ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

Privacy Preserving Health Data Mining

[1]Somy.M.S, [2]Gayatri.K.S, [3]Ashwini.B
[1,2,3]Dept. of CSE, Mar Baselios College of Engineering & Technology, Kerala, India

Abstract

Data mining is exploring of large quantities of data and analyses it into understandable patterns. In healthcare, many factors have motivated the use of data mining applications. Health data mining is a process of extracting previously unknown information from a large volume of health data. Main aim of health mining is to improve patients care. Health data mining has application such as fraud detection and abuse, decision management in customer relationship, best treatments and practices, better and modest health care services. Releasing patient specific medical reports may potentially reveal sensitive information of individual patients. Many previous privacy models and algorithms have been proposed but they cannot preserve the structure of mined data while, anonymizing. Here presents a new system to solve the problem of publishing medical report and proposed a solution to anonymize collection of medical reports. This method preserves the quality of medical reports especially for the purpose of cluster analysis. It also suggests a method for the extraction of information from the raw medical data.

Keywords

Health Data Mining, LKC, Privacy, Anonymity

I. Introduction

Dataminig is an analytic process mainly designed for exploring data. It is analyzing the data and summarizing it into useful information. It is extracting unknown information from a large database. The wide range of applications of Datamining include business, marketing, healthcare, scientific field etc.

One of the important applications of Data mining is in the field of medicine. Health mining is extracting previously unknown and hidden data and from large medical database. It is increasingly popular nowadays. Huge amount of medical data is very complex and its processing is also very difficult by ordinary method. Health Data mining provides efficient techniques and methodology for the processing of medical data. It helps to develop better diagnosis and treatment. Health mining application includes treatment effectiveness, health care management, drug discovery, electronic health records, customer relationship management, fraud and abuse [2] with medical data.

Even though Health Datamining has this much of applications, it affects individuals' privacy. Many organizations publish these medical data. The outside vendors and the insurance person can sell this data for various companies as commodity. Also the health data can be sold by the person who can access the cloud where this data is stored [1]. This may affect individuals privacy. Therefore preserving privacy is important.
To publish the medical data without affecting the privacy we need to anonymize the medical data. We have to anonymize the data before publishing. The fig. 1 shows the basic model for privacy.

Dataset
Data Anonymization
Data Publishing

Fig. 1: Privacy Model

Main aim of this paper is to publish the medical data for various purposes without affecting individual privacy. In this paper we suggest a method for anonymizing individual's information for the purpose of Datamining. Here we suggest a new method for information extraction from the electronic medical data such as Electronic Health data (EHR), Electronic Medical Record (EMR). The cluster application of data mining is focused in this paper. The rest of this paper is composed as. In section II we review related works, section III gives problem statement, section IV gives the overview of the proposed system lastly section V result and discussion took after by conclusion in section VI.

II. Related Works

Releasing of medical data without influencing individual privacy, we need to anonymize the document .Here in our work we concentrate on the clustering of data. Therefore the major areas which related to our work are information extraction, clustering and privacy.

There are various method for extracting information from medical reports .one method is by using the machine learning methods [5, 7]. They need large data set and also the training set should be very large to perform the information extraction using the machine learning method [1]. Another method for the information extraction is based on the lexical method and pattern matching [10, 12,]. But it also needed many domain specific thesaurus for information extraction [1].

The privacy techniques related to our work are discussed as follow; Sweeney [3] proposed a new definition of privacy called k anonymity. It is an anonymity model which provides a framework for algorithm and systems which release information in such a way that the information after anonymization should be restricted. It protected the data which refers to the particular entity. It mainly avoids the linkage attack. Anonymity is achieved by using suppression and generalization. Generalization is replace the original value with a less specific vale, for ex. age 38 can be generalized as 30-40.Suppression is suppressing or hiding the corresponding value using *. By generalizing and suppressing of the individual's information which needed to publish the joining attacks is prevented. The main drawback of this method is the degradation of information quality for high dimensional data and lack of protection against attribute disclosure.

