# Privacy Preserving Data Mining in Electronic Health Record using K-anonymity and Decision Tree

Anvita Srivastava*
CSE Department, RGUKT RK Valley
Cuddapah, India
e-mail: anvita.iti@gmail.com

Gaurav Srivastava
CSE Department, RGUKT, RK Valley
Cuddapah, India
e-mail: gauravsrignp@gmail.com

**Abstract—In this paper, we present an accurate and efficient privacy preserving data mining technique in Electronic Health Record (EHR) by using k –anonymity and decision tree C4.5 that is useful to generate pattern for medical research or any clinical trials. It is analyzed that anonymization offers better privacy rather than other privacy preserving method like that randomization, cryptography, perturbation and encryption methods. K- Anonymity is useful for prevention of identity of patient in medical research; it is useful to avoid linking attack using suppression and generalization process. The objective of this research work is to propose and implement privacy preserving data mining approach, which best suits with EHR systems without impeding the flow of control. The data stored in EHR system is highly confidential which contains information about patient disease. To ensure confidentiality, we are using anonymization of identity revealing attribute before publishing it for other utility purpose. The experimental results show the validity of the proposed approach. Our approach is useful to preserve utility and privacy in healthcare datasets.**

**Keywords -** Privacy Preserving; K-anonymity; Decision Tree.

## I. INTRODUCTION

In past several years, it is being observed that emergence of data mining technique is showing conflict with to privacy assurance. The easily accessibility of electronic form of data may cause privacy threat. Hence, it is required to address privacy issue of sensitive data. Recently, various techniques have been proposed by different researcher for data transformation, which assures privacy constraint and also preserves the utility of data. The anonymization with decision tree approach is one of the widely used approaches [1-9]. Now-a-days, anonymization approach has emanated as an important technique to satisfy privacy constraint when releasing confidential data. This interest in anonymization techniques has resulted in a plenty of methods for anonymizing data under different utility and privacy assumptions. The research work in the field of data utility for anonymized data is done to very less extent. It requires proposing techniques which would be more effective and increase utilization of anonymized data. Privacy is recognized as an essential requirement for Electronic Health Record (EHR) systems [2]. EHRs are very effective for clinical investigation and for healthcare research.

Electronic health record (EHR) contains highly confidential data and its confidentiality must be maintained. The data in EHR should not disclose the identity of the patient. In order to satisfy privacy constraint, anonymization of health record must be done and then it should be published for other utility purpose. Recently, k-anonymity is a very popular and effective approach to preserve privacy of the patient identity and also prevent the data from linking attack. Decision tree classification is useful to classify pattern [10-13]. In this research work, we have proposed Privacy Preserving Data Mining (PPDM) approach for medical research "k-anonymity with decision tree". The main goal of this research is to provide tradeoffs between privacy and utility.

The paper is organized into nine sections. Section 2 and 3 provides an overview of Electronic Health Record and Data Mining. Section 4 deals with the privacy preserving technique in data mining and EHR. Sections 5 and 6 provide brief overview of K-anonymity and Decision Tree approaches. Section 7 discusses the proposed methodology and experimental results are discussed in section 8. Concluding remarks are given in section 9.

## II. ELECTRONIC HEALTH RECORD

In the continuously developing world, EHR plays a very important role to improve global healthcare services. Presently existing technologies provides support of invention in health care systems and identify a technique to combined different technologies and also the need of technology. Different type of standardized electronic health records are remote healthcare, aggregate public health data, genomic medicine, and telemedicine. In short EHR

must be secure, robust and it must contain real time availability and patient based information [14-16]. An EHR system should include following features:

- Electronic health information contains detail about patient in a longitudinal manner, where health information means detail of health record of an individual patient.
- Fast and real time access to individual and globally available data by authenticated users only.
- It must contain features such as repository of valuable information, which could be useful in decision making and also improves the quality and efficiency of healthcare systems.
- It should provide adequate process for betterment of healthcare sector.

The architecture of Electronic Health Record is shown in figure 1.

