sustainability

MDPI

Article

Federated Learning Approach to Protect Healthcare Data over Big Data Scenario

Gaurav Dhiman [1,2], Sapna Juneja [3], Hamidreza Mohafez [4,*], Ibrahim El-Bayoumy [5], Lokesh Kumar Sharma [6], Maryam Hadizadeh [7], Mohammad Aminul Islam [8,*], Wattana Viriyasitavat [9] and Mayeen Uddin Khandaker [10]

1 Department of Computer Science, Government Bikram College of Commerce, Patiala 147001, India; gdhiman0001@gmail.com
2 University Centre for Research and Development, Department of Computer Science and Engineering, Chandigarh University, Gharuan, Mohali 140413, India
3 Department of Computer Science, KIET Group of Institutions, Delhi NCR, Ghaziabad 201206, India; sapnajuneja1983@gmail.com
4 Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur 50603, Malaysia
5 Public Health and Community Medicine, Tanta Faculty of Medicine, Tanta 31527, Egypt; ibrahim.elbayoumy@med.tanta.edu.eg
6 School of Computer Science Engineering and Technology, Bennett University, Greater Noida 201310, India; lokesh.gbu@gmail.com
7 Centre for Sport and Exercise Sciences, Universiti Malaya, Jalan Universiti, Kuala Lumpur 50603, Malaysia; mary@um.edu.my
8 Department of Electrical Engineering, Faculty of Engineering, Universiti Malaya, Jalan Universiti, Kuala Lumpur 50603, Malaysia
9 Department of Statistics, Chulalongkorn University, Bangkok 10330, Thailand; hardgolf@gmail.com
10 Centre for Applied Physics and Radiation Technologies, School of Engineering and Technology, Sunway University, Bandar Sunway 47500, Malaysia; mayeenk@sunway.edu.my
* Correspondence: h.mohafez@um.edu.my (H.M.); aminul.islam@um.edu.my (M.A.I.)

Abstract: The benefits and drawbacks of various technologies, as well as the scope of their application, are thoroughly discussed. The use of anonymity technology and differential privacy in data collection can aid in the prevention of attacks based on background knowledge gleaned from data integration and fusion. The majority of medical big data are stored on a cloud computing platform during the storage stage. To ensure the confidentiality and integrity of the information stored, encryption and auditing procedures are frequently used. Access control mechanisms are mostly used during the data sharing stage to regulate the objects that have access to the data. The privacy protection of medical and health big data is carried out under the supervision of machine learning during the data analysis stage. Finally, acceptable ideas are put forward from the management level as a result of the general privacy protection concerns that exist throughout the life cycle of medical big data throughout the industry.

Keywords: big data; healthcare; mobile device; patient clinical records; federated learning

check for updates

Citation: Dhiman, G.; Juneja, S.; Mohafez, H.; El-Bayoumy, I.; Sharma, L.K.; Hadizadeh, M.; Islam, M.A.; Viriyasitavat, W.; Khandaker, M.U. Federated Learning Approach to Protect Healthcare Data over Big Data Scenario. *Sustainability* **2022**, *14*, 2500. https://doi.org/10.3390/su14052500

Academic Editors: Saqib Iqbal Hakak and Thippa Reddy Gadekallu

Received: 27 December 2021
Accepted: 10 February 2022
Published: 22 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

In addition, the continual integration of medical technologies and information technologies has provided a steady impetus for the collection of medical data, laying the groundwork for the application and growth of big data technology in the medical field. As a type of big data with features such as a large amount of information and rapid expansion, a diverse data structure, and high application value, medical information falls into this category. The collection, management, and analysis of large amounts of medical data, as well as the effective discovery of the data's potential worth, have all played an important part in the advancement of clinical scientific research, clinical decision support, and drug development [1,2].

As a result, the growth of big data in health care is highly regarded both within the country and throughout the global community. Some developed countries have platforms that are relatively advanced in comparison with others [3]. My government is currently focusing on data collection, owing to a late start, and its ability to evaluate and handle data is severely limited. While benefiting from valuable information obtained from medical data infuses new vitality into clinical scientific research and other areas, such as health management and public health, it also brings about the problem of privacy leakage. Examples include a study undertaken by Greenbone Networks from the middle of July to the beginning of September that examined thousands of online medical care systems from all over the world [4,5]. Thus, over 24 million patient data records from various countries may now be seen and downloaded with reasonable ease using the Internet [6–8]. Patient data records that have been leaked contain specific personal and medical information, such as your name, date of birth, examination date, survey items, attending physician's name, and image information of test results [9]. The compromised patient data records contain information about your health as well. To harm someone's reputation, their names and images can be published; the leaked data can be linked with other data to conduct phishing and social engineering [10,11]; and the data can be read and automatically processed to search for valuable identity information, such as the use of ID numbers to steal an individual's identity [12,13].

How to improve the utilisation rate of medical data and tap the value hidden in them while protecting patients' privacy is a critical factor limiting their development at the moment [14]. As a result, throughout the entire life cycle of medical and health big data, it is critical to make full use of data while strictly preventing privacy leakage and striving to strike a balance between data utilisation and privacy protection [15,16].

2. Sources and Characteristics of Medical Big Data

Electronic data in medicine and health are growing at an unprecedented rate as a result of the growth of information technology in the medical industry [17]. A wide range of data types are available, including patient disease diagnostic and treatment information, physical health information, and data from medical clinical experiments [18–20]. These different types of complicated and large-scale medical data are brought together to demonstrate the features of big data, which are collectively referred to as "big medical data." Figure 1 primarily summarises the sources of medical big data as well as their properties [21,22].

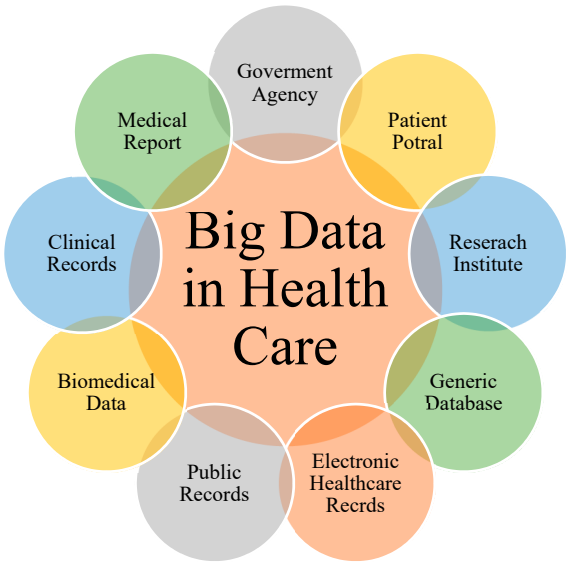

Figure 1. Big data management in healthcare.

