International Journal of Research in Engineering, Science and Management
Volume-3, Issue-7, July-2020
journals.resaim.com/ijresm | ISSN (Online): 2581-5792

Security and Privacy Preserving Deep Learning Framework that Protect Healthcare Data Breaches

S. Sreeji[1*], S. Shiji[2], M. Vysagh[3], T. Ambika Devi Amma[4]

[1]*PG Scholar, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Palakkad, India*

[2,3]*Assistant Professor, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Palakkad, India*

[4]*Professor & Principal, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Palakkad, India*

*Corresponding author: sreejisreenivasan123@gmail.com

***Abstract*: Big healthcare data security and privacy are a big concern increasing year-by-year. Heterogeneous data called big data, plays overwhelming role in medical industry. More than 750 data breaches occurred in 2015.The top data security breaches occurred from health care industry. The most important data security issue occurs during sharing sensitive data to train the system. There are several methods to protect the privacy of such healthcare data. Among them a distributed deep learning method called SplitNN, is the one which does not share raw data or model details with collaborating institutions (hospitals). Another method is sequentially sharing models in cyclic in order to train deep neural networks. Another approach is synchronous optimization approach which is empirically validated and shown to converge faster and to better test accuracies. The existing systems uses anonymization techniques to protect the privacy. The proposed deep learning framework keep patient's original data in local platforms and send gradient values to the client and back propagate the data without any anonymization. The learning performance improves by using data from different platforms (hospitals) during training.**

***Keywords*: Anonymization, Bigdata, Back Propagation, Data Security, Gradient, SplitNN, Synchronous Optimization.**

1. Introduction

The healthcare industry is witnessing an increase in sheer volume of data in terms of complexity, diversity and timeliness, which termed as BIGDATA. Big Data is a high volume, and high speed, high variety data quality, which needs new varieties of processing to enhance increased decision-making insight discovery, and process optimization. Bigdata implies many changes compared to traditional techniques in three ways: the amount of data (volume), the rate of generation and transmission of data (velocity) and the heterogeneity of the types of structured and unstructured data that it can handle (variety). These are known as 3Vs of bigdata. It is known that the new technologies arises many problems, in bigdata not only issues having issues with 3Vs but with privacy and security of bigdata. Researchers can use data in EHR systems to create deep learning models that will predict certain health-related outcomes such as the probability that a patient will contract a disease. Deep learning is an assistance to the healthcare professionals to analyse any diseases and predict accurately. But the major issue while developing such deep learning framework is security and privacy. It is important to secure existing healthcare big data environments due to increasing threats of breaches and leaks from confidential data. The distributed deep learning method called SplitNN, which proposes a training of distributed deep learning models without sharing model architectures and also not sharing raw data to prevent undesirable scrutiny by other entities. In order to train deep neural networks while preserving the privacy of dataset proposes a method that sequentially sharing models in cyclic order during training procedures.

A. Note on the technology

The phrase big data mean a massive volume of structured and unstructured data. It helps in decision making, and process automation. Bigdata helps to gain complete answers because more data is available, which means a complete different approach to tackle a problem. Machine learning helps the systems to automatically learn and improve from experience. It can access data and use it learn for themselves. To extract higher level features from the raw input in deep learning uses multiple layers. Artificial neural networks are used in modern deep learning. The data in each level learns to represent in more abstract and composite form.

B. Security and privacy in healthcare

In healthcare Adoption of big data increases security and patient privacy concerns. Data centres stores data with varying levels of security. The inflow of large data sets from diverse sources places an extra burden on storage, processing and communication, proposes big data healthcare cloud. But one of the main threat while data storing in cloud is the service provider is completely manages and monitor. The organization may have less control over their infrastructure. Another example is the UNC Health Care (UNCHC), has implemented a new system using natural-language processing allowing clinicians to rapidly access and analyse unstructured patient data. The automations helped to enhance patient care workflow .it increases probability of security and privacy breaches of healthcare. The healthcare organizations maintain, transmit store large amounts of data to support efficient and proper care. The lack of technical support and minimal security is an issue. Healthcare industry is the main area disclosed to data breaches. Attackers uses the private data and release the sensitive data to the public and thus data breach happens. The implementation of security measures is a complex task. The most threat to hospitals are,

- 25 large breaches of PHI, compromising 16,612,985 individual patient records.
- In a single year incident 3,620,000 breached patient records.
- 40% of large breach incidents involves unauthorized disclosure.