Machanavajjhala [4] proposed an extension of the k-anonymity model called l-diversity which is a group based anonymization which preserves privacy in data sets by reducing the granularity of a data representation. It does not require knowledge of sensitive and non sensitive attribute .The l-diversity model overcomes some of the disadvantages in the k-anonymity model. The protected identities of the k-level individuals are not equivalent to protecting the corresponding homogenous sensitive values that were generalized or suppressed. The l-diversity model helps for intra-group diversity.The larger value of l protects more knowledge is needed for sensitive attribute. It has an instance level knowledge. A main disadvantage of this method is it cannot be used for multiple sensitive attribute and cannot efficiently use the preserved data. The concept of utility is less achieved. The data quality is degraded when the data is high dimensional.

Li [6] proposed another model for privacy called t-closeness. It is an improvement of k anonymity and l diversity. It mainly distributes the sensitive attribute in the equivalence class and this is close to distribution of attribute in the entire table [6]. Information gain in this method is limited only to individual specific information. It uses a method called Earth Mover Distance which considers the semantic closeness of attribute value. The Earth Mover distance metric compute the distance between the two distributions. t-closeness approach is more useful in the case of numeric attributes. It reveals that not all values of an attribute are equally sensitive. It focused on the advantage of anonymization instead of generalization. The t-closeness is calculated such that if the distance between the distributions of a sensitive attribute in this class and the distance in the whole table is no more than a threshold t. It provides more protection to sensitive attributes. But the EMD measure is not perfect and also the information gain is unclear. N Mohammed [8] introduces a new technique which is different from the above methods called LKC privacy. The usual approach to ensure privacy is to generalize the quasi identifier and also removing the explicit identifiers. In the case of high dimensional data, the data suppression affects the quality. To overcome this problem a new technique is developed called LKC Privacy technique [3]. It exploits the limited prior knowledge of the attacker: in real-life, it is difficult for an attacker to acquire all the information in QID of a target victim. Thus, we can assume that the adversary's anterior knowledge is bounded by at most L values of the QID attributes of the victim. This LKC privacy model [1,8] guarantees that the probability of successfully identifying a medical document of a target patient based on qid values is at most 1/K, and the probability of successfully identifying a sensitive value of a target patient from a medical document is at most C. L, K, and C, are user-specified thresholds.

Divya [13] suggest various clustering techniques which can be used for health care application, which includes the k means algorithm, c means, hierarchical clustering algorithms, and density based clustering techniques.

III. Problem Statement

We need to extract the information from the medical reports like Electronic Health Record (EHR), Electronic Medical Record (EMR). In order to publish this data we have to anonymize the data according to the user needs. Here we are considering the clustering of medical data. So we need to perform the clustering in the anonymized.

Suppose a hospital want to publish the medical data to the public without affecting the person specific information. The data can be given to the third party without affecting the privacy. Publishing of person specific data may cause threat to individual privacy. There are many attacks which may cause for the identification of individual data from the large database [1].

Suppose an attacker aims to identify the patient's specific medical document, the intruder is able to identify patient using two possible attacks [1, 8].

A. Identity Linkage

This occurs when the numbers of records are small; the attacker may able to identify the target patient's document from group of document. For example attacker knows the details as<Male, Lawyer, 56> from the below table it can be easily identifiable.

Table 1: Linkage

Id	Age	Gender	Job	Disease
1	56	Male	Lawyer	Cancer
2	30	Female	Teacher	Fever
3	45	Male	Lawyer	HIV
4	37	Female	Mechanic	accident

B. Attribute Linkage

This takes place when the attacker can infer the value of sensitive attribute with a higher confidence. For example attacker knows the data such as<male, lawyer> from the table 3.2 he can identify the person's disease with a very high confidence.

Table 2: Attribute Linkage

Id	Age	Gender	Job	Disease
1	56	Male	Lawyer	Cancer
2	30	Female	Teacher	Fever
3	45	Male	Lawyer	HIV
4	37	Female	Mechanic	accident

To prevent the linkage attack every combination of the quasi identifier shared with other documents. This can be achieved by the LKC privacy technique.