Clinical big data: This part of data is mainly generated during the patient's medical treatment and constitutes extensive primary medical and health data. In seeking medical treatment, the patient produces a series of private data [23]. First, you need to provide detailed personal information, such as name, age, address, and phone number. The doctor will directly record according to experience judgments or generate electronic medical record data, medical image data, and drug use records during the diagnosis and treatment—part of the clinical data [24,25]. In addition, relevant cost information, medical insurance usage, and so forth will also be involved in the medical treatment process. This information will also be recorded. Under the condition of big data, these data can generate new value through system analysis. However, these also directly contain a large amount of personal information [26]. Once obtained by an illegal third party, they will now threaten patients' privacy [27].

Health big data: With the intelligentisation of life, wearable devices and mobile phone applications have penetrated people's lives. The information they obtain can help everyone monitor and record detailed personal physical data and browse and consult on significant websites [28]. Illness, health, and other related content behaviours will expose personal preference data. These data are connected to medical institutions through the Internet to form the content of electronic health records, which are used to monitor everyone's health [29,30]. These real-time data, which record the detailed health status of individuals, are aggregated through the Internet, which may expose a series of sensitive information, such as health status, location, and personal preferences [31].

Biological big data: High-throughput sequencing technology is rapidly evolving, increasing the data output capability of life-science-related research organisations, resulting in a variety of gaps across the spectrum of genomics, transcriptomics, proteomics, and metabolomics research, among others [32,33]. The enormous value of these biological data serves to successfully encourage the advancement of biological research, which has applications in agriculture, health, and medicine, among other fields of endeavour [34]. However, when genetic testing data are combined with pathological data, it is much easier to identify specific individuals. While patient privacy is jeopardised, it is also simple to generate genetic prejudice, resulting in a negative impact on the patient on both counts [35].

Extensive data of process and operation: During the process and operation of various medical institutions, a large amount of data will be generated accordingly, for example, operation costing data, medicines, consumables, equipment procurement data, drug research and development data, and consumer purchase behaviour data [36,37]. The transaction records of drugs or related devices in the data also often expose the user's private information, such as the user's physical condition and financial status. Therefore, it is also content that cannot be ignored in privacy protection [38].

Medical big data conform to the typical characteristics of big data—large scale, rapid growth, diverse structures, and great value. In addition, big medical data have other unique properties [39].

High sensitivity: Medical big data often directly record the patient's detailed personal information and physical health. Compared with other data, they are more sensitive and have higher requirements for privacy protection [40].

Incompleteness: There is sometimes a chasm between the collection and processing of medical and health data. The data in the medical database, on the other hand, are vast, and it is still difficult to record all disease information completely [41]. Furthermore, because electronic medical records have yet to gain widespread acceptance, a substantial amount of data is derived from manual records. Furthermore, deviations and incompleteness in record content, ambiguity in verbal expression, and insufficient data storage are all sources of incomplete medical health large datasets [42,43].

Timeliness: There is a progressive change in the time of the patient's treatment and disease process, and the waveforms and image data of medical testing have a specific time sequence [44]. The patient's health status is not static but always in dynamic change, which means that the corresponding value of its sensitive attribute is changing over time.

3. Privacy Protection of Medical Big Data

Medical big data collection, storage, sharing, and analysis involve a large number of users. As a result, every link has serious privacy leaks, and appropriate technical measures must be taken to address them. At the same time, some privacy issues pervade all links and can be resolved by implementing appropriate management measures. This article focuses on privacy leakage challenges and privacy protection technologies from various points in the medical big data life cycle [45]. Finally, it makes some reasonable recommendations from the management level of medical big data (Figure 2).

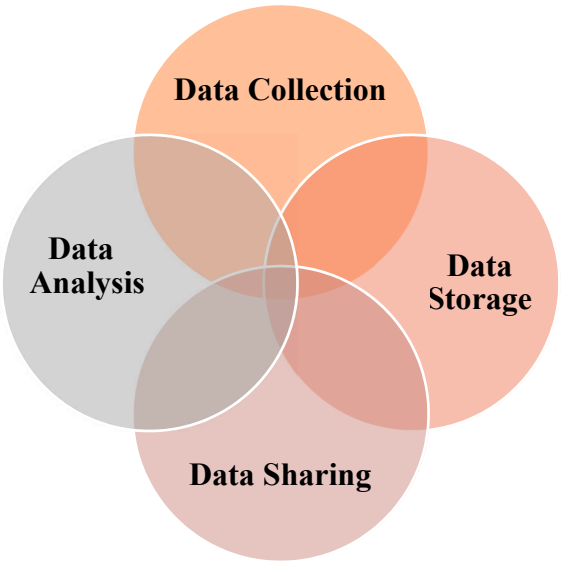

Figure 2. Management level of medical big data.

Data collection: Data collection is essential in the medical and health big data life cycle. The advancement of information technology has allowed the practise of medical health to permeate all facets of people's daily lives [46]. Medical information can come from a variety of sources, including information systems, wearable devices, and networks of medical facilities. It is necessary to bring together medical and health big data from multiple sources during the data collecting stage in order to build the groundwork for further data storage, sharing, and analysis in order to achieve success. At the opposite end of the spectrum from traditional data gathering, the collecting of medical data directly includes private information provided by the patient. In other words, the medical information is vulnerable. In light of the fact that medical and health big data are vulnerable, the question of how to efficiently hide content that may leak user privacy when the information is public is one that must be addressed as soon as possible.

