A Comprehensive Analysis of Privacy-preserving Techniques in Deep learning based Disease Prediction Systems

Andrew J¹, Shaun Shibu Mathew², Batra Mohit³

¹,²,³Computer Science and Engineering
Karunya Institute of Technology and Sciences
Coimbatore, India
onesimu@gmail.com¹, shaun.mathew95@gmail.com², mhtbatra005@gmail.com³

Abstract—With the rise in demand for deep learning models due to its ability to learn features from data, and predict, it is widely used in disease prediction systems. However, as patient medical records are considered to be highly confidential due to them consisting of personal information, its privacy-preservation is of prime importance. Conventional privacy-preserving techniques often tend to hinder the utilitarian aspect of the system. In this paper we carry out a comprehensive analysis of privacy-preserving techniques for disease prediction systems that use deep learning along with a comparison of the different privacy-preserving techniques. This paper also discusses the existing privacy-preserving approaches in deep learning. They are cryptographic approaches, attribute-based encryptions, homomorphic encryptions and other hybrid approaches.

Keywords—privacy-preservation, deep learning, disease prediction, homomorphic encryption, differential privacy

1. INTRODUCTION

As the good old saying goes “Health is Wealth” and thus this has become the most important thing in people’s lives. As a result, the ability to predict diseases will lower the factor of risk and enable people to lead a healthier life. In recent times, it has been observed that due to the rapid increase of data mining techniques there has been a rise in disease prediction systems. Data is constantly being shared to enhance the prediction models and to predict diseases. As bright as this may sound, many challenges could limit its use in a practical environment which includes information security and prediction efficiency.

Deep learning is a constituent of the machine learning kindred. Instead of following the traditional task-specific algorithms, they focus on learning data representations. Statistical models are created as output by each algorithm, from the information learned by hierarchically applying a non-linear transformation as its input. Deep learning enables the existence of prediction models that are used to predict disease.

Issues regarding privacy-preservation includes the honest-but-eager attitude of outsourced bodies that may attempt to acquire sensitive patient information from the intermediate or final results. Such opportunities to acquire patient sensitive information must be hindered by the prediction scheme to ensure maximum trust of the patients.

In this paper, we have analyzed different privacy-preserving techniques that are associated with disease prediction schemes which make use of deep learning. This includes the different concepts and algorithms used with respect to privacy-preservation. A comprehensive comparison of the different approaches is offered in this paper.
2. PRIVACY-PRESERVING APPROACHES

A. Cryptographic Approaches

A mechanism, which is clearly the first of its kind, is the K privacy. It was introduced in [1] and enables the presentation of multiple conflicting responses to a query by the data creators, whilst the utility evaluated as the absolute error from the actual original data is preserved. Although homomorphic encryption allows executing different operations on encrypted data, as a standalone implementation to compute exact aggregate statistics, it fails to deliver in terms with privacy protection. As a solution, such cryptographic implementations are combined with differentially private mechanisms to protect data owners. The K privacy mechanism proves to upgrade the strength of the privacy protection whilst maintaining utility. It also provides protection against pollution attacks.

In order to exponentially reduce the sensitivity of a query, [2] has introduced the use of a new differentially private decision forest algorithm that also reduces the number of queries that are required. This scheme, as a result, minimizes the amount of noise that must be appended in order to preserve privacy. This is made possible by raising a query that brings out most of the recurring label in some subset $x_i$ of the data which consists of a high probability. This query is used in all of the leaf nodes in every tree of the forest. The sensitivity is improved by the use of Smooth Sensitivity, which considers the specific data used in the query instead of assuming the worst-case scenario.

In order to bring about privacy preservation in cluster analysis of IoMT (Internet of Medical Things), [3] introduces an Efficient Differentially Private Data Clustering scheme (EDPDCS) that is based off the MapReduce framework. EDPDCS optimizes the allocation of privacy budgets and the selection of initial centroids. This brings about an improvement in the accuracy of differentially private K-means clustering algorithm. According to the minimal privacy budget of each iteration and the total privacy budget, the K-means algorithm’s iteration count is set to fixed value. An enhanced initial centroids selection method which picks out a minimal selection of the dataset and performs rough clustering initially to select the primary centroids, is suggested to upgrade the regulation of the clustering algorithm.