IV. Proposed Solution

In the proposed system we are anonymizing the medical data without affecting the quality of medical data. The data can be used for the data mining purpose; here we are considering the clustering of data. The data is anonymized using LKC privacy technique. The concept of LKC-privacy [1, 6] is that every combination of QIDj (where j=1, 2…..n) in maximum length L in the data table T is shared by at least K records, and the confidence of inferring any sensitive values in S is not greater than C, where L, K, C are thresholds and S is a set of sensitive values specified by the data holder [9, 17]. Here L is the knowledge of attacker, K is the number of records in each QID and C is the confidence bound [1]. In LKC privacy the probability of identity linkage to be ≤1/K and the probability attribute linkage to be < C, provided that the attackers knowledge does not excel L [1]. Here we are using this LKC privacy technique for the purpose of privacy.

The main stages of the proposed system is shown in the Fig. 2.

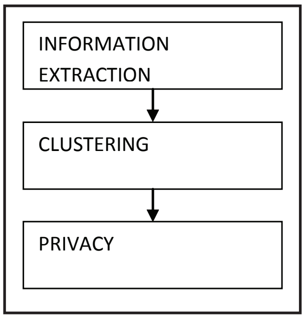


Fig. 2:

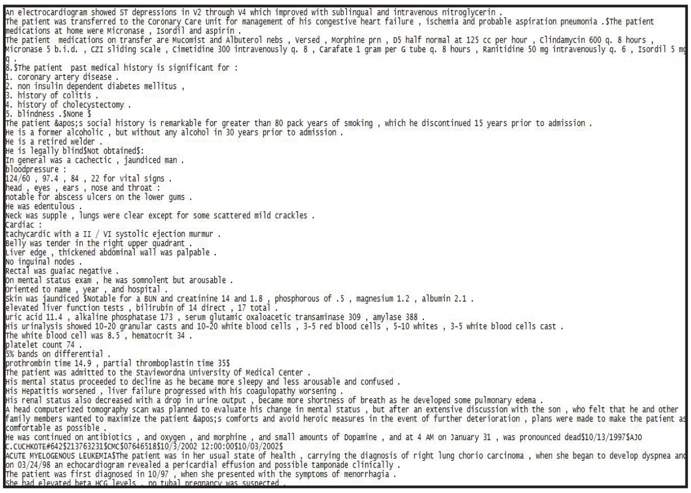
A. Information Extraction

The medical data which is obtained from the hospitals are incomplete. The real data may contain lack of attribute value, noisy data, some error data, lack of essential data and inconsistent data. This may affect the data mining results. To advance the quality of the mining task the data should be preprocessed. It will improve the data efficiency and also help to get better result. This will make the mining task ease. Medical records data are in the form of text document or unstructured xml format. We need to extract the information from the medical report.

Information extraction is extracting the information which is essential. Medical reports contain patient's personal information which is in an unstructured format. It includes patient's personal information such as names, phone numbers, places, age and also details of diseases [3]. These person specific data can be called identifiers. They are classified as Explicit Identifiers, Quasi Identifiers and Sensitive information [1]. Explicit identifiers refer to any term that would lead to identify a patient, such as name, patient IDs etc. Sensitive information includes any information that a patient may not want to disclose, such as diseases and medication taken. Quasi identifiers [3] are any information related to person other than explicit identifiers such as age, phone number, job, location etc.

In the case of Electronic Health Record, Electronic Medical Report these data can be used across different hospitals and also these data can be published to the outside world. So we need to infer each and every detail from the medical data. The information can be extracted in different ways such as pattern matching, lexicons and machine learning techniques which we discussed earlier [5, 7]. Semantic lexicon and extraction rules can also be used for the information extraction [10, 12].

