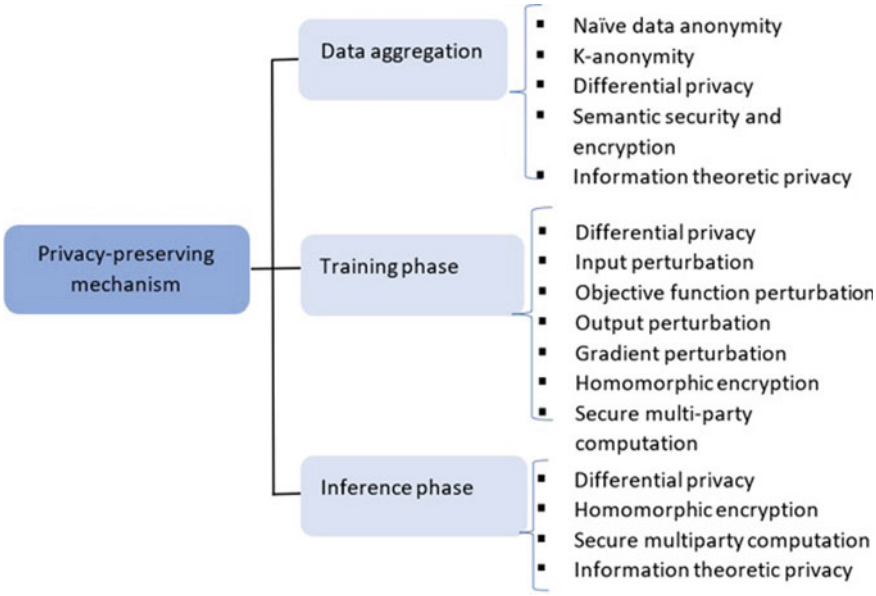

Fig. 2 Different privacy-preserving paradigms in the DL training phase

technique was implemented to protect patients while processing medical data, but it has repeatedly been shown to be useless. Perhaps the most egregious failure is the Netflix Prize instance, in which Narayanan and Shmatikov apply their de-anonymization technique on the Netflix Prize dataset. This dataset includes the anonymous movie ratings of 500,000 Netflix subscribers. They showed that an attacker with additional information about a user (from publicly available Internet Movie Database entries) can quickly identify the person and extract potentially sensitive information [30].

K-Anonymity: If each participant's information cannot be discriminated from the information of at least $(k-1)$ other participants in the dataset, the dataset has the k-anonymity property. K-anonymity states that there are at least k rows with an identical set of attributes available to the adversary for any given combination of attributes [31, 32]. k-anonymization, on the other hand, has been demonstrated to perform poorly when it comes to anonymizing high-dimensional datasets.

Differential Privacy: When two mechanisms with Privacy Budgets 1 and 2 are applied to the same datasets, together they use a privacy budget 1+2. As such, composing various differential private mechanism consumes a privacy budget that increases linearly [33, 34]. Without relying on a centralized server, differential privacy can be achieved by having each participant apply differentially private randomization to their data before sharing it. The approach "randomized response" is demonstrated to be locally differentially private [35], and the model is called the local model of differential privacy.

Semantic Security Encryption: A standard privacy requirement of encryption methods is semantic security [36], which specifies that an opponent given background information should have a cryptographically minimal advantage, a measure of how successfully an adversary can attack a cryptographic algorithm.

Information Theoretic Privacy: Information-theoretic privacy is a context-aware privacy solution. Context-aware approaches model the dataset statistics, as opposed to context-free solutions, which assume worst-case dataset statistics and adversaries. Privacy and fairness have been explored using information-theoretic methods, in which privacy and fairness are provided through information degradation, obfuscation, or adversarial learning, and mutual information reduction is used to verify them. Further, Generative Adversarial Privacy (GAP), a context-aware privacy system that generates private datasets using Generative Adversarial Networks (GANs) is discussed. An attacker tries to deduce secret attributes, whereas a sanitizer strives to eradicate them [37].

5.1 Learning Phase

The private training in deep learning and research articles on them can be categorized based on the guarantee that these methods provide differential privacy or semantic security and encryption. Privacy using encryption can be achieved by doing computation over encrypted data. Homomorphic Encryption (HE) and Secure Multi-party Computation (SMC) are two common methods for encryption of data.

