ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 5, May 2017

Privacy Preserving Top-K Disease Names Retrieval Method for Clinical Decision Support System

[1] Shruti B Karki, [2] Mrs. Anitha K, [3]Mrs. Nandini G
[1] PG scholar, Department of CSE, RRCE, Bengaluru-74
[2][3] Assistant Professor, Department of CSE, RRCE, Bengaluru-74

***Abstract*— Clinical Decision Support System, using the advanced Data Mining techniques help the clinicians to make proper decisions, has obtained a huge attention recently. One of the advantages of clinical decision support system include not only promoting diagnosis accuracy but also reducing diagnosis time. Even though the clinical decision support system is quite challenging, the twist of the system still faces many challenges including information security and privacy concerns. In this paper, we propose a new "Privacy Preserving Top-K Disease Names Retrieval Method for Clinical Decision Support System" that helps clinician complementary to diagnose the risk of patient's disease in a privacy-preserving way. A new cryptographic tool called Additive Homomorphic Proxy Aggregation scheme is designed to protect the privacy of past patient's historical data. The performance of the Proposed System is efficient in calculating the disease risk of the patients in the privacy preserving way.**

***Keywords:* Patient centric, Privacy preserving, Clinical Decision Support System (CDSS)**

I. INTRODUCTION

The evolution of data mining techniques over the last two decades, imposed an important impact on the human lifestyle by providing many features like predicting the behaviors and future trends on everything. This helped in converting the stored data into the useful information. Because of these reasons, the data mining techniques are very helpful in the Clinical decision Support System (CDSS). To decrease the time of diagnosis and to increase the diagnosis accuracy, a new method has to be imposed in the health industry. In the Clinical decision Support System(CDSS), the combination of data mining techniques are used to assist the physicians to diagnose the patients with similar symptoms. This received the major attention now a days. One of the Machine Learning tool called Naive Bayesian classifier is used to predict the diseases in the Clinical decision Support System (CDSS). This is more appropriate for medical diagnosis in the health care. The usage of Naïve Bayesian Classifier in the health care system offers many applications over the available traditional health care systems. This makes easier for the clinicians to predict the diseases of the patients. But its usage still limits in managing and understanding the information security and privacy challenges, during the disease decision phase of the patient. The main challenge is how to keep patient's medical data safe and away from the unauthorized parties. The usage of medical data available online can be more appropriate for the variety of the health care systems. But without the protection of the patients medical data, patient will be worried that his medical data can get leaked or damaged by the unauthorized parties. So the chances are that the patient can refuse to share his medical data to the CDSS for the diagnosis. Therefore it is very important to keep the medical data safe and away from the unauthorized parties. This paper mainly concentrates to address the privacy issues present in the CDSS. We propose a "Privacy Preserving Top-K Disease Names Retrieval Method for Clinical Decision Support System" which uses the algorithms like Additive Homomorphic Proxy Aggregation scheme, Naive Bayesian Classifier, Privacy Preserving Top-K disease names retrieval method, Privacy Preserving maximum out of n protocol that helps in predicting the diseases in a privacy preserving way. Firstly, we propose a secure method that makes the service provider to predict the disease of an individual without leaking the details of a disease to the other parties that are unauthorized. To train the Naïve Bayesian Classifier, the previous patients medical data is used. Using this medical data, the service provider can diagnose patient's disease in a privacy preserving way based on his symptoms. Hence the patients can rescue the results that are diagnosed according to his own priority without affecting the privacy of the service provider.

As the patient's historical medical data is used by the service provider, to maximize the privacy we bring out a new aggregation technique called Additive Homographic

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 5, May 2017

Proxy Aggregation (AHPA) scheme. It mainly helps the service provider to provide the useful data to the Naïve Bayesian Classifier without leaking the data in any way. Even though the cloud platform and the service provider crash, no unauthorized parties can get the information expect for the owner. Because only the aggregated data can be accessed by the service provider.

