# Privacy and Security in Intelligent Decision Support Systems

Saghi Khani

University of Windsor

December 2019

# Contents

# 1 Abstract

Decision support is a crucial function for decision-makers in many industries and help them to gather and interpret information and build a foundation for decision making. These kinds of systems use different high-quality approaches to keep users medical data secure. However, they still face many challenges, including information security and privacy issues. In this survey, different methods are proposed to keep medical data safe for both clients and service providers.

# 2 Introduction

In healthcare, Intelligent decision support systems (IDSS) can play a significant role. Different machine learning algorithms and professional data mining techniques are used to create intelligent decision support systems in medical areas to help physicians to diagnose diseases more accurately. Intelligent decisions usually are taken by healthcare service providers based on clinical guidance and evidence-based rules derived from medical science and also these systems make extensive use of artificial intelligence techniques. The advantages of intelligent decision support systems in healthcare are countless. For instance, reducing diagnosis time [Liu et al., 2015], improving diagnosis accuracy [Liu et al., 2015], lowering health care costs and decreasing medical appointments waiting time [Liu et al., 2018] are some of these systems advantages. In this survey, four papers are discussed which provide various privacy-preserving methods for both clients and service providers. The suggested methods are based on fully homomorphic cryptography and secure multiparty computations.

# 3 Literature Review

In this part, suggested methods in four papers are introduced and compared with each other, which each of them proposed a new approach for data privacy-preserving in medical decision support systems.

## 3.1 Privacy-Privacy Patient-Centric Clinical Decision Support System on Naive Bayesian Classification

In Privacy-Preserving Patient-Centric Clinical Decision Support System on Naive Bayesian Classification [Liu et al., 2015], the authors proposed a privacy-preserving patient-centric clinical decision support system, called PPCD which stores past patient's medical data in a cloud to train the Naive Bayesian Classifier without any data disclosure; then the trained classifier starts to compute the risk of different diseases for new patients and allow patients to retrieve the top-k disease names based on their preferences. In this system, to protect a patient's

old medical data, a new cryptography method called "Additive Homomorphic Proxy Aggregation" (AHPA) is designed. In addition, to protect both parties (clients and service providers), a privacy-preserving TOP-K disease names retrieval protocol is introduced. The customized TOP-K is much more efficient in terms of computation cost and communication overhead in comparison with the existed privacy-preserving top-k protocol. AHPA is based on El-Gammal homomorphic encryption and the Top-K protocol is based on paillier homomorphic encryption. Figure 1, presents the system model for this method.

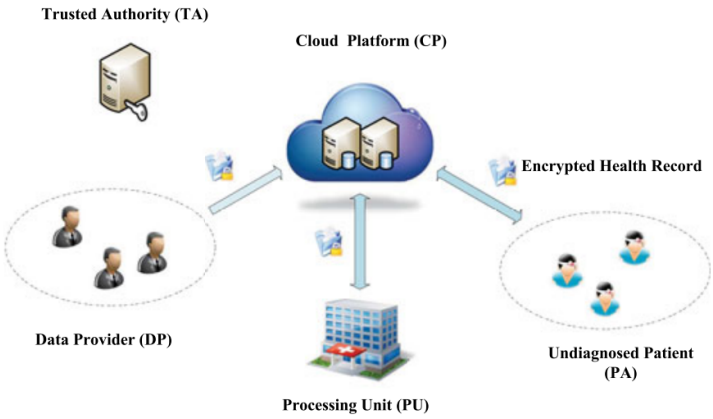


Figure 1: System model under consideration

In the proposed system, the authors mainly focus on how to have a secure trained naive Bayesian classifier and use it to make decisions about patients disease without leaking their private medical information. The whole PPCD model divided into five parties: Trusted authority (TA), Cloud Platform (CP), data provider (DP), processing unit (PU) and undiagnosed patient (PA).

1. Trusted Authority: TA is the indispensable part, which it distributes and manages all the private keys, and also it is trusted by all entities in the system.
2. Cloud Platform: CP contains unlimited storage space, which can store and manage all the data in the system. Moreover, it can do some computation and calculation over some data.
3. Data Provider: This part includes all the historical medical data from different patients to train the naive Bayes classifier.
4. Processing Unit: Hospitals or other companies which provide online direct-to-consumer service and give a prediction to users based on their disease symptoms.