C. Big data security in healthcare

Protecting the security and privacy of healthcare data various technologies are used. The technologies are:

1) Data Masking

Data masking is the most important technology used. Data masking means replacing sensitive data elements with an unidentifiable value. It is not an encryption technique so the original value cannot be returned from the masked value. De-identifying the data sets or masking personal identifiers is the strategy data masking uses. The benefit of data masking is that cost of securing a big data deployment is reduced. Masking reduces the need for applying additional security controls on that data while it resides in the platform because secure data is migrated from a secure source into the platform.

2) Authentication

The act of establishing or confirming claims made by or about the subject are true and authorised. In a healthcare system, verifying both healthcare information offered by providers and identities of consumers at the entry of every access.

3) Encryption

Preventing unauthorized access of sensitive data is called data encryption. In healthcare the encryption scheme should be efficient. It should be easy to use by patients and healthcare professionals. A suitable encryption algorithm for a secure storage is a difficult problem.

2. Related Works

Distributed platforms download the learning model from a centralized server. But the security of the server is an issue because in order to store huge amount of data requires cloud servers. The security of the cloud servers is an issue. The data in many cloud applications increases. The data is stored and accessed through the encryption technique and need to be accessed at anytime and anywhere by the providers or patients. Privacy-preserving distributed deep learning via Homomorphic re encryption [4] proposes a deep learning framework which assuming that the server and any learning participant do not collude. Once collude, server could decrypt and get data of all learning participants. Even if learning participant collude with the server no information should leaked. But requires more number of secure channels to communicate with the server. Thus the communication cost of the scheme will be high. Towards privacy-preserving deep learning based medical imaging applications [5] which employs Homomorphic encryption and no decryption is performing. And also not disclosing the real data. But the system provides an attractive solution for data privacy in Deep Neural Networks and having limitation in strengthening the security. Considerations for Privacy Preserved Open Big Data Analytics Platform [6] which discusses the aspects for considering building a practical open big data analytic platform.ie. Health care industry. There are four major roles involved in the development and usage of the Big Data Analytical platform having different privacy concerns. The major roles are Data Producer, Data Curator, Data analyser, Data Maker. Data Curators receives data from data producers. More risk involved in releasing data to data curator from data producer. The risk involves are Predictive Disclosure and Re-Identification Risk. Privacy-preserving scoring of tree ensembles: a novel framework for AI in healthcare [7] which develops a ML scoring service in which the healthcare organization send encrypted data to the ML scoring service. The customers will receiving only encrypted class labels without seeing the input. The service is used by two parties that are health system which possess a trained model and the second party is customers that uses the trained model for predictions. But the system having limitation in classifying the first parties input with the second parties' classifiers. Because the second part only knows the input data and the second part only knows the classifiers look like. Privacy preserving computations over healthcare data [8] uses an encrypted scheme which analyses the data in encrypted form. The healthcare industry collects the data from multiparty. Each data from different parties encrypted with different public keys. So requires a cloud storage for accessing data at any time and anywhere.so the sensitive data stored in stored, processed and shared within the cloud healthcare systems. As the data is stored in cloud need to be protected from unauthorized access. In order to protect it from unauthorized access a proxy re-encryption technique is used.by that can share the data with someone we want in the cloud environment. But the system cannot be applied in the machine learning model. Large-scale machine learning with stochastic gradient descent [9], the data sizes have grown faster than the speed of processors.so that the capabilities of statistical machine learning methods is limited to process such a huge amount of data. The computational complexity of learning algorithm gets increased to overcome such issue stochastic gradient and averaged stochastic gradient are asymptotically efficient. Privacy preserving data mining technique in Hadoop [11]. Privacy preserving using hadoop can cause privacy violation. So to protect the privacy a dummy variable is added as noise to the transaction data. By this generates the amount of fake frequent item set to hide the real frequent item set. This mining is performed in untrusted cloud platforms. A scalable two-phase top-down approach [12] for data anonymization using map reduce on cloud to anonymize large-scale data sets using the Map Reduce framework on cloud. Privacy preserving machine learning algorithms for big data systems [14] where the training data are distributed and each shared data portion is of large volume. The Proposed scheme is secure in the novel model. The training data are distributed and not able to achieve distributed feature selection.