Data storage: The data storage stage is concerned with the privacy risks associated with large-scale medical and health data storage and management. Because of the vast scale of medical and health data, big data must be stored on a cloud platform after collection. It is completely independent of the owner and the data saved on cloud platforms; however, the cloud storage service provider cannot be completely trusted. As a result, medical data stored on a cloud platform are not secure, and they are vulnerable to third-party information theft or alteration, which is unacceptable.

Data sharing: Data stored in different medical institutions can be maximised through data sharing. However, in a big data environment, while data sharing brings convenience, it also brings risks to patients. When the patient's data are stored on the cloud platform, the patient does not know who has accessed the data in the shared account, so there is a high risk of data leakage. In addition, the data cannot be tracked after the leakage, which is a more significant challenge for privacy protection.

Data analysis: Only through the analysis of large amounts of medical data can the advancement of medical fields, such as disease diagnosis and drug research and development, be aided, as well as the provision of better services to patients. After a series of clustering, correlation, and additional data analysis of medical data, sensitive patient information may still be exposed, even after anonymisation, encryption, and other processing. Privacy and security must prevent the leakage of sensitive information in the original data while also taking into account the results of data mining, analysis, and prediction.

4. Privacy Protection Technology in Medical Big Data Collection

Data collection is the essential step in the data life cycle. The collection of medical and health big data facilitates scientific research and cooperation between institutions, posing potential threats to data privacy.

The risk at this stage is link attacks based on data integration or other more complex attacks based on knowledge background. For example, medical data, such as a patient's diagnosis and treatment data, drug or medical device purchase records, and related social information on the Internet, can serve for data analysis and reflect user behaviour to a certain extent. Suppose the attacker intercepts these data from the network transmission and uses other external information comprehensively to infer the individual's identity. In that case, this brings a severe challenge to protecting patient privacy.

For the most part, traditional medical data privacy protection relies on anonymity technologies. The most important concept is to conceal the relationship between the data and the individual; however, simply removing individual attributes from the data can easily be disrupted via link assaults, which are quite effective. As a countermeasure to this attack strategy, it was suggested that *k*-anonymity be used. The idea is that the quasi-identifier in the data (unrecognisable qualities that can relate to multiple people, such as date of birth and postal code) can be used to match at least k different people in the database. In other words, a certain piece of information will not be able to identify itself from other k-1 personal information datasets. The *l*-diversity model was developed as a means of defending against homogeneity attacks and background knowledge attacks based on k-anonymity. According to the principle of k-anonymity, each sensitive attribute must have at least one well-performing value to be considered valid. The *t*-closeness model is a development of the l-diversity model that is more specific. When data representation is reduced to a finer granularity, the l-diversity model maintains privacy while also treating distinct attribute values differently by taking the distribution of attribute values into consideration. This is a method for obtaining some privacy, which means a reduction in the effectiveness of data mining, which is a trade-off.

However, existing anonymity technology has a flaw: it relies too heavily on the attacker's assumptions about his or her own prior information. It is unable to provide rigorous and sufficient evidence of the level of privacy protection it provides. The implementation of differential privacy in the medical industry has proven to be an effective solution to problems associated with anonymous technology. When using the differential privacy protection paradigm, there is no need to consider any prior knowledge obtained by the attacker from other sources. Additionally, differential privacy provides a precise mathematical description as well as a way of calculating privacy leaks in real time. This tool allows you to compare the availability of datasets processed with different parameters using the same or different parameters.

5. Anonymous Technology

To a certain extent, data anonymity safeguards the privacy of individuals' personal information. Some anonymity technologies that are better suitable for large amounts of medical data have been developed, which are based on the traditional anonymity protection methods of k-anonymity, l-diversity, and t-closeness models.

Ambigavathi et al. suggested a random k anonymity approach in answer to the problem of enormous data scale. In order to reduce the amount of time spent searching

for anonymous equivalence, a two-step clustering algorithm is employed to partition the original dataset into equivalence groups. In order to drastically reduce the computing cost of discovering anonymous equivalence classes, it is first necessary to divide the original dataset into numerous subdatasets and then establish equivalence classes in each of the subdatasets. As a result, the information loss associated with unknown datasets is significantly reduced. Usability is now more reliably ensured.

The collected medical and health data usually have many different sensitive attributes. Therefore, when operating high-dimensional data, the association and mixing of these additional sensitive attributes are also worth paying attention to. In this case, the (a,k)-anonymous privacy protection method will be more productive. Li et al. used the (a,k)-anonymity model as a privacy protection scheme for data collection and proposed a new data collection method based on anonymity in healthcare services using a client–server–user model for analysis. On the client-side, the concept of (a,k)-anonymity generates anonymous tuples to resist possible attacks. Furthermore, a bottom-up clustering method is used to create clusters that meet the basic level of anonymous privacy. On the server-side, the communication cost is reduced through generalisation technology, and anonymous data are compressed through the clustering and combination method based on upgrade so that the data meet a deeper level of privacy.

Due to the incompleteness of medical big data, to avoid the reduction of information availability caused by this feature, Pei Mengli proposed an anonymous algorithm, DAIMDL (data anonymity for incomplete medical data based on l-diversity), based on l-diversity DAIMDL algorithm groups' data records based on clustering. After optimising the grouping, the divided data groups are generalised. In the clustering stage, clustering is performed based on the distance calculation of information entropy to ensure that the information distance within the cluster is the smallest and the information distance between the collections is the largest; in the generalisation stage, the divided data groups are generalised, and finally, the quasi-identity in each group is obtained. Equivalence classes with the same character attribute value. The DAIMDL algorithm processing patient information can avoid discarding incomplete data records in the data table and reduce the information loss of medical data. At the same time, the sensitive attributes in the medical data are diversified and distributed, and there is no less than 1 different sensitive attribute value in each equivalence class group. Therefore, the obtained medical dataset meets the requirements of the l-diversity anonymous model.