HE is a program that allows one to compute encrypted data [38]. A client can transmit their data to a server in an encrypted format, and the server can compute over it without decrypting it, then send the client a ciphertext for decryption. Because HE is so computationally intensive, it has yet to be used in many production systems [39, 40]with confidentiality. SMC tries to create a network of computing parties that carry out a particular calculation while ensuring that no data leaks. Only an encrypted portion of the data is accessible to each party in this network.

Figure 3 describes a deep learning aggregated model for multi-institutional learning or known as federated learning. The randomization required for differential privacy can be inserted in five places that are input, objective function, gradient updates, output, and labels [41]. Input perturbations can be considered equivalent to using a sanitized dataset for training. Objective function perturbation and output perturbation are explored for machine learning tasks with convex objective functions. For instance, in the case of logistic regression, it is proved that objective perturbation requires sampling noise on the scale of $\frac{2}{n\epsilon}$ [42], and output perturbation required sampling noise on a scale of $\frac{2}{n\lambda\epsilon}$, where n is the number of samples and λ is the regularization coefficient. The proposed work is more practical and is a general objective perturbation approach that works for high-dimensional real-world data.

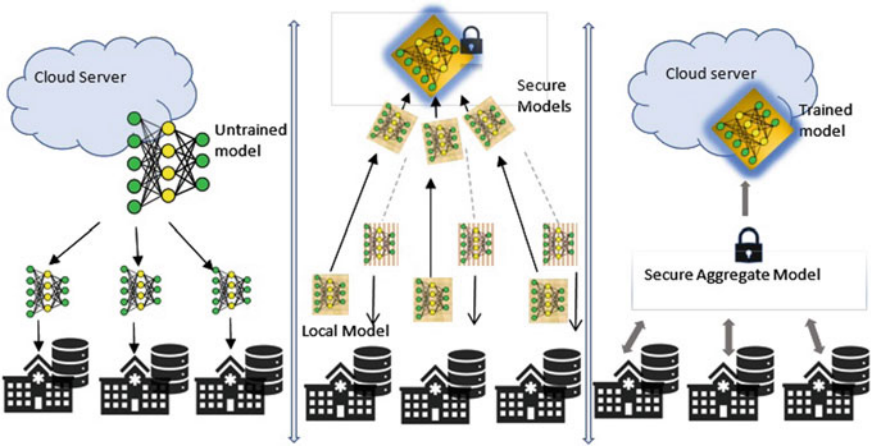

Fig. 3 The framework of deep learning on the cloud

5.2 Inference Privacy

Inference privacy, as opposed to training, aims to provide Inference-as-a-Service through a cloud-based health inferencing system. This cloud-based system can efficiently analyze data on cloud storage and can learn hidden patterns underlying medical data. The situation is only expected to carry out the inference task that has been assigned to it. The classification of literature for inference privacy is similar to that for training, except for one additional area called Information-Theoretic (IT) privacy. The context-aware mechanism is intelligent and sends only information sufficient enough for analysis purposes removing much information content about the patient.

Differential Privacy: The major reason in differential privacy is that it provides a worst-case guarantee, which necessitates applying high-intensity noise to all input segments. On pre-trained networks, this automatically degrades performance. Arden is a data nullification and differentially private noise injection strategy used for inference [43]. Arden divides the DNN between the edge and the cloud. The mobile device performs a simple data transformation, while the cloud data center handles the computationally intensive and complicated inference. Simultaneously, it employs data nullification and noise injection to make various requests indistinguishable, protecting clients' anonymity. The suggested method necessitates noisy retraining of the entire network, with noise inserted at various layers.

Homomorphic Encryption: CryptoNets is one of the first efforts in the field of HE inferences [44]. CryptoNet is a method for transforming a learned neural network into an encrypted one. This allows inference service clients to provide their data in an encrypted format and obtain the result without having to decode it. CryptoNets enables Single Instruction Multiple Data (SIMD) operations, which boost the deployed system's throughput. However, the latency of this technique is still

considerable for single queries. GAZELLE [45] is a system for secure and private neural network inference with shorter latency. It is a two-party computation system that combines homomorphic encryption and regular two-party computation techniques such as garbled circuits. Besides, it is three orders of magnitude quicker than CryptoNets because of its homomorphic linear algebra kernels, which transfer neural network operations to efficient homomorphic matrix vector multiplication and convolutions.