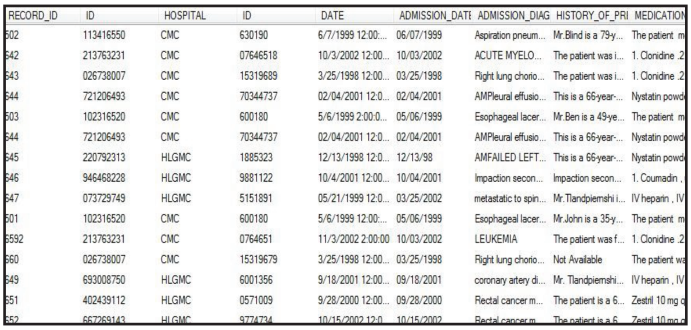
In our system we consider each and every part of the data. By using this we can easily identify the data which are under the same category. We converted the unorganized data into the useful data. The original data are in not in a structured format this can be seen in the fig. 3.



```
An electrocardiogram showed ST depressions in V2 through V4 which improved with sublingual and intravenous nitroglycerin .
The patient was transferred to the Coronary Care Unit for management of his congestive heart failure , ischemia and probable aspiration pneumonia .$The patient
medications at home were Micronase , Isordil and aspirin .
The patient  medications on transfer are Mucomist and Albuterol nebs , Versed , Morphine prn , D5 half normal at 125 cc per hour , Clindamycin 600 q. 8 hours ,
Micronase 5 b.i.d. , CZI sliding scale , Cimetidine 300 intravenously q. 8 , Carafate 1 gram per G tube q. 8 hours , Ranitidine 50 mg intravenously q. 6 , Isordil 5 mg
q .
8.$The patient  past medical history is significant for :
1. coronary artery disease .
2. non insulin dependent diabetes mellitus ,
3. history of colitis .
4. history of cholecystectomy .
5. blindness .$None $
The patient &apos;s social history is remarkable for greater than 80 pack years of smoking , which he discontinued 15 years prior to admission .
He is a former alcoholic , but without any alcohol in 30 years prior to admission .
He is a retired welder .
He is legally blind$Not obtained$:
In general was a cachectic , jaundiced man .
bloodpressure :
124/60 , 97.4 , 84 , 22 for vital signs .
head , eyes , ears , nose and throat :
notable for abscess ulcers on the lower gums .
He was edentulous .
Neck was supple , lungs were clear except for some scattered mild crackles .
Cardiac :
tachycardic with a II / VI systolic ejection murmur .
Belly was tender in the right upper quadrant .
Liver edge , thickened abdominal wall was palpable .
No inguinal nodes .
Rectal was guaiac negative .
On mental status exam , he was somnolent but arousable .
Oriented to name , year , and hospital .
Skin was jaundiced $Notable for a BUN and creatinine 14 and 1.8 , phosphorous of .5 , magnesium 1.2 , albumin 2.1 .
elevated liver function tests , bilirubin of 14 direct , 17 total .
uric acid 11.4 , alkaline phosphatase 173 , serum glutamic oxaloacetic transaminase 309 , amylase 388 .
His urinalysis showed 10-20 granular casts and 10-20 white blood cells , 3-5 red blood cells , 5-10 whites , 3-5 white blood cells cast .
The white blood cell was 8.5 , hematocrit 34 .
platelet count 74 .
5% bands on differential .
prothrombin time 14.9 , partial thromboplastin time 35$
The patient was admitted to the Staviewordna University Of Medical Center .
His mental status proceeded to decline as he became more sleepy and less arousable and confused .
His Hepatitis worsened , liver failure progressed with his coagulopathy worsening .
His renal status also decreased with a drop in urine output , became more shortness of breath as he developed some pulmonary edema .
A head computerized tomography scan was planned to evaluate his change in mental status , but after an extensive discussion with the son , who felt that he and other
family members wanted to maximize the patient &apos;s comforts and avoid heroic measures in the event of further deterioration , plans were made to make the patient as
comfortable as possible .
He was continued on antibiotics , and oxygen , and morphine , and small amounts of Dopamine , and at 4 AM on January 31 , was pronounced dead$10/13/1997$AJO
C.CUCHKOTE#642$213763231$CMC$07646518$10/3/2002 12:00:00$10/03/2002$
ACUTE MYELOGENOUS LEUKEMIA$The patient was in her usual state of health , carrying the diagnosis of right lung chorio carcinoma , when she began to develop dyspnea and
on 03/24/98 an echocardiogram revealed a pericardial effusion and possible tamponade clinically .
The patient was first diagnosed in 10/97 , when she presented with the symptoms of menorrhagia .
She had elevated beta HCG levels , no tubal pregnancy was suspected .
```