To procure for the retrieval of the disease names by the patients, we propose a privacy preserving top-k disease names retrieval method. This makes the patients to retrieve the disease names based on their priorities. With this scheme, neither the service provider will get any information about the symptoms of the patients nor the patient gets any information about the Naïve Bayesian Classifier. Even though there are other privacy preserving top-k protocols, our proposed system is deeper efficient in stipulation of communication overhead and computation cost.

Finally to justify the efficiency of the proposed system, we develop a custom simulator built in Java. The simulation demonstrates that our proposed system can efficiently aid the patients to diagnose the disease with high degree of successfulness along with maximizing the privacy without over turning the entire system.

The rest of the paper is categorized as mentioned. Section II is all about the related work, Section III is meant for System Architecture and Design Goals, Section IV provides with the explanation of the Proposed System and section V gives the Performance Analysis and finally we will provide the conclusion in Section VI.

II. RELATED WORK

"Computer-assisted Decision Support for the Diagnosis and Treatment of Infectious Diseases in Intensive Care Units" was developed by C. Schurink, P. Lucas, I. Hoepelman, and M. Bonten. This paper indicates Diagnosing nosocomial infections in critically ill patients admitted to intensive care units (ICUs). It is a challenge because signs and symptoms are usually non-specific for a particular infection. In addition, the choice of treatment, or the decision not to treat, can be difficult. The authors of this paper discussed the historical development, capabilities, and drawbacks of various computer-based decision-support models for infectious diseases, with special emphasis on Bayesian approaches. Although Bayesian decision-support systems are potentially profitable for medical decision making in infectious disease management, clinical experience with them is restricted and prospective estimation is needed to determine whether their use can recover the quality of patient care. The Some of the benefits include clinicians trust that the use of decision-support systems in medicine will improve the quality of patient care through better treatment choices and by accomplishing a better balance between costs (both financial and medical, such as side-effects of drugs) and benefits. The defect of this system is Clinicians are generally uncertain to use computerized instructions that require additional data entry and time and effort.

III. SYSTEM ARCHITECTURE AND DESIGN GOALS

This section provides the architecture of the proposed system along with design goals, Performance analysis and then provides the conclusion.

A. System architecture

The Architectural representation of the proposed system is as present in the Fig 1. It consists of the Processing unit that processes the clinical text. The processing unit is considered to be a hospital or a company. It helps in providing online direct-to-consumer service by providing individual risk prediction for the diseases based on the symptoms of a new patient. The Processing Unit Consists of the historical medical data that helps to train the naïve Bayesian classifier and then finds the disease risk of the undiagnosed patients. The Data Provider provides the historical data of the old patients. It consists of symptoms and diseases of all the patients. This mainly helps to train the Naïve Bayesian Classifier. All these data are outsourced to the cloud platform. The new patients are the one who has not diagnosed yet. These patients will give information about their symptoms. The patients can give the symptoms by their own or they can give the symptoms they have collected during doctor visit. The new patient can give heart rate, Sugar level, blood pressure, weight etc. these symptoms have to sent to the Processing Unit for the diagnosis. The trusted authority is the one that can be trusted by all other entities in the system. It distributes and manages the private keys involved in the process.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 5, May 2017

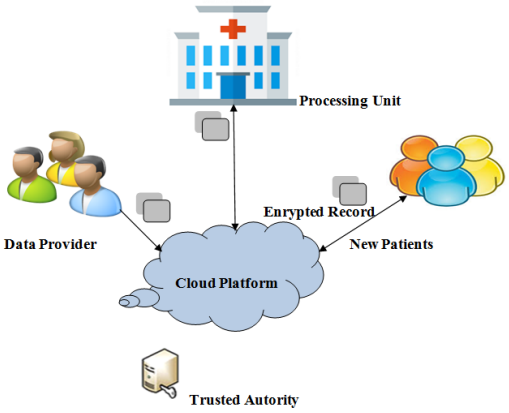

Fig.1 System Architecture

B. Design Goals

In order to accomplish the secured medical decision for new patients under
the preceding model, our system design will achieve privacy and performance guarantees as follows.