Secure Multi-party Computation: Unlike GAZELLE, which utilizes Additively Homomorphic Encryption (AHE) to speed up linear algebra directly, MiniONN [46] uses AHE in a preprocessing stage. When compared to CryptoNets, MiniONN shows a considerable performance boost without sacrificing accuracy. However, it is a two-party computation system that does not support multi-party calculation. Further, Chameleon, a two-party computation system whose vector dot product of signed fixed-point integers increase prediction performance in heavy matrix multiplication classification algorithms. Chameleon improves on MiniONN by 4.2 seconds in latency. The majority of work in the field of SMC for deep learning is focused on speeding up computation.

Information Theoretic Privacy: Privacy-preserving strategies that rely on information-theoretic approaches often assume a nonsensitive task, the task that the service is supposed to perform, and strive to degrade any unnecessary information in the input data [47–49]. Anonymization techniques for securing temporal sensory data through obfuscation are proposed [50] which offer a multi-objective loss function for training deep autoencoders to extract and conceal user identity-related data while maintaining the sensor data's value. The training method instructs the encoder to ignore user identifiable patterns and tunes the decoder to shape the output without regard for the training set's users.

6 Cloud Environment and Privacy-Preserving Framework

Cloud computing incorporates many techniques like grid, utility, and distributed computing to provide infrastructure to host data and provide software services. Recent healthcare trends, which emphasize accessing information at any time and from any location, favor sending healthcare data to the cloud. Although the cloud has many advantages, it also poses new privacy and security risks to health data [51].

6.1 Cloud Privacy Threats

To protect patient privacy, various measures need to be taken to safeguard medical data on the cloud. Four primary components that can violate the privacy of patients are data representation, data transmission, data distribution and processing, and data storage.

Data representation refers to any user that processes medical data as input, retrieve, and visualization. This could generally be a client browser. Data transmission refers to the transmission of medical data from the client machine to the health record system on the cloud and vice versa. Data distribution and processing refers to the system that handles medical data, and gives a representation of health records on the cloud. This part of the system also manages the efficient storage of data in a distributed manner. In data storage, all kinds of medical data related to the patient are permanently stored on the cloud in a distributed manner. This is based on a database management system and other storage solutions to provide query processing.

6.2 Privacy-Preserving Framework

The word "privacy-preservation" refers to a lot more than just keeping data secret. Spoofing identification, tampering with data, repudiation, and information exposure are all challenges to data privacy in the cloud [52]. In a spoofing identity attack, the attacker impersonates a legitimate user, whereas data tampering entails harmful content change and modification. Users who deny executing an action with the data are considered repudiation threats. The exposing of information to entities with no permission to access it is known as information disclosure [53]. An EMR is a record of patient medical records information. Some other type of medical records can be EHR. The Health Information and Management System Society (HIMSS) and ISO/TS 18308 standards describe these types of medical records [54]. A healthcare institution, such as a hospital, creates and manages EMR. These records are proof of overall patient health, treatment, and other results to track the patient's treatment. The EHR is created and maintained inside a single institution or community. It's a digital record that can be shared throughout a community, region, or state's many institutions. Data from EMRs can be fed into EHRs. When EMR data is shared with other institutions, EHRs are established. Figure 4 depicts the conceptual framework for sharing medical data in the cloud while maintaining privacy.

Vertical Partitioning of Medical Data: This component in any system preserves the privacy of the medical data by dividing data and then storing it at various locations in the cloud. The EMR 'T' is partitioned into three different tables T_p, T_a, and T_e. The T_p is a plaintext table; T_a is an anonymized table where many of the personal identifiers of the personal has been removed. The third table T_e is a table of explicit identifiers and quasi-identifier. After the portioning, these three tables are moved to cloud storage in separate locations.