Data are constantly updated, inserted, and deleted due to the continuous update characteristics of big medical data. The static anonymity technology is still in use, and there is no doubt that new privacy leaks will occur. The standard privacy protection model includes a safe and anonymous method for incremental datasets based on l-diversity, but it can only solve the data insertion operation. The m-invariance plan, proposed in the literature, can dynamically publish data insertion and deletion. To protect privacy to the greatest extent possible, the concept of pseudo-tuples is added in addition to satisfying m-invariance-related rules. Simultaneously with the release of the data, an auxiliary table is also released to record the statistical information of inserting pseudo-tuples. Shi et al. further considered that the target's specific quasi-identification attributes and sensitive attributes would change (such as disease recovery or deterioration and changes in physical indicators) and proposed a dynamic update scheme. This scheme uses the Laplace noise mechanism to protect the sensitive attributes of the result set, saves the quasi-identifying attributes and sensitive attributes separately, gives the recipient different results according to his or her permissions, and finds a solution that can guarantee the availability of information, but also the best cluster for privacy protection.

Anonymity technology can better prevent the leakage of sensitive patient data and ensure the authenticity of the data. It has received widespread attention in practical applications, but there is still room for improvement. The current research on the balance between privacy and usability is mainly focused on reducing information loss. How to find a reasonable balance point is a problem that needs further investigation. Most of the

anonymisation methods used are greedy algorithms, and the execution efficiency is not high. Therefore, it is necessary to study efficient anonymisation algorithms to cope with the ever-increasing issue of large-capacity data release.

Furthermore, there is no unified measurement and evaluation standard for anonymisation technology. Therefore, we need to devote ourselves to this research to objectively and reasonably evaluate anonymisation technology. In addition, how to efficiently realise personalised anonymity and quickly and accurately select the quasi-identifier of the data table according to the actual application and how to solve the anonymisation of multiple data tables in a distributed environment are all issues worthy of deep consideration and research.

6. Differential Privacy Technology

Differential privacy is a privacy model that can successfully withstand the majority of privacy assaults and provide verifiable privacy assurances, as opposed to anonymous privacy models. While maximising the availability of medical data, it also ensures that the leakage of patient privacy is kept within the expected control range of a given situation. The amount of noise introduced into the dataset as a result of using differential privacy technology is determined by the sensitivity of the query function. This has no relationship with the size of the dataset. Because of the query function's low sensitivity, it is possible to introduce a small amount of noise to achieve the goal of privacy protection while also significantly increasing the availability of medical data when dealing with large-scale medical data. In light of these developments, differential privacy appears to be a promising strategy for protecting the privacy of medical data.

In order to guarantee data privacy while still maintaining the correctness of data queries, differentiable privacy technology has been developed. The authors [45] created a heuristic hierarchical query approach and then suggested a private partition algorithm for differential privacy in order to reduce computational overhead and query mistakes [12]. Electronic health records and genetic data are the primary subjects of research towards differential privacy in the medical field. Several authors have published papers in which they first encrypt the data before employing the differential noise process to interfere with them, thereby safeguarding both genomic privacy and the privacy of distributed clinical data. Apart from that, they are committed to merging biology and informatics at the bedside (i2b2) frameworks, improving privacy while simultaneously minimising network overhead.

In a similar vein, the author [13] used the usual differential privacy protection method and two-way decryption method from the literature to safeguard the genetic data from any potential attacker. This resulted in an improvement in both the secrecy and execution speed of the i2b2 framework in electronic genome data records, according to the author. A different private aggregation technique, described in [11], was created by the author, which aggregates health device data while also providing timely incentives to the users. Differential privacy, the Boneh-Goh-Nissim encryption system, and Shamir secret sharing are all used in this technique to improve user security and privacy. The model is constructed with the help of the Java JPBC package, which reduces the amount of work required.

7. Privacy Protection Technology in Medical Extensive Data

The collection of medical data and the promotion of electronic medical records have created a solid data foundation for the application of machine learning in the medical field. Only through the analysis and processing of large amounts of medical data can valuable knowledge and rules for disease diagnosis, treatment, and medical research be unearthed. However, some data appear to be unrelated. Some sensitive information may be mined using data mining technology: data that do not involve personal privacy when they appear independently may be sufficient to analyse after matching with personal information.

By conducting data mining on medical data, some information and patterns that could not be identified may be exposed and leaked to untrusted third parties. Therefore,

it is necessary to analyse and process data to protect the privacy and limit the mining of sensitive knowledge in big data. Although big medical data have gone through a series of cleaning operations, the confidentiality of patients cannot be directly obtained from the dataset. Still, after mining a large amount of collected information, some sensitive information may be leaked through the mining results. Therefore, how to protect privacy in applying machine learning to the medical field is a problem worthy of research in the analysis of medical and health big data.

(a) Confidential Calculation

Confidential computing emphasises the transfer of data and the confidentiality of calculations during the machine learning training process in order to ensure data privacy protection. Secure multiparty computing and a trusted execution environment are some of the current ways for implementing confidential computing.

The trusted execution environment considers hardware security to be a mandatory need, and a safe region is formed on the computing chip by itself to ensure the integrity of the data and code operating on the chip.

Intel's SGX (software guard extensions) technology was used to create the trusted execution environment, which was then used to create the rare disease gene data analysis system PRINCESS (privacy-protecting rare disease international network collaboration via encryption through software guard extensions), which was proposed [11]. Secure distributed calculations on encrypted data are performed, and an allele association study for Kawasaki disease is carried out on a family-based basis. The PRINCESS method delivers safer and more accurate analysis than alternatives, such as homomorphic encryption and corrupted circuits, while also being more efficient. The SGX technology has also been used to develop a genetic data analysis framework called PRESAGE (privacy-preserving genetic testing via software guard extension) [47] as well as a novel and safe genetic relationship analysis method called PREMIX (privacy-preserving estimation of individual admixture) [12]. However, SGX based on hardware security will be vulnerable to side-channel attacks of specific algorithms [12].

Homomorphic encryption is a type of encryption that allows confidential calculations to be performed on ciphertext without requiring a key. Only the key can be used to decode the message and return it to plaintext. When it comes to genetic data analysis, homomorphic encryption technology is typically used due to the sensitive nature of the data. HEALER (homomorphic computation of exact logistic regression) or example, is a rare illness research platform that uses genetic data to assess rare variants with a small sample size while still maintaining the confidentiality of human genome data. Generally, genomic data owners only supply encrypted sequences for genetic data analysis [11], and public, commercial clouds can do sequence analysis without decryption on the data they receive. The result can only be decrypted by the data owner or a designated representative who has access to the decryption key and can perform the necessary operations. [12] All genotype and phenotypic data in the genome-wide association analysis project are completely homomorphically encrypted, which allows the cloud to do meaningful calculations on the encrypted data. While homomorphic encryption has a wide range of potential applications in practise, the huge processing burden associated with it means that the current method can only be applied to the inference portion of datasets [13].