1. The Proposed System Should fulfill Privacy-Preserving Requirements:

As mentioned earlier, if CDSS will not consider the privacy requirements, patient's sensitive information (symptom and disease information) will be disclosed to Processing Unit, Cloud Platform, and unauthorized third parties in the patient's medical decision. It will make the patients to unknowingly their own data to the CDSS. Along with that the Processing Unit being a benefit company, that prohibits its own data from leaking to other third parties. For this reason, the proposed system should accomplish the privacy of a Processing Unit and the new patients.

2. The Proposed System Should Accomplish the Computation Efficiency:

The patients always have limited computational assets, which cannot support overturning of the computation. To device patient-centric diagnosis revival from Cloud Platform in time, the proposed system should deal with the computation efficiency. So, it is important to allow New Patients to rescue the diagnosis results in real time.

IV. PROPOSED SYSTEM

Naïve Bayesian Classifier, Paillier Homomorphic encryption and Secure Multiplication (SM) protocol is used as the basis of the proposed Privacy Preserving Top-K Disease Names Retrieval Method for Clinical Decision Support System.

Naïve Bayesian Classifier [9] is a commonly used Machine Learning algorithm which is proved to be very effective in many of the applications including medical diagnosis.

We use Paillier Homomorphic Encryption [8] as one of the building block to implement our method. This algorithm includes steps like key generation, encryption and decryption. Key generation is nothing but generating the secrete key and through that secret key the encryption and decryption is carried out. The Paillier Homomorphic Encryption have properties like Additive Homomorphism, Scalar-Multiplicative Homomorphism and Self-Blinding.
As the Paillier Homomorphic encryption supports only additive homomorphism, and it cannot achieve the multiplication of the plaintext, so we use SM protocol as a building block to design our system. The proposed system consists of 3 main phases.

A. Phase 1

In this phase, Data Provider should provide his historical medical data to the Processing Unit for training naive Bayesian classifier and these data should be sent to the Cloud Platform for inventory purpose. We first compose new cryptographic tool called AHPA (Additive Homomorphic proxy Aggregation) scheme to securely aggregate the message to solve the scam problem between Processing Unit and Cloud Platform. Then, we use the tool to train Naive Bayesian classifier privately.

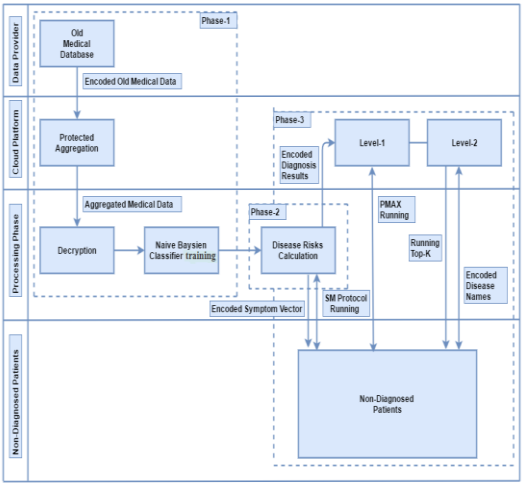

Fig.2 Procedure of the Proposed System

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 5, May 2017

B. Phase 2

In this phase we compute the disease risk of a patient in a privacy preserving way. Suppose a patient has n measured symptoms called X1,….,Xns, he decrypts these symptoms and sends them to the Processing Unit through Cloud Platform. Once these encrypted symptoms are received, Processing Unit cannot decrypt the patient's encrypted symptoms since he has not given with a patient's private key. The Processing Unit can use the Bayes's theorem to calculate patient's probability of having disease. In some situations, Cloud Platform may know the disease names according to the order arranging of the cipher text even without decryption. Once some encrypted disease names are rescued, the privacy about this patient will be leaked to the Cloud Platform. In order to protect patient's privacy, all the cipher texts must be permuted before outsourcing.