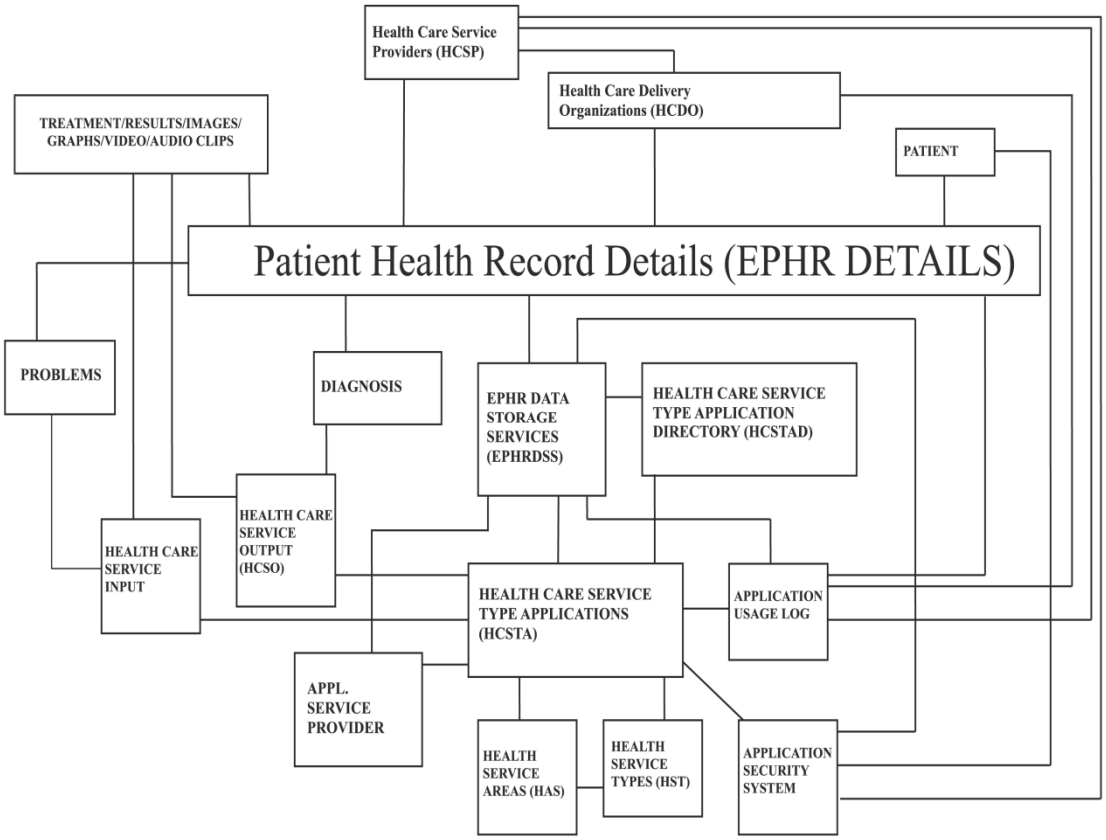



Figure 1. Architecture of Electronic Health Record

## III. DATA MINING

Knowledge discovery process, or data mining, is the technique of analysis huge amounts of data to find out the hidden patterns and also helpful in predictive information in an automated sequence [6,10, 16-18]. This process is useful in organization to make decision better. Data mining process uses modelling techniques, database technology, machine learning, and statistical analysis to:

- Discover hidden pattern and to make prediction easier.
- On actual historical data produce predictive models.
- Can make data analysis more accessible to end users.
- Directly handles large databases directly.
- Strong focus on decisions and their implementation.
- Results can be easier to interpret than e.g. regression models.
- Semi-automation of analysis

In other words we can say that data mining is the process which reveals hidden pattern in vast data.Data mining process or knowledge discovery process follows step by step process which is given below [6]:-

Step 1:- Firstly remove noise and inconsistent data from the information repository. These processes known as data cleaning process.

Step 2:- In second step multiple data sources may be merged, known as data integration process.

Step 3:- In this step relevant data are retrieved from the database and perform analysis task. These processes are known as data selection process.

Step 4:- In this step data are transformed into appropriate way for mining by using aggregation operation, known as data transformation process.

Step 5:- Intelligent technique apply to extract useful data pattern, known as data mining process. This is a very essential and important process.

Step 6:- Identify appropriate and true interesting pattern which represent knowledge based on interestingness measurement, known as pattern evaluation process.

Step 7:- In this final step knowledge representation and visualization methods are applied to show the mined information knowledge to the end user, known as knowledge representation step.

## IV. PRIVACY PRESERVING DATA MINING

PPDM consists of those techniques and methodologies of data mining, which would be used to satisfy privacy constraint and other side it also maintains the utilization of data for data mining. This sector comprises of those researches in which it is being studied how to find hidden patterns and information in large data sets, while maintaining the privacy constraint associated with data. PPDM technique is solely based on description of privacy. Description of privacy defines the different attributes of data. It depicts which attribute is sensitive and hence required to ensure confidentiality constraint [1, 3, 18-25]. The block diagram of PPDM is shown in figure 2.