(b) Federated Learning

Federated learning is essentially a framework for distributed learning, whereas machine learning is not a category of distributed learning. The purpose of switching to federated learning was to generate a machine learning model using datasets that are distributed across multiple devices and halting any loss of data at the same time [47]. The centralised data integration training of multiple medical institutions can frequently outperform separate training using the data of one institution. Nonetheless, each medical institution hopes that its data are secure, and centralised data integration frequently results in complex privacy and data security issues. Additionally, there are other issues. The data

owner can also obtain the training model via federated learning without providing the data directly. The model's training effect can also be guaranteed, which is nearly identical to the training effect after data integration. The federated learning technology effectively protects medical and health data privacy through parameter exchange. The data and models are kept locally and will not be transmitted by themselves, so there is no possibility of leakage at the data level.

The computational formula of federated learning can be shown as:

$$\text{Arg Min } L(a,b,c) = \sum P_k L_k(a,b,c) \quad (1)$$

where k = total number of clients;

P_k = weight of k^{th} client;
a = input;
b = output;
c = parameter to be learned

In the paper [19], the author jointly analysed the data of multiple hospitals to find the phenotype of a specific patient population under the condition that each hospital's data do not leave the local area. It can be seen from the results of the study that there is a big difference between the results obtained by using the data of one hospital alone and the data analysis of the two hospitals combined. However, using the federal learning method, when the data do not come out of the hospital, the accuracy and phenotype of the discovery aspect are similar to the centralised training model while respecting privacy [20].

8. Comparative Analysis

In this paper, a comparative analysis was conducted between federal optimisation scheme (cPDS) and Princess algorithm, which provides safe transmission of healthcare data over a big data environment [48]. Comparison was performed over the evaluation parameter end to end encryption delay, decryption possibility, and encryption time.

It is seen that delay was increased with increased load over the network, as shown in Figure 3 and Table 1.

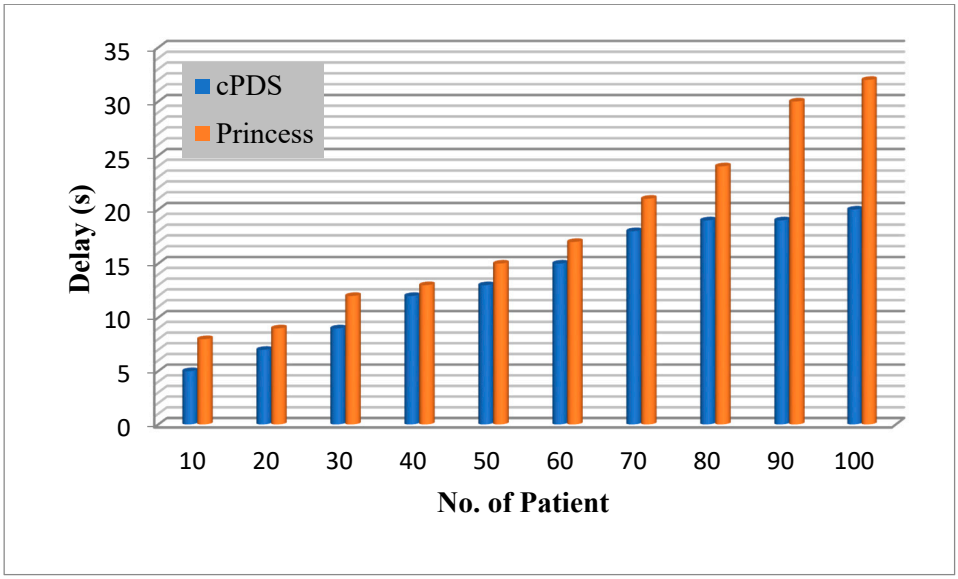

Figure 3. End-to-end delay over clinical data transmission.

Table 1. End-to-end delay over clinical data transmission.

No. of Patients	Delay (s)	
	cPDS	Princess
10	5	8
20	7	9
30	9	12
40	12	13
50	13	15
60	15	17
70	18	21
80	19	24
90	19	30
100	20	32

It is seen that decryption possibility dynamically varied over the different patient scenarios but generally decreased over large patient detail, as shown in Figure 4 and Table 2.

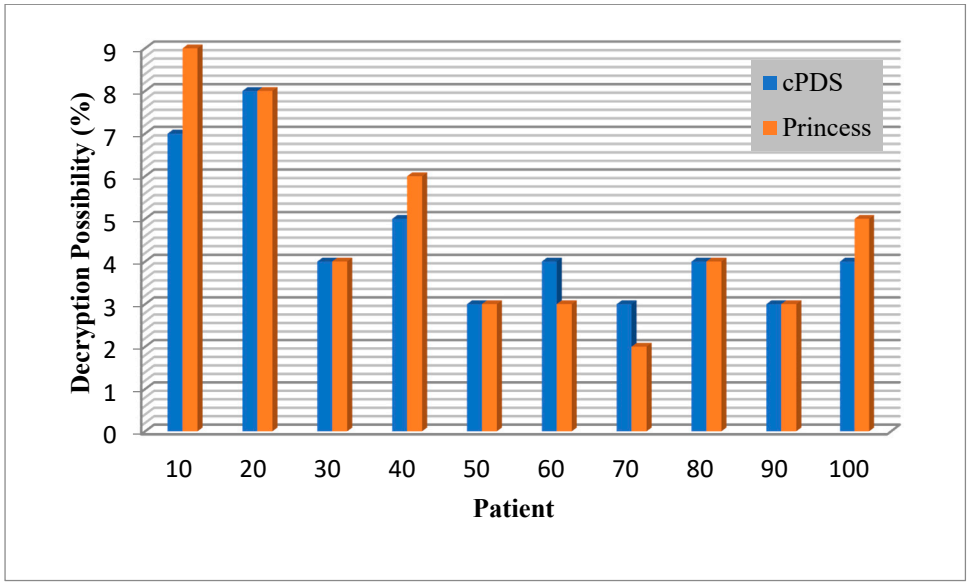

Figure 4. Decryption possibility over encrypted clinical data.