C. Phase 3

This phase is mean for the patients to rescue the Top-K diagnosed results in a privacy preserving way. When these diagnosed, results have been computed, Cloud Platform needs to find the encrypted top-k diagnosis results and sends them back to the patent. This procedure can be divided into two steps.

Step1: Cloud Platform needs to judge whether a patient suffers from some specific diseases according to the probabilities.

Step 2: Cloud Platform needs to select top-k possible disease names according to diagnosis probabilities.

In order to accomplish these two steps, the below mentioned methods have to be designed called privacy-preserving maximum out of n protocol (PMAXn), and privacy-preserving top-k disease names retrieval protocol (TOP-K), respectively.

1. Privacy-Preserving Maximum Out of n Protocol: The main objective of PMAXn is to allow Cloud Platform to obtain a new cipher text tuple TU with maximum HU from the n encrypted tuples T1, . . . , Tn . Neither the plaintext information nor cipher text relationship will be leaked to both Cloud Platform and patient during running this protocol.

Pseudo code for Privacy-Preserving Maximum Out of n Protocol

Input: Let Cloud Platform has n tuples T_1,.., T_n and patient has a private key SK.

Output: The maximum tuple T_U among T_1,.., T_n Initialize set S_b as S_b = { T_1,.., T_n } From i=1 to $\log_2$ n , initialize S_a as ∅

if $j \in \frac{Sb}{2}$ then T_j is calculated using the equation T_j =PMAX (T_{2j-1}, T_{2j})

then adding T_j to set S_a

then set S_a= $S_{b,}$ S_b will now have only one entity T_U. Now return T_U.

2. ***Privacy-preserving top-k disease names retrieval protocol (TOP-K)***

This method will run for *k* loops. For each loop, *PMAXn* will be executed to allow Cloud Platform to get the maximum tuple *T*MAX from set *Sa*. Then, for every entity that belongs to *Sa*, we have to choose random numbers such that *Rj* ∈ ZN

Pseudo code for Privacy-preserving top-k disease names retrieval protocol (TOP-K)

Input: Cloud Platform has n_d cipher text T_1, . . . , T_n, (k< n_d) and patient has a private key SK.

Output: Cloud Platform gets Top-K diseases names. Firstly initialize S_a^1 such that S_a^1= { T_1, . . . , T_n} , now calculate P_{IDj} = E_{PKc}(0)

From i=1 to k we have to do, Run T_{MAX}=$PMAX_n$ (T_1, . . . , T_n). After this we get a tuple T_{MAX} with highest probability. Here T_1, . . . , T_n ∈ S_a^1 . From j=1 to n, we have to randomly choose *Rj* ∈ ZN. Now we have to calculate

V_j=($E_{PKc}(H_{MAX})$· $E_{PKc}(H_j)^{N-1}$)^{Rj}. Now we have to permute n encrypted data using πi and it has to be sent to the patient.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 5, May 2017

At Patient: Decrypt πi , using the SK and we can denote as β_j=D_{SKc}(V πi)

If β_j=0 then denote A π_i= E_{PKc}(0) and B π_i= E_{PKc}(1)

Otherwise denote Aπi= E_{PKc}(1) and B π_i= E_{PKc}(0)

Reurn A π_i and B π_i to Cloud Platform.

At Cloud Platform: Obtain A_j and B_j using the permutation π_i. Now refresh P_{IDj} and E_{PKc} using the SM protocol. Now return P_{IDj} to the patient

V. PERFORMANCE ANALYSIS

In this section, we consider the performance of the Proposed System in terms of computation cost and communication overhead.