In order to exchange person-specific information, [4] brings about a privacy-preserving data mashup model. This is a very similar application like that in a disease prediction system. The privacy preservation contraption used here can be adopted for the same. Anonymization algorithms, such as the Top-Down Specialization (TDS) and Differentially private anonymization based on Generalization (DiffGen) for relational data mashup from multiple data providers are employed in this scheme. However, a limitation of this model is the absence of a typical method to monetize the value of personal data, typically when multiple parties are involved in fetching person-specific information from the same population.
The EPDP (an Efficient and Privacy-Preserving Disease Prediction scheme) was introduced in [5] in order to resolve most of the popular drawbacks faced in a disease prediction scheme. In this scheme, privacy is attained by encrypting the historical medical data of every validated patient using the Okamoto-Uchiyama (OU) cryptosystem in advance to outsourcing. In order to generate the Bloom filter and the query element, a keyed-cryptographic hash function is used by the undiagnosed patient and the healthcare provider.

[6] discusses about the potential privacy preserving techniques that can be used in a modern IoT-based healthcare system. It focuses on the multiple aspects of a system, such as the predictive model, patient’s health record, etc. The sought after result is to bring about a predictive model that is beneficial in every aspect. A cryptographic service provider can be introduced into the framework depending on the level of privacy and the analytics desired. After confirming the correctness of the system and stabilizing the corresponding cost of computation, storage and communication, privacy techniques such as data perturbation, homomorphic encryption schemes and differential privacy can be introduced into the framework.

B. Attribute Based Encryption

There is an increased demand for wearable technology along with other electronic devices for collecting and diagnosing health data. This data being stored in the cloud gives access to doctors and hospitals to keep track of a patient’s past and present health situation. In order to assure patient confidentiality, it is highly significant that the data is restricted to only authorized personals. In [7], a secure and efficient fine-grained access architecture for collecting and accessing data generated by the BAN (Body Area Network).

On top of only granting authorized personals to access data, this scheme also allows certain physicians to even write records. The ABE (Attribute-Based Encryption) applications in healthcare system of PHR record sharing in cloud computing is considered in this paper. The model is enhanced to bring about a more secure and efficient data access at the user terminal. This is brought forward by introducing the “matching” algorithm before decryption and modifying the key generation in LSSS based CP-ABE construction.

A fully integrated IoT platform for prediction of ventricular arrhythmia which makes use of ECG signals that is also secure and ultra-low power is introduced in [8]. In this scheme, the communication channels are protected by a chip-specific ECG key that is extracted from the ECG signals by the suggested processor. On homogenizing a prevailing design-for-trust solution with the generated ECG key, this scheme that is put forward, can offer protection even at the hardware level, which foils hardware security threats much like that of reverse engineering and counterfeiting.

C. Homomorphic Encryption

An uncomplicated yet highly reasonable classifier is the Naive Bayes (NB). The STC (Substitution-Then-Comparison) attack, which is an easy-to-perform but difficult-to-detect attack. [9] focuses primarily on making an effort to construct a privacy-preserving NB classifier that is resistant to the STC attack. In order to hide the privacy of both the user and the server, the proposed essential technique incorporates the use of a “double-binding” technique along with the integration of additively homomorphic encryptions and oblivious transfer.

Being able to predict diseases and to get precise treatment options at a preliminary phase helps people to steer a healthier life. The DPSS (Disease Prediction Support System) with the inclusion of machine learning algorithms are able to predict diseases on the basis of physiological signs and symptoms. Due to the increasingly high popularity of such DPS Systems, data security and privacy are now prime concerns that have to be dealt with. As a result, to help preserve patient sensitive data, the Privacy-Aware Disease Prediction Support System (PDPSS) [10] was developed using FKNN-CBR (Fuzzy k-Nearest Neighbor
classifier and Case-Based Reasoning classifier). In order to secure the patient’s sensitive data, the paillier homomorphic cryptosystem is used. Over standard evaluation results, the FKNN-CBR has manifested better performance when compared to Naïve Bayesian, SVM and SLNN.

The use of Privacy Preserving Collaborative Filtering (PPCF) on ADD (Arbitrary Distributed Data) has a couple of limitations. A few being that they only consider two parties and that their off-line model generation has a high cost of computation. As a solution to these drawbacks, [11] introduces PPCF scheme on ADD based on polynomial aggregation techniques and multi-party random masking. For the preservation of privacy, three protocols have been used. Only homomorphic encryption’s additive property is used as a result of calculating the length of vector X by using the Paillier homomorphic encryption system. Upon evaluation, it has been concluded that this scheme is secure and that the accuracy and coverage of prediction generation have been upgraded, which is attributable to the collaboration of different parties.