Table 2. Decryption possibility over encrypted clinical data.

No. of Patients	Decryption Possibility	
	CPDS	Princess
10	7	9
20	8	8
30	4	4
40	5	6
50	3	3
60	4	3
70	3	2
80	4	4
90	3	3
100	4	5

It is seen that encryption time dynamically varied over the different patient scenarios but generally increased and was constant over higher load over large patient detail, as shown in Figure 5 and Table 3.

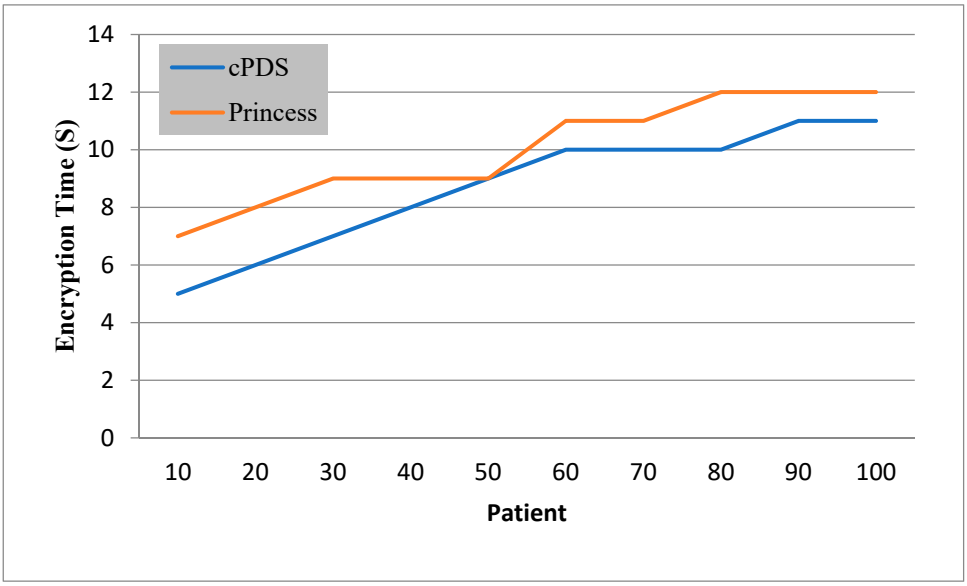

Figure 5. Encryption time for encrypting clinical data.

Table 3. Encryption time for encrypting clinical data.

No. of Patients	Encryption Time	
	CPDS	Princess
10	5	7
20	6	8
30	7	9
40	8	9
50	9	9
60	10	11
70	10	11
80	10	12
90	11	12
100	11	12

9. Conclusions

Ensuring a high utilisation rate of medical big data and extracting data value while effectively protecting user privacy is a critical issue in today's medical research field. This article first introduces the complex sources of medical and health big data and the distinguishing characteristics that set them apart from big general data. Then, beginning with the medical big data life cycle, it introduces the privacy protection issues in each link, categorises privacy protection technologies, briefly discusses the desirability and limitations of various technologies, and investigates the privacy of big medical data—the direction of future data protection technology development. In general, more documents have been written about problems and solutions connected to large medical data, whereas fewer technologies are actually being used in the field. The need for fine-grained and tailored privacy protection is becoming increasingly essential, and this will be the critical content of future research.

In the life cycle of medical and health big data, privacy protection technology can prevent the leakage of privacy to a certain extent. However, suppose there are no scientific and reasonable management measures. In that case, you will still face technical problems that are difficult to control, such as improper manual operations, malicious internal personnel, damaged infrastructure, and unclear relevant laws and regulations.

Establish privacy and security policies and management standards, and improve laws and regulations. The entire life cycle of medical and health big data is inextricably linked

to the operation and management of personnel, such as doctors in the medical department, who have direct access to patients' personal information and test results. Exposed, not only is the patient's physical condition revealed, but so are his or her home address, living habits, and other details. To prevent the malicious use and leakage of patients' sensitive information, strict management standards should be formulated. The staff involved in each link should be trained on privacy and safety regulations implemented in its operation and management.

Improve laws and regulations on the privacy protection of medical and health significant data. Laws are mandatory and are a powerful weapon to protect patient privacy and reduce data leakage. Therefore, the government should speed up the legislative work on the privacy protection of medical and health big data, and further improve the protection system and strengthen the crackdown on malicious data theft. In addition, considering that the transmission of medical and health big data is global, it is also essential to establish and improve a set of international standard laws on the protection of medical and health big data.

Infrastructure should be supervised in real time. The privacy and security of medical and health big data also rely on the security of various infrastructures in the life cycle. For example, once a cloud platform that stores medical data is damaged or maliciously attacked, the data may be lost or tampered with. In the entire life cycle of medical and health big data, multiple infrastructures are involved, and the privacy and security of each link cannot be underestimated. Real-time supervision and protection are required to respond to emergencies for the first time.

Author Contributions: Conceptualization, G.D.; Methodology S.J. and H.M; Software, I.E.-B.; Validation, L.K.S.; Resources, H.M. and M.H.; Writing—Original Draft Preparation H.M. and S.J.; Writing—review and editing, H.M. and G.D.; Project Administration, H.M. and M.A.I.; Supervision, W.V. and M.U.K.; Funding Acquisition, M.A.I. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the Malaysian Ministry of Higher Education through an FRGS grant: FRGS/1/2020/TK0/UM/02/33. This research was also funded by the Universiti of Malaya Research Grant (RU013AC-2021).

Data Availability Statement: The data will be available from authors upon request.