5. Undiagnosed Patient (PA): PA saves some symptoms which the physician collected during the diagnosing process or the patient can provide them directly.

## 3.2 Efficient Privacy-Preserving Online Medical Primary Diagnosis Scheme on Naive Bayesian Classification

In Efficient Privacy-Preserving Online Medical Primary Diagnosis Scheme On Naive Bayesian Classification [Liu et al., 2018], the authors suggested P-Diag, which stands for primary diagnosis scheme to keep sensitive personal health data securely. Also to improve the Naive Bayes classifier, an efficient privacy-preserving classification scheme with lightweight polynomial aggregation technique is proposed. The whole model is shown in the figure below:

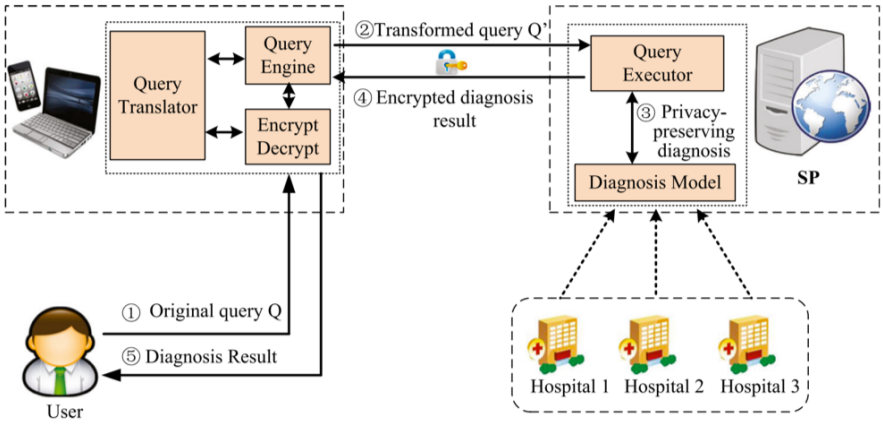


Figure 2: System model under consideration

As can be seen, the suggested model has two main parts: a user and a service provider (SP). In the client side, the user creates his original query vector includes his sensitive information, and the query is encrypted with an arbitrary asymmetric encryption algorithm in his side. In the next step, the query is transformed to the service provider. The service provider considered as an authorized data analysis organization that has a naive Bayes classifier to classify clinical datasets provided by hospitals, and it provides online medical primary diagnosis service for registered users. The naive Bayes classifier in SP analyzes the encrypted data and transfer the encrypted diagnosis result to the client. In this model, digital signature and hash function are used to validate messages including queries, public keys and timestamps.

### 3.3 CINEMA: Efficient and Privacy-Preserving Online Medical Primary Diagnosis With Skyline Query

In Efficient and Privacy-Preserving Online Medical Primary Diagnosis With Skyline Query [Hua et al., 2018], the authors presented an efficient and privacy-preserving online medical primary diagnosis (CINEMA) framework which users have access to their medical diagnosis service without divulging their medical data. CINEMA can ensure that user's health data and healthcare SP's diagnosis model are kept secure. This suggested model is based on skyline query [Borzsony et al., 2001]. Skyline query is an operator that returns a set of interesting points which are the best tradeoffs between the different dimensions of a large database. The authors believe that medical data are high dimensional and skyline operator would work better with these kinds of data. The suggested model, like the previous model, consisted of two parts: a user and a service provider.