Being able to detect peculiarities on complex, large and dynamic data is considered to be highly indispensable with today’s increased dependance on cloud-computing infrastructures to process data. They pose a threat in terms of privacy and security. [12] introduces a framework that exploits cloud resources to offer a privacy preserving anomaly detection service for sensitive data. A lightweight homomorphic encryption (Domingo-Ferrer) is made use of in order to sensor the data before it is analyzed in the cloud. Nonetheless, due to the contemporary limitations of homomorphic encryption, some operations in the anomaly detection system will have to be done decrypted. Therefore, these operations are carried out on a trusted private server.

[13] explores the conceivable chain of events for homomorphic encryptions as a result to ensure the privacy of sensitive medical data. The differentiating element of this model is that for real-life predictive analysis, a leveled homomorphic encryption scheme is used. “Leveled” implies that the homomorphic encryption scheme is unable to accurately and securely conduct an arbitrary computation. As an alternative, the scheme is only able to compute functions of a certain complexity, which is selected prior to the computation. In order to determine safe parameters for the practical homomorphic encryption scheme, an automatic parameter selection module is introduced. This guaranteed the flawlessness and safety of the outcome while the functions used in predictive analysis were analyzed. It also includes Cox proportional hazard regression and logistic regression.

In [14], the POMP (efficient and Privacy-Preserving Online Medical Prediagnosis) scheme is introduced for the cloud context. This primarily makes use of the BGN (Boneh-Goh-Nissim) homomorphic encryption. This scheme allows for the online prediagnosis service to be directly applied on the ciphertext. In order to escalate the prediagnosis process, the Bloom Filter along with the preprocessing technique is used. Upon real world evaluation, the scheme had proved to be highly efficient and secure.

[15] introduces the PCD (Privacy-preserving Predictive Clinical Decision) scheme. This scheme is based off of the RNN (Recurrent Neural Networks) in order to create predictive clinical decisions. A modified version of the Paillier Homomorphic encryption, that allows multiplication function, is used to ensure privacy of the patient’s data. In order to achieve real-time disease prediction, the scheme deploys a sequential RNN model and an averaged RNN model. The scheme upon evaluation, has proved to be highly efficient and accurate whilst preserving privacy.

The PPDM scheme, i.e., Privacy-Preserving Data Mining, which is introduced in [16], allows the identification of the most important factors associated with diseases using linear multiple regression. A substantial overhead was found upon evaluation due to the use of public homomorphic encryptions. However, [16] claims that with the use of modern encryption technologies like the lattice encryption, the suggested multiple regression protocol is practical to carry out arbitrary epidemically analysis over distributed datasets when in a secure network.
D. Hybrid Approaches

Privacy issues were also found to surface from the neural network models and sensitive data used. The realization of the prediction of the privacy-preserving neural network, as a result, has become a subject that is highly sought after. While making use of two non-colluding servers, an outsourcing model for a privacy-preserving prediction was introduced in [17]. The privacy-preserving prediction service is offered to the public clients by the servers, who were securely supplied with a neural network model that had already existed by the original neural network owner. From their evaluation, it was evident that the model owner’s computational overhead was progressively related with the model size. However, it had proved to be highly efficient. While being fully non-interactive, the system had also satisfied the security definition of data and model privacy.

A Restricted Sensitive Attributes-based Sequential Anonymization (RSA-SA) approach for privacy-preserving data stream publishing is put forward in [18]. Furthermore, two alternative privacy restrictions were also introduced, these being the semantic and sensitivity diversities. On using the RSA-SA approach, these restrictions obstruct the Sensitive Attributes (SAs) that were published. The input tuples’ each value is replaced by the RSA-SA approach with a set of sensitive values containing both the original and randomly selected noise values. The released tuples’ anonymized SA value is represented by the sensitive-values set of every SA. This approach offers an invulnerable system in terms of privacy preservation.

In [19], a framework in fog–cloud computing for hybrid privacy-preserving clinical decision support system, called HPCS is introduced. In order to securely monitor patients’ health condition, a real-time, lightweight data mining method is made use of by a fog server in HPCS. HPCS consists of seven parties which include the Key Generation Center (KGC), Global Computation Service Provider (GSP), Cloud Storage Center (CSC), Fog Server (FS), Local Computation Service Provider (LSP), Monitored Patient (MP) and Health Service Provider (HP). In order to achieve a secure and efficient model, the HPCS uses three phases. These three phases are user physiological status monitoring and processing, real time processing in fog servers and high accuracy clinical decision in cloud server.