Acknowledgments: This work was financially supported by the Malaysian Ministry of Higher Education through an FRGS grant: FRGS/1/2020/TK0/UM/02/33. This research was also funded by the Universiti of Malaya Research Grant (RU013AC-2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mian, M.; Teredesai, A.; Hazel, D.; Pokuri, S.; Uppala, K. Work in Progress–In-Memory Analysis for Healthcare Big Data. In Proceedings of the 2014 IEEE International Congress on Big Data, Anchorage, AK, USA, 27 June–2 July 2014; pp. 778–779. [CrossRef]
2. Rahman, F.; Slepian, M.; Mitra, A. A novel big-data processing framwork for healthcare applications: Big-data-healthcare-in-a-box. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 3548–3555. [CrossRef]
3. Patil, H.K.; Seshadri, R. Big Data Security and Privacy Issues in Healthcare. In Proceedings of the 2014 IEEE International Congress on Big Data, Anchorage, AK, USA, 27 June–2 July 2014; pp. 762–765. [CrossRef]
4. Lambay, M.A.; Mohideen, S.P. Big Data Analytics for Healthcare Recommendation Systems. In Proceedings of the 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 3–4 July 2020; pp. 1–6. [CrossRef]
5. Kaur, M.J.; Mishra, V.P. Analysis of Big Data Cloud Computing Environment on Healthcare Organizations by implementing Hadoop Clusters. In Proceedings of the 2018 Fifth HCT Information Technology Trends (ITT), Dubai, United Arab Emirates, 28–29 November 2018; pp. 87–90. [CrossRef]
6. Mande, R.; JayaLakshmi, G.; Yelavarti, K.C. Leveraging Distributed Data over Big Data Analytics Platform for Healthcare Services. In Proceedings of the 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 11–12 May 2018; pp. 1115–1119. [CrossRef]

7. Panahiazar, M.; Taslimitehrani, V.; Jadhav, A.; Pathak, J. Empowering personalized medicine with big data and semantic web technology: Promises, challenges, and use cases. In Proceedings of the 2014 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 27–30 October 2014; pp. 790–795. [CrossRef]
8. Ambigavathi, M.; Sridharan, D. Big Data Analytics in Healthcare. In Proceedings of the 2018 Tenth International Conference on Advanced Computing (ICoAC), Chennai, India, 13–15 December 2018; pp. 269–276. [CrossRef]
9. Katarya, R.; Jain, S. Exploration of Big Data Analytics in Healthcare Analytics. In Proceedings of the 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, 22–23 April 2020; pp. 1–4. [CrossRef]
10. Viceconti, M.; Hunter, P.; Hose, R. Big Data, Big Knowledge: Big Data for Personalized Healthcare. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1209–1215. [CrossRef] [PubMed]
11. Vuppalapati, C.; Ilapakurti, A.; Kedari, S. The Role of Big Data in Creating Sense EHR, an Integrated Approach to Create Next Generation Mobile Sensor and Wearable Data Driven Electronic Health Record (EHR). In Proceedings of the 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, UK, 29 March–1 April 2016; pp. 293–296. [CrossRef]
12. Li, X.; Sedeh, R.S.; Wang, L.; Yang, Y. Patient-record level integration of de-identified healthcare big databases. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 1784–1786. [CrossRef]
13. Koppad, S.H.; Kumar, A. Application of big data analytics in healthcare system to predict COPD. In Proceedings of the 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, India, 18–19 March 2016; pp. 1–5. [CrossRef]
14. Bochicchio, M.; Cuzzocrea, A.; Vaira, L. A Big Data Analytics Framework for Supporting Multidimensional Mining over Big Healthcare Data. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 508–513. [CrossRef]
15. Li, L.; Bagheri, S.; Goote, H.; Hasan, A.; Hazard, G. Risk adjustment of patient expenditures: A big data analytics approach. In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6–9 October 2013; pp. 12–14. [CrossRef]
16. Balan, S.; Shristiraj, N.; Shah, V.; Manjappa, A. Big data analysis of youth tobacco smoking trends in the United States. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 4727–4729. [CrossRef]
17. Sterling, M. Situated big data and big data analytics for healthcare. In Proceedings of the 2017 IEEE Global Humanitarian Technology Conference (GHTC), San Jose, CA, USA, 19–22 October 2017; p. 1. [CrossRef]
18. Grover, P.; Johari, R. Review of big data tools for healthcare system with case study on patient database storage methodology. In Proceedings of the 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), Noida, India, 14–15 January 2016; pp. 698–700. [CrossRef]
19. Otoum, S.; Guizani, N.; Mouftah, H. Federated Reinforcement Learning-Supported IDS for IoT-steered Healthcare Systems. In Proceedings of the ICC 2021-IEEE International Conference on Communications, Montreal, QC, Canada, 14–23 June 2021; pp. 1–6. [CrossRef]
20. Aich, S.; Sinai, N.K.; Kumar, S.; Ali, M.; Choi, Y.R.; Joo, M.I.; Kim, H.C. Protecting Personal Healthcare Record Using Blockchain & Federated Learning Technologies. In Proceedings of the 2021 23rd International Conference on Advanced Communication Technology (ICACT), Online, 7–10 February 2021; pp. 109–112. [CrossRef]
21. Nair, R.; Soni, M.; Bajpai, B.; Dhiman, G.; Sagayam, K.M. Predicting the Death Rate Around the World Due to COVID-19 Using Regression Analysis. *Int. J. Swarm Intell. Res. (IJSIR)* **2022**, *13*, 1–13. [CrossRef]
22. Sharma, S.; Gupta, S.; Gupta, D.; Juneja, S.; Singal, G.; Dhiman, G.; Kautish, S. Recognition of Gurmukhi Handwritten City Names Using Deep Learning and Cloud Computing. *Sci. Program.* **2022**, *2022*. [CrossRef]
23. Juneja, A.; Juneja, S.; Kaur, S.; Kumar, V. Predicting Diabetes Mellitus With Machine Learning Techniques Using Multi-Criteria Decision Making. *Int. J. Inf. Retr. Res.* **2021**, *11*, 38–52. [CrossRef]
24. Zeidabadi, F.A.; Doumari, S.A.; Dehghani, M.; Montazeri, Z.; Trojovsky, P.; Dhiman, G. AMBO: All Members-Based Optimizer for Solving Optimization Problems. *CMC-Comput. Mater. Contin.* **2022**, *70*, 2905–2921. [CrossRef]
25. Juneja, S.A.; Juneja, A.; Anand, R. Healthcare 4.0-Digitizing Healthcare Using Big Data for Performance Improvisation. *J. Comput. Theor. Nanosci.* **2020**, *17*, 4408–4410. [CrossRef]
26. Balakrishnan, A.; Kadiyala, R.; Dhiman, G.; Ashok, G.; Kautish, S.; Yadav, K.; Maruthi Nagendra Prasad, J. A Personalized Eccentric Cyber-Physical System Architecture for Smart Healthcare. *Secur. Commun. Netw.* **2021**, *2021*. [CrossRef]
27. Juneja, S.; Juneja, A.; Dhiman, G.; Jain, S.; Dhankhar, A.; Kautish, S. Computer Vision-Enabled Character Recognition of Hand Gestures for Patients with Hearing and Speaking Disability. *Mob. Inf. Syst.* **2021**, *2021*. [CrossRef]
28. Balakrishnan, A.; Ramana, K.; Dhiman, G.; Ashok, G.; Bhaskar, V.; Sharma, A.; Gaba, G.S.; Masud, M.; Al-Amri, J.F. Multimedia Concepts on Object Detection and Recognition with F1 Car Simulation Using Convolutional Layers. *Wirel. Commun. Mob. Comput.* **2021**, *2021*. [CrossRef]
29. Das, S.R.; Sahoo, A.K.; Dhiman, G.; Singh, K.K.; Singh, A. Photo voltaic integrated multilevel inverter based hybrid filter using spotted hyena optimizer. *Comput. Electr. Eng.* **2021**, *96*, 107510. [CrossRef]