Simple vanilla configuration for split learning:

Each client trains a partial deep network up to a specific layer known as the cut layer, the outputs at the cut layer are sent to a server which completes the remaining training without looking at raw data. Without sharing raw data it completes a round of forward propagation. The gradients [1] will be back propagated from its last layer to cut layer. The gradients from cut layer sent to client centres. Without looking raw data this process continues until the distributed split learning network is trained. The fig.1 shows the simple vanilla split learning. But these simple configurations will not suffice for practical setups of collaboration across health entities.

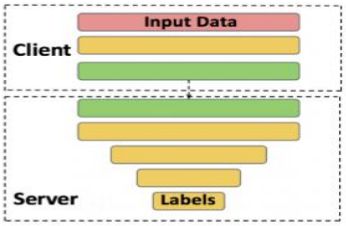


Fig. 1. Simple vanilla split learning

Cyclic parameter delivery learning:

As data set has been increased, for deep learning training mainly using parallelization strategies. To distributely train deep neural networks data parallelization [2] is commonly used which partitions the input samples. If the sharing data is very sensitive, so it have constraints while sharing with multiple nodes. But it is hard to apply such conventional procedures. Instead of train dataset propose a method to sequentially share models on training. Using the data it owns after updating the model parameter θe. This updated model parameter θe passes to the next agent. But it faces an issue that the training datasets are collected by multiple agencies but cannot be shared by other agencies due to some legal and ethical issues.

Revisiting synchronous stochastic optimization:

The third approach, synchronous optimization [3] with backup workers where the parameter servers wait for all workers to send their gradients, aggregate them, and send the updated parameters to all workers afterward. Synchronous distributed stochastic optimization suffer from their respective weaknesses of stragglers and staleness. And also faces some communication overhead between parameter servers.

3. Methodology

The proposed system is a client server system. The system consists of a training phase and prediction phase. In training phase gets the classifiers from the server and fit it with weights which is from the client and the send back the weight to the server. The server receives the trained data. Thus the system prepares an efficient prediction model. Using the developed prediction model hospitals predict the chances of heart diseases. The various hospitals upload the text files which contains patient's details.

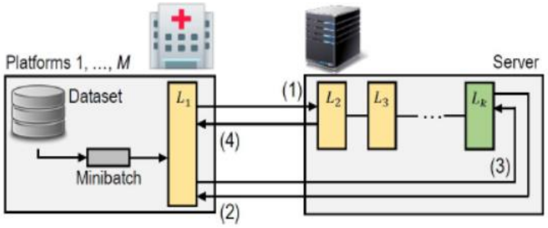


Fig. 2. System architecture

The existing systems mainly uses anonymization techniques to protect the patient details. In the system data servers of hospitals are trained by a central server. Hospitals want to predict the heart disease using a trained model, it needs dataset from another hospitals.so by preserving the confidentiality of the data, the dataset from each hospitals are acquiring by this method. Each hospitals gives its data to a central server after applying a weight L1 using this L1 of each hospital server apply another weights L2….Lk then L1…Lk send to each hospitals then they generate gradients, it again send to server and the server performs back propagation. This repeats until an efficient prediction model develops.