Fig. 3: Raw Data

Here we extracted the information from the raw data by the duplication of the header and also consider the repeated data. Thus we get each and every data which is useful. In order to convert the unstructured data into structured data we used the tagging so that they can be easily extracted from the whole data and can be used for the further processing. The fig 4 shows the organized data that we get from the unstructured one.



RECORD_ID	ID	HOSPITAL	ID	DATE	ADMISSION_DATE	ADMISSION_DIAG	HISTORY_OF_PRI	MEDICATION
502	113416550	CMC	630190	6/7/1999 12:00:...	06/07/1999	Aspiration pneum...	Mr.Blind is a 79-y...	The patient m
642	213763231	CMC	07646518	10/3/2002 12:00...	10/03/2002	ACUTE MYELO...	The patient was i...	1. Clonidine .2
643	026738007	CMC	15319689	3/25/1998 12:00...	03/25/1998	Right lung chorio...	The patient was i...	1. Clonidine .2
644	721206493	CMC	70344737	02/04/2001 12:0...	02/04/2001	AMPleural effusio...	This is a 66-year-...	Nystatin powd
503	102316520	CMC	600180	5/6/1999 2:00:0...	05/06/1999	Esophageal lacer...	Mr.Ben is a 49-ye...	The patient m
644	721206493	CMC	70344737	02/04/2001 12:0...	02/04/2001	AMPleural effusio...	This is a 66-year-...	Nystatin powd
645	220792313	HLGMC	1885323	12/13/1998 12:0...	12/13/98	AMFAILED LEFT...	This is a 66-year-...	Nystatin powd
646	946468228	HLGMC	9881122	10/4/2001 12:00...	10/04/2001	Impaction secon...	Impaction secon...	1. Coumadin ,
647	073729749	HLGMC	5151891	05/21/1999 12:0...	03/25/2002	metastatic to spin...	Mr.Tlandpiemshi i...	IV heparin , IV
501	102316520	CMC	600180	5/6/1999 12:00:...	05/06/1999	Esophageal lacer...	Mr.John is a 35-y...	The patient m
6592	213763231	CMC	0764651	11/3/2002 2:00:00	10/03/2002	LEUKEMIA	The patient was f...	1. Clonidine .2
660	026738007	CMC	15319679	3/25/1998 12:00...	03/25/1998	Right lung chorio...	Not Available	The patient wa
649	693008750	HLGMC	6001356	9/18/2001 12:00...	09/18/2001	coronary artery di...	Mr. Tlandpiemshi...	IV heparin , IV
651	402439112	HLGMC	0571009	9/28/2000 12:00...	09/28/2000	Rectal cancer m...	The patient is a 6...	Zestril 10 mg q
652	667269143	HLGMC	9774734	10/15/2002 12:0	10/15/2002	Rectal cancer m	The patient is a 6	Zestril 10 mg q

Fig. 4: Information Extraction

B. Clustering

Clustering is the grouping of object which is in the same class. The clustering is mainly based on the similarity. It is an unsupervised learning method. Here we consider that the custodian want to perform the cluster analysis. The cluster analysis has wide application in the field of medicine. The data can be used to predict the likelihood diseases, easily findout the similar symptoms. Doctors can be easily indentified a disease or medication for the patient on the basis of clustered result. It can be used for the medical decision making [1, 16].

In this research we are using the clustering techniques such as kmeans, c-means, agglomerative clustering, single linkage. We are applying the clustering method before the privacy and also after giving the privacy. We describe the uses of clustering algorithm in medical data also this can be used for the analysis of the quality of information after giving the privacy. Thus we can analyze the efficiency of the privacy technique.