Disease name	IUB	NRPO
YES	59/59(100%)	50/50(100%)
NO	45/61(73.77%)	60/70(85.71%)
Overall	104/120(86.67%)	110/120(91.67%)

Table 1. The efficiency of the Proposed System over Real Dataset AID

A. Computation Cost: We calculate the computation cost of Proposed System by using a custom simulator built in Java. This experiment was run on a test machine with one 2.5-GHz two-core processor and 6-GB RAM. In the experiment, we consider two datasets. One real dataset is used from the UCI machine learning repository called Acute Inflammations Dataset AID[4]. We use this dataset to test the performance of the Naive Bayesian classifier by using Proposed System. We also use synthetic dataset to test all factors which affect the performance of the Proposed System.

B. Communication Overhead: In this section, we describe the communication cost in our Proposed System. The privacy parameter of Paillier encryption system used is 2048 bits. In order to train Naive Bayesian classifier, all historical medical data should be encrypted and sent to Cloud Platform which costs O(l • (ns + nd)) to transmit. Then, Cloud Platform wants to aggregate all the historical medical data into one vector. This aggregated vector costs O(ns + nd) to transmit to Processing Unit. So the total communication cost of Phase 1 is O(l • (ns + nd)). After computation, the encrypted diagnosed probability and encrypted disease names should be sent to Cloud Platform to store which costs O(nd). So, the total communication cost of Phase 2 is O(ns + nd) and it also costs Cloud Platform (ns + 2nd) • 2048 bits to store Patient's cipher text. In Phase 3, it first costs O(nd) to calculate PMAXn (log2 nd rounds). After that, it takes O(k • (nd)2) (k(log2nd + 1) rounds) to achieve TOP-K in Step 2. Therefore, the total communication overhead is O(k • (nd)2) (k • log2nd + k + 1rounds) in Phase 3.

VI. CONCLUSION

In this paper, we have proposed a "Privacy Preserving Top-K Disease Names Retrieval Method for Clinical Decision Support System". By using the advantages of Cloud Computing techniques new patients can make use of big medical dataset stored in Cloud Platform to train Naive Bayesian Classifier, and then apply the classifier for diagnosis without compromising the privacy of Data Provider. In addition, the patient can securely retrieve the top-k diagnosis results according to their own priority in our system. Since all the data are processed in the encrypted form, our system can achieve patient-centric diagnose result retrieval in privacy preserving way.

REFERENCES

[1] A. Shiryaev, "Bayes formula," in Encyclopedia of Mathematics. New York, NY, USA: Springer, 2011.

[2] Y. Lindell and B. Pinkas, "A proof of security of YAOS protocol for two-party computation," J. Cryptol., vol. 22, no. 2, pp. 161–188, 2009.

[3] A. Shiryaev, "Bayes formula," in Encyclopedia of Mathematics. New York, NY, USA: Springer, 2011.

[4] Acute inflammations data set, UCI machine learning repository.(2009).

[5] R. S. Ledley and L. B. Lusted, "Reasoning foundations of medical diagnosis,"Science, vol. 130, no. 3366, pp. 9–21, 1959.

[6] Y. Elmehdwi, B. K. Samanthula, and W. Jiang, "k-nearest neighbor classification over semantically secure encrypted relational data,"

[7] C. Schurink, P. Lucas, I. Hoepelman, and M. Bonten, "Computer-assisted decision support for the diagnosis and

treatment of infectious diseases in intensive care units," Lancet Infectious Dis., vol. 5, no. 5, pp. 305–312, 2005.

[8] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in Proc. Adv. Cryptol. Int. Conf. Theory Appl. Cryptograp. Techn., Prague, Czech Republic, May 2–6, 1999, pp. 223–238.

[9] R. Lu, H. Zhu, X. Liu, J. K. Liu, and J. Shao, "Toward efficient and privacy-preserving computing in big data era," IEEE Netw., vol. 28, no. 4, pp. 46–50, Jul./Aug. 2014.