Merging Process: All the stored medical data in the cloud can be accessed through the merging process. The recipient has to merge partitioned medical data into one dataset. The data recipient can access T_p directly for medical data research or analysis with the approval of the data owner. When the plaintext table T_p is unable to meet the information needs of data consumers, this component is utilized to integrate T_p with T_a and T_e, resulting in two additional medical dataset access paradigms. Similarly,

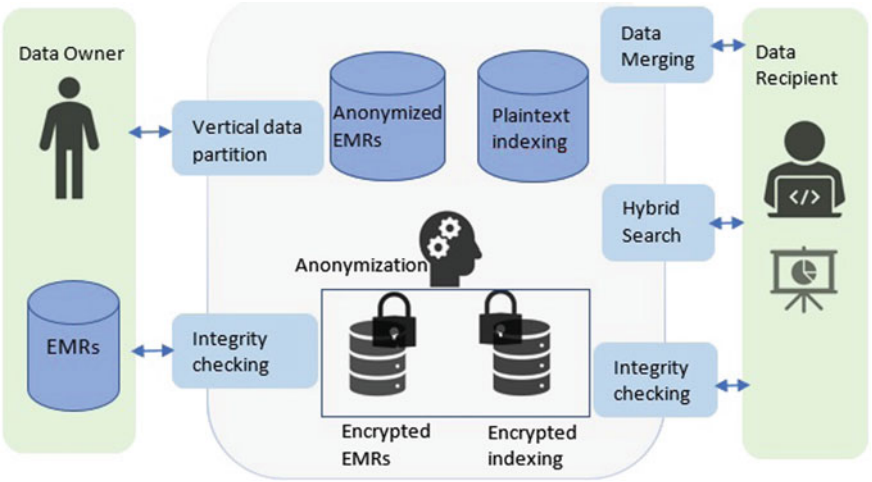

Fig. 4 Conceptual framework of a cloud environment for medical data

the data recipient can access anonymized medical dataset and full version medical dataset access through original EMR table T.

Integrity Checking This component is often used by both the data owner and the data recipient to confirm that the data saved in the cloud is equivalent to what was originally recorded. Two integrity checking systems are given due to the varied requirements of data owners and data recipients.

Hybrid Search This component is used by the data recipient to achieve record-level medical data access, i.e., to locate one or more interested EMRs in the shared medical dataset. For the implementation of information retrieval over remote medical data storage, a hybrid search technique is proposed that combines encrypted and plaintext search methods.

In any cloud-based healthcare system, the entities are mainly patients, physicians, pharmacists, pathologists, nurses, laboratory staff, insurance companies, reports, and cloud service providers. Various ways to maintain e-Health cloud privacy are based on adversarial models. One model considers cloud servers to be untrustworthy entities that might potentially reveal sensitive health information. Furthermore, untrusted cloud servers are vulnerable to both internal and external attacks. The adversary may not only use forged credentials to access the encrypted health data, but they may also obtain access to the data as privileged users. Inside enemies may pose a danger to health data kept on trusted cloud servers in the second scenario. Parts of the data, for example, may be stored by a doctor, who may then share the data with unauthorized parties, resulting in information leakage [55].

Additionally, the entity identities must be protected. The third strategy entails the usage of semi-trusted cloud servers, while semi-trusted cloud services are considered trustworthy that collect information about health data and collaborate with unauthorized users [55]. The intruder in these cases might steal or can manipulate

the patient's personal information; even in worse cases might even sell it to third parties for monetary gains. For instance, a physician's prescription medicine may be disclosed to representatives of pharmaceutical companies, or insurance company spending information may be falsified. To deal with such cases, the healthcare system should provide a mechanism and guarantee to protect private sensitive information belonging to a patient.

7 Vertical Partitioning Approach

If the data volume is high, the data is growing at a high rate and it is of varied kinds; the healthcare system needs partitioning of data to store them because they cannot be stored in one repository. Many large-scale solutions split data so that it can be maintained and retrieved separately. Scalability, conflict reduction, and performance optimization can all be aided by partitioning. It can also be used to categorize data according to usage trends.

There are three aspects to the EMR conceptual model such as patient data (name, birth date, address); profile of patient (medical history and reports); and data such as symptoms, diagnoses, and treatments, for each patient's hospital visit [56, 57]. Overall data like patient data is in proper format and structure, but patient profiles and clinical reports are semi-structured that include a lot of text information. Consider that 'T' is a table that holds EMR data, to maintain data privacy; T's properties can be categorized into three distinct groups like Quasi-Identifiers (QID), Explicit Identifiers (EID), and Medical Information (MI).

All collected information that together identifies a patient or individual is called QID attributes. EID refers to information such as name, social security number, and phone number that can uniquely identify individual patient records. MI refers to information containing clinical, reports, and medical data about patients and is called medical information. Knowing that MI contains all the sensitive patient information, data analysis of such data might reveal the identity of the person and so vertical partitioning use can be done to hide the details by keeping the table in a partitioned manner.