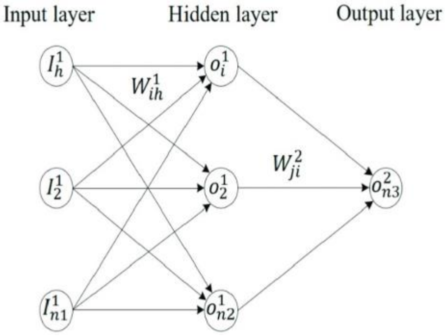


Fig. 3. Structure of a multilayer perceptron (MLP) algorithm

The MLP algorithm (multilayer perceptron) is a class of feedforward artificial neural network (ANN). For training MLP

Table 1
Comparative Analysis

Title	Method	Result accuracy %
Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks (RFNN)	RFNN	96.63
Prediction of Heart Disease Using Neural Network	ANN+BPNN	81
Neural Network Diagnosis of Heart Disease	Multilayer BPNN	85
Heart Disease Diagnosis using Extreme Learning Based Neural Networks	EXTREME LEARNING ANN	79
Prediction of Heart Disease Using Multilayer Perceptron Neural Network	MULTILAYER PERCEPTRON ANN	92

uses a supervised learning technique called backpropagation. The three-layer network, which consist of first layer, last layer and middle layer. First layer is input layer and last layer is output layer and middle layer is hidden layer. We feed the input data into the input layer and take the output from the output layer. Able to increase the number of the hidden layer as much as want, to make the model more complex according to our task. Fig. 3 represents the structure of a MLP algorithm. MLP consists of several fully connected layers. All units in a layer are connected to previous layer. In a fully connected layer, each unit parameters are independent of the rest of the units in the layer. Each unit possess a unique set of weights. Hospitals upload data to the server in the form of csv files and client data get classifiers from the server and train the dataset with the weights and sent back the data to the server. The patient details are uploaded without performing any anonymization. Anonymization of healthcare data means hiding the identifiers from the details or removing the identification details from EHR.

The neural network in this system accepts 11 clinical features as input. It is trained using back-propagation algorithm. The algorithm is used to predict that there is a presence or absence of heart disease in the patient. The system provides highest accuracy of 98% comparative to other systems. Obtained accuracy with this system is better and efficient than other system.

1) Data source

The performance of the system is evaluated on heart disease database that was taken from dataset repository of kaggle. This database consists of records with each having 11 clinical attributes that include name, age, name, systolic blood pressure, idl, adiposity, family history, type, obesity, alcohol,. In this database out of 303 records 164 belong to healthy category and 139 belong to heart disease.

2) Performance evaluation

The system for prediction for heart disease using multilayer perceptron neural network is implemented in DJANGO framework. In this system the database is divided in to two sets randomly that is training set and testing set. Out of total records 70% records are used for training and testing is done by using remaining 30% records. The evaluation of performance of the system is done by computing the percentage value of parameter like Accuracy.

Accuracy= [(TP+TN) / (TP+TN+FP+FN)]*100

Where, TP = classifies number of samples as true while they were true.

TN =classifies number of samples as false while they were actually false

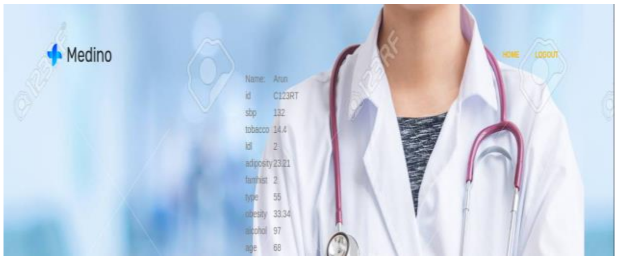
FN = classifies number of samples as false while they were actually true.

FP = classifies number of samples as true while they were actually false.

The prediction system in this paper gives higher accuracy of 98% for 5 neurons in hidden layer. This shows that the prediction system shows higher performance. Table 1 shows the comparative analysis of various methods used in different works and the accuracy of prediction result in percentage.

4. Result and Discussion

The novel prediction system bring out 98% of accuracy. Moreover, the trained model of prediction system developed without disclosing the healthcare data to public. Collaborating the details from multiple hospitals in terms of big data can bring a model with more accuracy compared to that of models that combining details from fewer hospitals.