The PPDP (efficient and Privacy-Preserving Disease Prediction) scheme discussed in [20] promises an efficient and secure scheme for disease prediction. They are subjected to two main phases. The first phase focuses on the disease learning aspect of the system. The cypher texts for all the health data along with the prediction model of every disease is generated and sent to the cloud server. The cloud server then executes the learning procedure on the encrypted data. The second phase includes the disease prediction phase. During this phase, the hospital sends the encrypted EHR (Electronic Health Record) of the patient who wishes to get diagnosed to the cloud server. With the already existing encrypted prediction models, the cloud server calculates the information without decrypting the data. This ensures a highly secure and efficient disease prediction scheme.
Table 1: Comparison of different privacy-preserving techniques

| Reference | Privacy-Preserving Techniques                                      | Advantage(s)                                                                 | Disadvantage(s) | Dataset                  |
|-----------|-------------------------------------------------------------------|-------------------------------------------------------------------------------|-----------------|--------------------------|
| [1]       | Differential Privacy, K-Anonymity, Homomorphic Cryptosystem       | 1. Highly scalable                                                          | —               | 1. Heart related Dataset |
| [2]       | Differential Privacy, Smooth Sensitivity                          | 1. Classifier with over 85% predictive accuracy                              | —               | 1. WallSensor            |
|           |                                                                   |                                                                                |                 | 2. PenWritten            |
|           |                                                                   |                                                                                |                 | 3. GammaTele             |
|           |                                                                   |                                                                                |                 | 4. Mushroom              |
|           |                                                                   |                                                                                |                 | 5. Claves                |
|           |                                                                   |                                                                                |                 | 6. Nursery               |
| [3]       | Differentially private k-means clustering algorithm               | 1. Optimizes the allocation of privacy budgets.                              | —               | 1. Blood Dataset         |
|           |                                                                   | 2. Improved accuracy of differentially private K-means clustering algorithm.  |                 | 2. Adult Dataset         |
| [5]       | Okamoto-Uchiyama Cryptosystem, Keyed-cryptographic hash function  | 1. Reduced communicational and computational overheads.                       | 1. Message integrity | 1. Acute inflammations dataset |
| [7]       | Attribute-Based Encryption, “matching” algorithm                  | 1. Fast decryption for the users                                             | —               | —                        |
|           |                                                                   | 2. Reduces private key overhead                                               |                 | —                        |
| [8]       | Specific ECG key                                                  | 1. Lessening of 62.2% in power consumption.                                  | 1. Prone to hardware attacks which includes side-channel attacks on cryptographic algorithms. |
|           |                                                                   | 2. Reduction of 16.0% in area when compared to other state-of-the-art processors. |                 | 2. Low protection when it comes to the insertion of |
| Reference | Privacy-Preserving Techniques | Advantage(s)                                                                 | Disadvantage(s)                                                                 | Dataset                                      |
|-----------|-------------------------------|-------------------------------------------------------------------------------|--------------------------------------------------------------------------------|---------------------------------------------|
|           |                               | different classes of hardware Trojans.                                        |                                                                              |                                             |
| [9]       | Double-Binding, Homomorphic Encryption | 1. Avoids Information Leakage  
2. Independent from fully homomorphic encryption and garbled-circuits framework. | 1. When used on a multi-class NBC will cause information leakage.               | 1. Breast Cancer Wisconsin (Original) Dataset  
2. Statlog Heart Dataset |
| [10]      | Paillier Homomorphic Cryptosystem, FKNN-CBR | 1. Offers 90.42% specificity, 99.28% sensitivity and 96.74% prediction accuracy when compared to existing baselines. | 1. Computational and communicational costs are still considerably high.         | 1. Indian Liver Patient dataset |
| [11]      | Paillier Homomorphic ENcryption | 1. Comparatively faster online prediction and off-line model generation.       | —                                                                             | 1. Healthcare  
2. Movieslens |
| [12]      | Homomorphic Encryption (Domingo-Ferrer) | 1. Scalable  
2. Enables lightweight arithmetic computations to be performed. | 1. Communication overheads are still considerably high.  
2. Adapted homomorphic techniques needs further enhancement. | 1. Intel dataset |
| [13]      | Homomorphic Encryption, Automatic parameter selection module | —                                                                            | 1. Scalability                                                                | 1. Real Dataset |
3. CONCLUSION

As it has been observed that the privacy of patient medical records are required to be maintained and handled securely, the privacy-preservation of patient medical data is to be dealt with utmost importance. In this paper, we have given out a comprehensive analysis of the different privacy preserving techniques that use deep learning along with a comparison of the different approaches taken to preserve privacy. These approaches include homomorphic encryption, differential privacy, attribute-based encryption and other hybrid approaches.