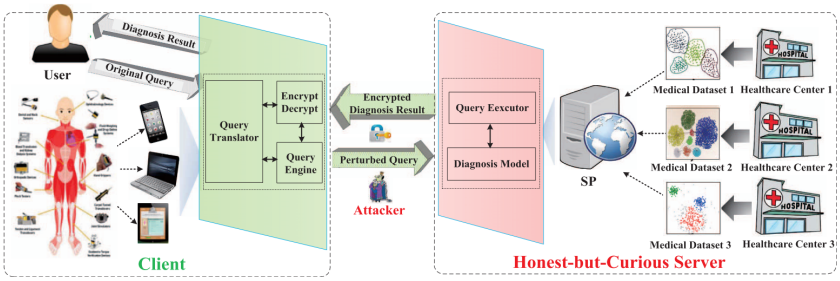


Figure 3: System model under consideration

As shown in the figure above, the registered user sends a query request based on information collected from wearable sensors or devices. This query includes sensitive medical data. Before sending the query to the SP, the medical data is transformed into ciphertext and operated without decryption during the diagnosis process to avoid privacy leakage. An arbitrary encryption method is used to encrypt the query and also the encrypted query will perturb before sends to the SP. After some processes in the SP, the encrypted answer without any decryption will send to the user. The whole encryption and decryption operation is in the client-side.

### 3.4 Reliable Medical Recommendation Based on Privacy-Preserving Collaborative Filtering

In Reliable Medical Recommendation Based on Privacy-Preserving Collaborative Filtering [Hou et al., 2018], the authors proposed a Privacy-Preserving Medical Recommendation (PPMR) algorithm, which can protect patients' treatment information and demographic information during the online recommenda-

tion process without compromising recommendation accuracy and efficiency. The model involves two privacy-preserving operations: Private Neighbor Selection and Neighborhood-based Differential Privacy Recommendation. Private Neighbor Selection is based on the k-anonymity concept, which means that it collects users' historical behaviours and users' basic information to identify the users of k nearest neighbours. Private Neighbor Selection helps to select k neighbours from a candidate list in a secure way. Before doing any anonymization, direct identifiers (name, ID number, etc.) need to be suppressed from the dataset. However, some of the attributes that remain in the anonymized dataset may be quasi-identifiers, which may lead to indirect reidentification. Therefore a method called Optimal Lattice Anonymization (OLA) is used for quasi-identifiers anonymization.
Neighbourhood-based Differential Privacy Recommendation is introduced to predict the rating by aggregating the ratings on those items that identified neighbour users rated. Moreover, the k-nearest neighbors (KNN) algorithm is used as the classifier in this method.

# 4 Comparison

## 4.1 PPCD Performance Evaluation

The authors used paillier homomorphic cryptosystem. As the main advantage, these cryptosystems have high security and users can retrieve their medical diagnosis results without any data disclosure. In addition, their information is safe when they sending data to the SP. However, the performance of the homomorphic cryptosystems are often a disadvantage. Medical information is massive data, therefore, it has high overhead in computation and communication.

## 4.2 PDiag Performance Evaluation

The two figures below depict the overhead computation varying with the number of disease classes in user and SP. The authors believe that the dimension of the query is 20 the average dimension of statistic vectors is 100. With comparing Figs 3 and 4, it is evident that with the increase of the numbers of disease classes, the computation overhead of CDSS is much higher than that of the PDiag scheme. Although the computation overhead of our proposed PDiag scheme also increases when the number of disease classes is large, it is still much lower than that of CDSS.

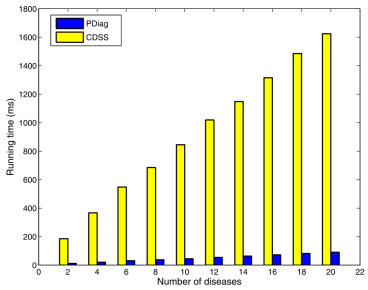


Figure 4: Average running time of SP in PDiag and CDSS

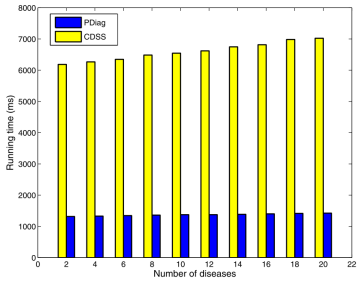


Figure 5: Average running time of user in PDiag and CDSS

PDiag is a fast model because it is based on a lightweight polynomial aggregation techniques and does not use homomorphic cryptosystems. However, the Service Provider in the presented work is not efficient in case of having many users and it is considered as a disadvantage.

## 4.3 CINEMA Performance Evaluation

CINEMA framework can achieve efficient medical diagnosis with low computation complexity for users and service providers. The advantage of CINEMA in comparison with CDSS is less overhead as can be seen in the figures above, the computation overhead varying with the dimension of the query vector and the number of vectors in SP and the user, and the authors assume the dimension of each vector is 10.