C. Privacy

The privacy technique which we used here is LKC Privacy. The LKC privacy technique [8] can be used for the high dimensional data. It does not affect the linkage attacks[1].It is mainly based on the assumption that in the real world an intruder do not have a complete knowledge about the attributes. This is exploited in this technique. The L is the knowledge of the attacker, K is the no of attributes in the record and c is the confidence bound of the attacker. LKC privacy is mainly based on certain taxonomy.

In this research we are giving certain taxonomy in the record. Here we consider the data can be used inside and outside the hospitals so the user can identify the area which he need to anonymize. The data custodian (Hospital) can decide the attribute s which he does not want to publish [1]. He can preserve the person specific information which affects the patient privacy. The person specific information which is should be anonymized is given by HIPAA [14] so without violating is terms the data is anonymized based on the requirement.

V. Result and Discussion

The dataset used for the study purpose is the publically accessible dataset in i2b2 National Center for Biomedical Computing. The information is extracted from the data set. The dataset is first preprocessed and then it is converted to the text format in order to perform the information extraction. The extracted information is in a structured format. After the data is extracted the clustering are applied in the selected fields in the extracted information.

Clustering technique used here are the kmeans, Cmeans, hierarchical agglomerative clustering, single linkage clustering. The best clusters are formed in the Hierarchical Clustering. The analyses of the different cluster formation are shown in the fig. 5. For analysis we use the silhouette index. The below fig. 5 shows the result of cluster analysis. The best clusters are obtained from the Hierarchical agglomerative clustering.

The cluster result can be used for the analysis of the privacy quality. The data quality or information quality always reduced after the privacy is given to the data.

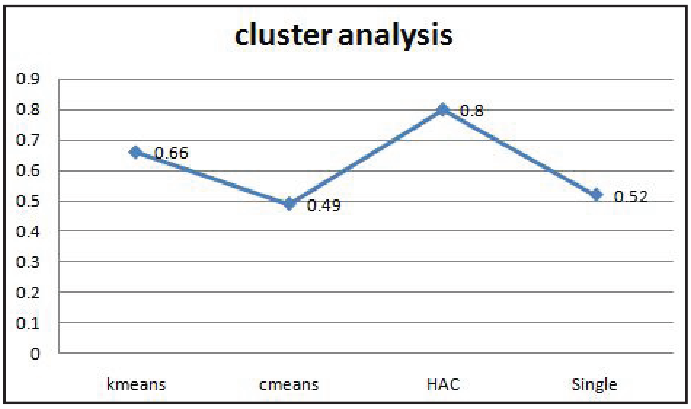


Fig. 5: Clustering Algorithms

The data quality is reduced because of the original data is replaced by the anonymized value. In this system obtained a good information quality even after the anonymization. The information utility is calculated on the basis of clustering results. The Fig. 6 shows the analysis of the privacy technique.

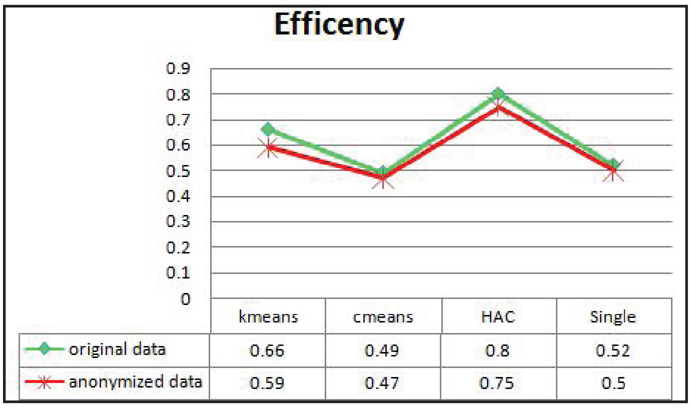


Fig. 6: Privacy

From the above fig we can see the efficiency of the privacy technique. The above graph is plotted on the result of clustering analysis of original data analysis and the anonymized data. From this it is clear that the clustering is almost similar.