Three vertical partitioned tables T_p, T_a, and T_e are created from the original EMR table T. T_p contains attributes from MI, T_e contains attributes from EID, and T_a has information from QID. Their precise values are encrypted and saved in ciphertext. Allow A to signify the set of all attributes $\{A_1, A_2, \cdots, A_m\}$ and $t[A_j]$ to denote the value of attribute A_i for tuple t. The attributes in A are categorized as EID, QID, and MI, with EID equal to $\{A_1, A_2, \cdots, A_{|EID|}\}$, QID equal to $\{A_1, A_2, \cdots, A_{|QID|}\}$, and MI equal to $\{A_1, A_2, \cdots, A_{|MI|}\}$.

8 Conclusion

In light of the developing cloud computing realm and deep learning, models use in bioinformatics; protecting patient information has become a serious issue. EHRs and personal health records keep all kinds of personal and clinical information. Privacy considerations are critical when dealing with such sensitive medical data; yet, with today's sophisticated cloud apps, privacy concerns are a distant second. The necessary technical steps have not been widely addressed and are not included in standard solutions. The chapter introduces the notion of deep learning as well as several ways for protecting patient privacy from insider threats. The chapter's primary objective is to provide a detailed overview of the secondary use of medical data, deep learning for data analytics, and storage solutions on cloud platforms. A cloud provider can host an effective health record system even if it can't link people and their health information. In addition, the chapter goes through four essential parts of cloud storage: horizontal and vertical partitioning, integrity, and privacy-preserving query processing. Data splitting makes it possible to store information straightforwardly and effectively. It also enables more flexible access and lowers the cost of data storage by integrating cryptographic techniques and statistical analysis.

References

1. Solares, J.R.A., Raimondi, F.E.D., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., Gomes, A.C.P., Payberah, A.H., Zottoli, M., Nazarzadeh, M., Conrad, N., Rahimi, K., Salimi-Khorshidi, G.: Deep learning for electronic health records: a comparative review of multiple deep neural architectures. J. Biomed. Inform. **101**(1), 103337 (2020)
2. Azencott, C.A.: Machine learning and genomics: precision medicine versus patient privacy. Philos. Trans. R. Soc. A: Math. Phys. Engin. Sci. **376**(2128), 20170350 (2018)
3. Si, Y., Du, J., Li, Z., Jiang, X., Miller, T., Wang, F., Zheng, W.J., Roberts, K.: Deep representation learning of patient data from Electronic Health Records (EHR): a systematic review. J. Biomed. Inform. **115**(3), 103671 (2021)
4. Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. **13**(6), 395–405 (2012)
5. Jiang, M., Chen, Y., Liu, M., Rosenbloom, S.T., Mani, S., Denny, J.C., Xu, H.: A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J. Am. Med. Inform. Assoc. **18**(5), 601–606 (2011)
6. Office for Civil Rights, H.H.S.: Standards for privacy of individually identifiable health information. Final rule. Feder. Regist. **67(157)**, 53181–53273 (2002)
7. McGraw, D., Mandl, K.D.: Privacy protections to encourage use of health-relevant digital data in a learning health system. NPJ Digital Med. **4**(1), 2 (2021)
8. Liu, Y., Zhang, L., Yang, Y., Zhou, L., Ren, L., Wang, F., Liu, R., Pang, Z., Deen, M.J.: A novel cloud-based framework for the elderly healthcare services using digital twin. IEEE Access **7**, 49088–49101 (2019)
9. Jungkunz, M., Köngeter, A., Mehlis, K., Winkler, E.C., Schickhardt, C.: Secondary use of clinical data in data-gathering, non-interventional research or learning activities: definition, types, and a framework for risk assessment. J. Med. Internet Res. **23**(6), e26631 (2021)
10. Xue, J., Xu, C., Bai, L.: DStore: A distributed system for outsourced data storage and retrieval. Futur. Gener. Comput. Syst. **99**, 106–114 (2019)