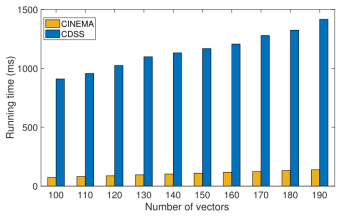


Figure 6: Average running time of SP in CINEMA and CDSS

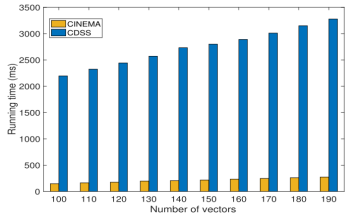


Figure 7: Average running time of user in CINEMA and CDSS

CINEMA uses perturbation which unlike k-anonymity[Sweeney, 2002] and l-diversity [Machanavajjhala et al., 2006], does not rely on sensitive client's data, unlike randomization approaches, this method does not distort the data that might cause misdiagnosis, or unlike homomorphic encryption techniques, this method does not have a high overhead of computation. However, the model is not perfect and requires the implementation of Skyline Query in databases which can be considered as a disadvantage.

### 4.4 OLA Performance Evaluation

The main advantage of OLA is that in comparison with current de-identification methods, it has faster performance and has less information loss in medical datasets. However, as a disadvantage, it may face re-identification based on quasi-identifiers.

## 5 Conclusion

In conclusion, four methods compare in this survey. PPCD uses Additive Homomorphic Proxy Aggregation (AHPA) based on paillier homomorphic encryption and the customized TOP-K, which is base on El-Gammal encryption. PDiag uses an arbitrary asymmetric encryption algorithm, and the whole encryption process is based on lightweight polynomial aggregation techniques. CINEMA uses Skyline Query, which is a query that returns a set of interesting points which are the best tradeoffs between the different dimensions of an extensive database. OLA workes based on k-anonymization, and it uses the KNN algorithm as the classifier while other studies used naive Bayes classifier. Based on studies in [Borzsony et al., 2001] and [Hou et al., 2018] a table like below concluded.

| Method | Scalability | Performance | Accuracy | Privacy |
|---|---|---|---|---|
| Cryptography | Low | Low | High | High |
| Perturbation | High | Moderate | Moderate | Moderate |
| Randomization | High | High | Low | Low |
| Anonymization | High | High | Low | Moderate |

I believe that although cryptography is a weak method in terms of performance and scalability, it is the best approach to keep medical data safe. Physicians need accurate medical data to diagnose different diseases, and cryptography is the most precise method in comparison with others like anonymization or perturbation. However, in my opinion, we can use different approaches based on our need in various situations.

# References


[Borzsony et al., 2001] Borzsony, S., Kossmann, D., and Stocker, K. (2001). The skyline operator. In *Proceedings 17th international conference on data engineering*, pages 421–430. IEEE.

[Hou et al., 2018] Hou, M., Wei, R., Wang, T., Cheng, Y., and Qian, B. (2018). Reliable medical recommendation based on privacy-preserving collaborative filtering. *Computers, Materials & Continua*, 56(1):137–149.

[Hua et al., 2018] Hua, J., Zhu, H., Wang, F., Liu, X., Lu, R., Li, H., and Zhang, Y. (2018). Cinema: Efficient and privacy-preserving online medical primary diagnosis with skyline query. *IEEE Internet of Things Journal*, 6(2):1450–1461.

[Liu et al., 2015] Liu, X., Lu, R., Ma, J., Chen, L., and Qin, B. (2015). Privacy-preserving patient-centric clinical decision support system on naive bayesian classification. *IEEE journal of biomedical and health informatics*, 20(2):655–668.

[Liu et al., 2018] Liu, X., Zhu, H., Lu, R., and Li, H. (2018). Efficient privacy-preserving online medical primary diagnosis scheme on naive bayesian classification. *Peer-to-Peer Networking and Applications*, 11(2):334–347.

[Machanavajjhala et al., 2006] Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). l-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24. IEEE.

[Sweeney, 2002] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.