VI. Conclusion

A new method for preserving privacy in health data mining is proposed. The proposed system has mainly three stages- Information extraction, Clustering and Privacy. Our system extracts the information from medical data. The data is completely extracted and the required data is obtained. The extracted data is clustered using different clustering algorithms. The data anonymized in order to ensure privacy. The clustering algorithms K-Means, Cmeans, Hierarchical clustering, and single linkage are used here. The anonymization is given using the LKC privacy. The anonymization process can be done according to custodian's requirement. The challenge of this work is, only textual medical data can be extracted using this technique. We proposed a technique for the anonymization of medical data and the anonymization purpose can be done according to data requirement.

Reference

[1] A. Hood, C. M. Fung, Farmhand Irbil,"Privacy-Preserving Medical Reports publishing for Cluster Analysis", International Conference on New Technologies, Mobility and Security (NTMS), 2014, pp. 1-8

[2] H.C.Koh, G.Tan,"Data mining applications in health care", Journal of Healthcare Information Management Vol. 19, No. 2, pp. 64-73.

[3] P. Samurai, L. Sweeney,"Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", Technical Report, SRI International, 1998.

[4] A.Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam,"l diversity: Privacy beyond k-anonymity," ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 1, pp. 3, 2007

[5] E. Aramaki, T. Imai, K. Miyo, K. Ohe,"Automatic deidentification by using sentence features and label consistency," in i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006, pp. 10-11.

[6] Ninghui Li, Tiancheng Li,"t-Closeness: Privacy Beyond k-Anonymity and l--Diversity", Proceedings of the 32nd international conference on Very large data bases, pp. 139–150. VLDB Endowment, 2006.

[7] G. Szarvas, R. Farkas, R. Busa-Fekete,"State-of-the-art anonymization of medical records using an iterative machine learning framework", J. Am. Med. Inform. Assoc., Vol. 14, pp. 574-580, Sep-Oct, 2007.

[8] N. Mohammed, B. C. M. Fung, P. C. K. Hung, C. Lee, "Anonymizing healthcare data: A case study on the blood transfusion service", Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 1285-1294.

[9] N.Mohammed, B.C.M. Fung, Hung, "Privacy Preserving Data Mashup", Proceedings of Extending Database Technology (EBDT), March 2009, pp. 24-46.

[10] E. Fielstein, S. Brown, T. Speroff,"Algorithmic de-identification of medical exam text for HIPAA privacy compliance: preliminary findings", Med Info, Vol. 1590, 2004.

[11] Deléger, Louise, Cyril Grouin, Pierre Zweigenbaum, "Extracting Medical Information from Narrative Patient Records: The Case of Medication-Related information.", Journal of the American Medical Informatics Association : JAMIA17.5 (2010): 555–558. PMC.

[12] F. P. Morrison, L. Li, A. M. Lai, G. Hripcsak, "Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes?" J. Am. Med. Inform. Assoc., vol. 16, pp. 37-39

[13] Divya Tomar, Sonali Agarwal," A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology Vol. 5, No. 5 (2013), pp. 241-266

[14] U.S. Department of Health and Human Services Office for Civil Rights, "HIPAA Administrative Simplification Regulation Text", March 2013.

[15] Sunita Sarawagi,"Information Extraction", Databases Vol. 1, No. 3 (2007) pp. 261–377, 2008

[16] David Dilts, Oseph Khamalah, Ann Plotkin,"Using Cluster Analysis for Medical Resource Decision Making", Cluster Analysis for Resource Decisions, Vol. 15, No. 4, Oct-Dec 1995.

[17] B.C.M.Fung, K.Wang,"Attack Models and Privacy Models", Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2010.

Somy.M.S received the B.Tech degree in Information Technology from Mar Baselios college of Engineering and Technology in 2012. She is currently doing M.Tech degree in Computer Science and Engineering at Mar Baselios college of Engineering and Technology. Her current research interests include Datamining, Bigdata.