4. REFERENCES

[1]. Joy, J., et al., K Privacy: Towards improving privacy strength while preserving utility. Ad Hoc Networks, 2018. 80: p. 16-30.

[2]. Fletcher, S. and M.Z. Islam, Differentially private random decision forests using smooth sensitivity. Expert Systems with Applications, 2017. 78: p. 16-31.

[3]. Guan, Z., et al., Achieving data utility-privacy tradeoff in Internet of Medical Things: A machine learning approach. Future Generation Computer Systems, 2019.

[4]. Khokhar, R.H., et al., Privacy-preserving data mashup model for trading person-specific information. Electronic Commerce Research and Applications, 2016. 17: p. 19-57.

[5]. Yang, X., et al., An Efficient and Privacy-Preserving Disease Risk Prediction Scheme for E-Healthcare. IEEE Internet of Things Journal, 2018: p. 1-1.

[6]. S. Sharma, K.C.a.A.S., Toward Practical Privacy-Preserving Analytics for IoT and Cloud-Based Healthcare Systems. IEEE Internet Computing, 2018. 22: p. 42-51.

[7]. Liu, X., et al., Secure and efficient querying over personal health records in cloud computing. Neurocomputing, 2018. 274: p. 99-105.

[8]. Yasin, M., et al., Ultra-Low Power, Secure IoT Platform for Predicting Cardiovascular Diseases. IEEE Transactions on Circuits and Systems I: Regular Papers, 2017. 64(9): p. 2624-2637.

[9]. Gao, C.-z., et al., Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack. Information Sciences, 2018. 444: p. 72-88.

[10]. D, M., et al., Hybrid Reasoning-based Privacy-Aware Disease Prediction Support System. Computers & Electrical Engineering, 2019. 73: p. 114-127.

[11]. Kaur, H., N. Kumar, and S. Batra, An efficient multi-party scheme for privacy preserving collaborative filtering for healthcare recommender system. Future Generation Computer Systems, 2018. 86: p. 297-307.

| Reference | Privacy-Preserving Techniques | Advantage(s) | Disadvantage(s) | Dataset |
|-----------|-------------------------------|--------------|-----------------|---------|
| [17]      | Non-Colluding Servers         | 1. Realizes non-interactivity for the client | —               | 1. Real Dataset |
[12]. Alabdulatif, A., et al., Privacy-preserving anomaly detection in cloud with lightweight homomorphic encryption. Journal of Computer and System Sciences, 2017. 90: p. 28-45.

[13]. Bos, J.W., K. Lauter, and M. Naehrig, Private predictive analysis on encrypted medical data. Journal of Biomedical Informatics, 2014. 50: p. 234-243.

[14]. Guo, W., et al., A Privacy-Preserving Online Medical Prediagnosis Scheme for Cloud Environment. IEEE Access, 2018. 6: p. 48946-48957.

[15]. J. Lin, J.N.a.H.L., PCD: A privacy-preserving predictive clinical decision scheme with E-health big data based on RNN. 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Atlanta, GA, 2017: p. 808-813.

[16]. Kikuchi, H., et al., Privacy-Preserving Multiple Linear Regression of Vertically Partitioned Real Medical Datasets, in 2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA). 2017: p. 1042-1049.

[17]. Ma, X., X. Chen, and X. Zhang, Non-interactive privacy-preserving neural network prediction. Information Sciences, 2019. 481: p. 507-519.

[18]. Abdelhameed, S.A., S.M. Moussa, and M.E. Khalifa, Restricted Sensitive Attributes-based Sequential Anonymization (RSA-SA) approach for privacy-preserving data stream publishing. Knowledge-Based Systems, 2019. 164: p. 1-20.

[19]. Liu, X., et al., Hybrid privacy-preserving clinical decision support system in fog–cloud computing. Future Generation Computer Systems, 2018. 78: p. 825-837.

[20]. Zhang, C., et al., PPDP: An efficient and privacy-preserving disease prediction scheme in cloud-based e-Healthcare system. Future Generation Computer Systems, 2018. 79: p. 16-25.

[21]. Mv Ngo Tien HoA, High Speed And Reliable Double Edge Triggered D-Flip-Flop For Memory Applications”.Journal of VLSI Circuits And Systems, 1 (01), 13-17,2019

[22]. MN BORHAN,”Design Of The High Speed And Reliable Source Coupled Logic Multiplexer”,Journal of VLSI Circuits And Systems 1 (01), 18-22,2019