11. Manogaran, G., Shakeel, P.M., Fouad, H., Nam, Y., Baskar, S., Chilamkurti, N., Sundarasekar, R.: Wearable IoT smart-log patch: an edge computing-based Bayesian deep learning network system for multi access physical monitoring system. Sensors **19**(13), 3030 (2019)
12. Li, D., Huang, L., Ye, B., Wan, F., Madden, A., Liang, X.: FSRM-STS: Cross-dataset pedestrian retrieval based on a four-stage retrieval model with Selection Translation Selection. Futur. Gener. Comput. Syst. **107**(6), 601–619 (2020)
13. Avasthi, S., Chauhan, R., Acharjya, D.P.: Processing large text corpus using N-gram language modeling and smoothing. In: Proceedings of the Second International Conference on Information Management and Machine Intelligence, Springer Singapore, pp. 21-32 (2021)
14. Hutchings, E., Loomes, M., Butow, P., Boyle, F.M.: A systematic literature review of researchers' and healthcare professionals' attitudes towards the secondary use and sharing of health administrative and clinical trial data. Syst. Control Found. Appl. **9**(1), 1–27 (2020)
15. Ozyurt, F.: Efficient deep feature selection for remote sensing image recognition with fused deep learning architectures. J. Supercomput. **76**(11), 8413–8431 (2020)
16. Santhanavijayan, A., Naresh Kumar, D., Deepak, G.: A semantic-aware strategy for automatic speech recognition incorporating deep learning models. In: Proceedings of the Intelligent System Design, pp. 247–254. Springer, Singapore (2021)
17. Jain, R., Gupta, M., Taneja, S., Hemanth, D.J.: Deep learning based detection and analysis of COVID-19 on chest X-ray images. Appl. Intell. **51**, 1690–1700 (2021)
18. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in pharmacogenetics: An End-to-End case study of personalized warfarin dosing. In: Proceedings of the 23rd USENIX Security Symposium, pp. 17–32 (2014)
19. Bolton, R.J., Hand, D.J.: Statistical fraud detection: a review. Stat. Sci. **17**(3), 235–255 (2002)
20. Thompson, S.A., Warzel, C.: Twelve million phones, one dataset, zero privacy. In: Ethics of Data and Analytics, pp. 161–169. Auerbach Publications (2022)
21. Schiff, J., Meingast, M., Mulligan, D. K., Sastry, S., Goldberg, K.: Respectful cameras: detecting visual markers in real-time to address privacy concerns. In: Protecting Privacy in Video Surveillance, pp. 65–89 (2009)
22. Senior, A., Pankanti, S., Hampapur, A., Brown, L., Tian, Y.L., Ekin, A., Connell, J., Shu, C.F., Lu, M.: Enabling video privacy through computer vision. IEEE Secur. Privacy **3**(3), 50–57 (2005)
23. Geetha Mary, A., Acharjya, D.P., Iyengar, N.C.S.: Improved anonymization algorithms for hiding sensitive information in hybrid information system. Int. J. Comput. Netw. Inf. Secur. **6**(6), 9–17 (2014)
24. Avasthi, S., Chauhan, R., Acharjya, D.P.: Extracting information and inferences from a large text corpus. Int. J. Inf. Technol. **15**(1), 435–445 (2023)
25. Cate, F.H.: The failure of fair information practice principles. In: Consumer Protection in the Age of the Information Economy, pp. 341–377. Routledge (2016)
26. Mendes, R., Vilela, J.P.: Privacy-preserving data mining: methods, metrics, and applications. IEEE Access **5**, 10562–10582 (2017)
27. Li, X.B., Qin, J.: Anonymizing and sharing medical text records. Inf. Syst. Res. **28**(2), 332–352 (2017)
28. Sweeney, L.: k-anonymity: A model for protecting privacy. Internat. J. Uncertain. Fuzziness Knowl.-Based Syst. **10**(05), 557–570 (2002)
29. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. J. Privacy Confident. **7**(3), 17–51 (2016)
30. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: Proceedings of the IEEE Symposium on Security and Privacy, pp. 111–125 (2008)
31. Aggarwal, C.C.: On k-anonymity and the curse of dimensionality. In: Proceedings of the VLDB Conference, Trondheim, Norway vol. 5, pp. 901–909 (2005)
32. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data **1(1)**, 3–es (2007)
33. Dwork, C., Rothblum, G.N., Vadhan, S.: Boosting and differential privacy. In: Proceedings of the IEEE 51st Annual Symposium on Foundations of Computer Science, pp. 51–60 (2010)