Name: Arun
id C123RT
sbp 132
tobacco 14.4
ldl 2
adiposity 23.21
famhist 2
type 55
obesity 33.34
alcohol 97
age 68

Fig. 4. Prediction system

Fig. 4 depicts the novel prediction system. The comparison done with other heart disease prediction system proved that our system with multilayer perceptron algorithm providing security and privacy of the data without any anonymization and the accuracy of the system is graphically higher when compared with other prediction model

5. Conclusion and Future Scope

The three approaches discussed in this paper is facing many issues regarding the configuration and accuracy of data. Even facing some legal and ethical issue while sharing data with multiple platforms (hospitals). The existing approaches using anonymization techniques to protect the privacy and security of the healthcare data. The proposed system is not using any anonymization methods. Without masking the personal identities of a patient the data is sending to the training model in server. Thus proposing a novel distributed deep learning framework for accurate prediction model in hospitals without losing security and privacy. The proposed system is focused on clinical attributes only. The future research work focus on working with additional risk factor attributes like Number of major vessels (0-3) coloured by fluoroscopy, Exercise induced angina, resting electrocardiographic results. The slope of the peak exercise ST segment etc. which is quite advantageous for improving the prediction.

References

[1] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, “Split learning for health: Distributed deep learning without sharing raw patient data,” 2018.
[2] J. Jeon, D. Kim, and J. Kim, “Cyclic parameter sharing for privacy preserving distributed deep learning platforms,” in Proc. ICAIIC, 2019.
[3] J. Chen, X. Pan, R. Monga, and S. Bengio, “Revisiting distributed synchronous SGD,” 2016.
[4] F. Tang, W. Wu, J. Liu, H. Wang, and M. Xian, “Privacy-preserving distributed deep learning via homomorphic re-encryption,” Electronics, vol. 8, no. 4, p. 411, 2010.
[5] A. Vizitiu, C. I. Ni¸t˘a, A. Puiu, C. Suciu, and L. M. Itu, “Towards privacy-preserving deep learning based medical imaging applications,” in 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA). IEEE, 2019, pp. 1–6.
[6] H. Surendra and H. Mohan, “Considerations for privacy preserved open big data analytics platform,” in 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). IEEE, 2016, pp. 445–449.
[7] K. Fritchman, K. Saminathan, R. Dowsley, T. Hughes, M. De Cock, A. Nascimento, and A. Teredesai, “Privacy-preserving scoring of tree ensembles: A novel framework for ai in healthcare,”in2018 IEEE International Conference on Big Data (BigData). IEEE, 2018, pp. 2413–2422.
[8] S. Rao, S. Suma, and M. Sunitha, “Privacy preserving computations over healthcare data,” in 2015 Second International Conference on Advances in Computing and Communication Engineering. IEEE, 2015, pp. 510–514.
[9] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in Proceedings of COMPSTAT’2010. Springer, 2010, pp. 177–186.
[10] J. Sedayao, R. Bhardwaj, and N. Gorade, “Making big data, privacy, and anonymization work together in the enterprise: experiences and issues,” in 2014 IEEE International Congress on Big Data. IEEE, 2014, pp. 601–607.
[11] K. Jung, S. Park, and S. Park, “Hiding a needle in a haystack: Privacy preserving apriori algorithm in mapreduce framework,” in Proceedings of the First International Workshop on Privacy and Secuirty of Big Data. ACM, 2014, pp. 11–17.
[12] Sameesha Vs "A Scalable Two Phase Top Down Specialization Approach for Data Anonymization Using Mapreduce On Cloud". International Journal of Computer Trends and Technology, 45(1):45-49, March 2017.
[13] E. Mohammadian, M. Noferesti, and R. Jalili, “Fast: fast anonymization of big data streams,” in Proceedings of the 2014 International Conference on Big Data Science and Computing. ACM, 2014, p. 23.
[14] K. Xu, H. Yue, L. Guo, Y. Guo, and Y. Fang, “Privacy-preserving machine learning algorithms for big data systems,” in 2015 IEEE 35th international conference on distributed computing systems. IEEE, 2015, pp. 318–327.
[15] L. Wei, H. Zhu, Z. Cao, X. Dong, W. Jia, Y. Chen, and A. V. Vasilakos, “Security and privacy for storage and computation in cloud computing,” Information Sciences, vol. 258, pp. 371–386, 2014.
[16] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, “Big healthcare data: preserving security and privacy,” Journal of Big Data, vol. 5, no. 1, p. 1, 2018.