30. Dhiman, G.; Kaur, G.; Haq, M.A.; Shabaz, M. Requirements for the Optimal Design for the Metasystematic Sustainability of Digital Double-Form Systems. *Math. Probl. Eng.* **2021**, *2021*. [CrossRef]
31. Juneja, S.; Dhiman, G.; Kautish, S.; Viriyasitavat, W.; Yadav, K. A Perspective Roadmap for IoMT-Based Early Detection and Care of the Neural Disorder, Dementia. *J. Healthc. Eng.* **2021**, *2021*. [CrossRef] [PubMed]
32. Juneja, A.; Juneja, S.; Bali, V.; Mahajan, S. Multi-Criterion Decision Making for Wireless Communication Technologies Adoption in IoT. *Int. J. Syst. Dyn. Appl.* **2020**, *10*, 1–15. [CrossRef]
33. Uppal, M.; Gupta, D.; Juneja, S.; Dhiman, G.; Kautish, S. Cloud-Based Fault Prediction Using IoT in Office Automation for Improvisation of Health of Employees. *J. Healthc. Eng.* **2021**, *2021*. [CrossRef] [PubMed]
34. Kansal, L.; Gaba, G.S.; Sharma, A.; Dhiman, G.; Baz, M.; Masud, M. Performance Analysis of WOFDM-WiMAX Integrating Diverse Wavelets for 5G Applications. *Wirel. Commun. Mob. Comput.* **2021**, *2021*. [CrossRef]
35. Vaishnav, P.K.; Sharma, S.; Sharma, P. Analytical review analysis for screening COVID-19 disease. *Int. J. Mod. Res.* **2021**, *1*, 22–29.
36. Chatterjee, I. Artificial intelligence and patentability: Review and discussions. *Int. J. Mod. Res.* **2021**, *1*, 15–21.
37. Kumar, R.; Dhiman, G. A comparative study of fuzzy optimization through fuzzy number. *Int. J. Mod. Res.* **2021**, *1*, 1–14.
38. Piao, Y.; Ye, K.; Cui, X. A Data Sharing Scheme for GDPR-Compliance Based on Consortium Blockchain. *Future Internet* **2021**, *13*, 217. [CrossRef]
39. Rumbold, J.M.M.; Pierscionek, B. The Effect of the General Data Protection Regulation on Medical Research. *J. Med. Internet Res.* **2017**, *19*, e47. [CrossRef]
40. Juneja, A.; Juneja, S.; Soneja, A.; Jain, S. Real time object detection using CNN based single shot detector model. *J. Inf. Technol. Manag.* **2021**, *13*, 62–80. [CrossRef]
41. Upadhyay, H.; Juneja, S.; Juneja, A.; Dhiman, G.; Kautish, S. Evaluation of Ergonomics-Related Disorders in Online Education Using Fuzzy AHP. *Comput. Intell. Neurosci.* **2021**, *2021*. [CrossRef]
42. Upadhyay, H.K.; Juneja, S.; Maggu, S.; Dhingra, G.; Juneja, A. Multi-criteria analysis of social isolation barriers amid COVID-19 using fuzzy AHP. *World J. Eng.* **2021**. [CrossRef]
43. Gadekallu, T.R.; Pham, Q.-V.; Huynh-The, T.; Bhattacharya, S.; Maddikunta, P.K.R.; Liyanage, M. Federated Learning for Big Data: A Survey on Opportunities, Applications, and Future Directions. *arXiv* **2021**, arXiv:2110.04160. Available online: http://arxiv.org/abs/2110.04160 (accessed on 17 October 2021).
44. Agrawal, S.; Sarkar, S.; Alazab, M.; Maddikunta, P.K.R.; Gadekallu, T.R.; Pham, Q.-V. Genetic CFL: Hyperparameter Optimization in Clustered Federated Learning. *Comput. Intell. Neurosci.* **2021**, *2021*, 1–10. [CrossRef] [PubMed]
45. Juneja, S.; Juneja, A.; Dhiman, G.; Behl, S.; Kautish, S. An Approach for Thoracic Syndrome Classification with Convolutional Neural Networks. *Comput. Math. Methods Med.* **2021**, *2021*. [CrossRef]
46. Agrawal, S.; Chowdhuri, A.; Sarkar, S.; Selvanambi, R.; Gadekallu, T.R. Temporal Weighted Averaging for Asynchronous Federated Intrusion Detection Systems. *Comput. Intell. Neurosci.* **2021**, *2021*. [CrossRef]
47. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–19. [CrossRef]
48. Health Care Analytics—2 Multi Class Classification AV Janatahack Series: Healthcare Analytics II. Available online: https://www.kaggle.com/vetrirah/av-healthcare2/activity (accessed on 17 October 2021).