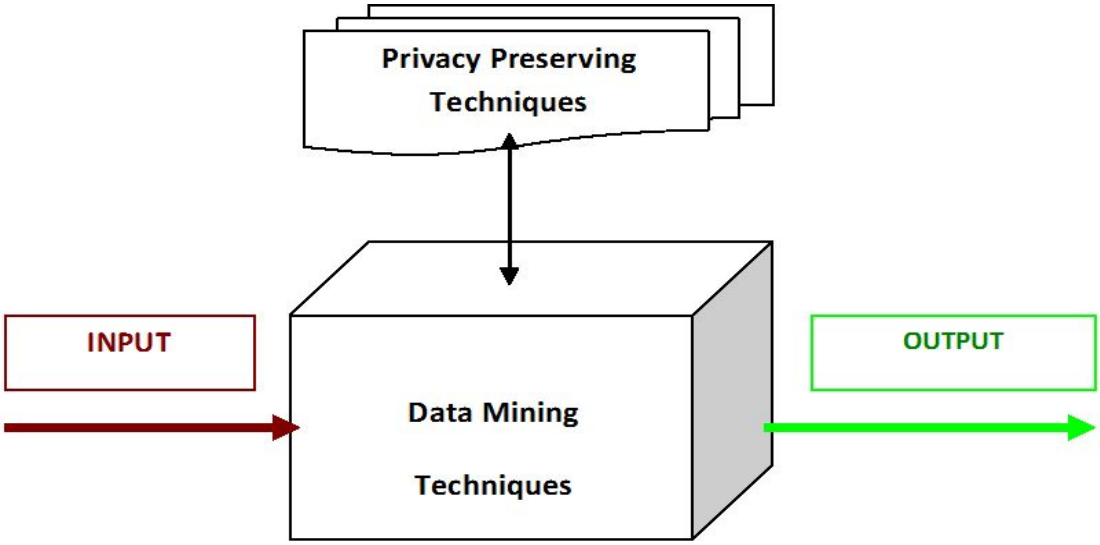



Figure 2: Block diagram of PPDM technique

### *A. Privacy Preserving Data Mjining In Electronic Health Record*

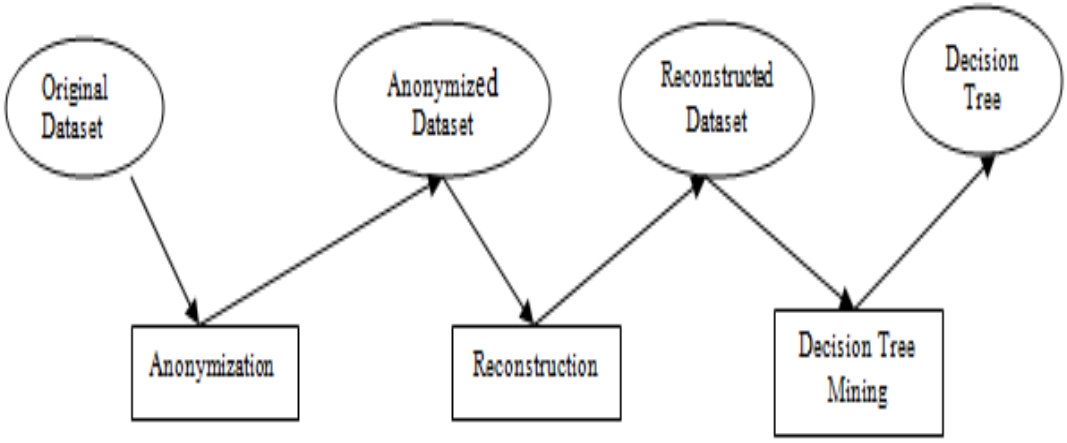
A PPDM technique is highly useful in healthcare systems. It facilitates healthcare organizations to extract valuable information from large data without disclosing the identity of any patient. Thus it provides ability to medical research systems to make analysis of large data sets to gain knowledge and to find out curative measures of deadly diseases without breaching privacy of patients. PPDM uses various techniques to achieve privacy constraints. One such technique involves perturbation approach in which we use a function which perturbs the confidential information regarding any patient's record. This technique allows perturbation to that extent only from which user can create data set which resembles to original data set to certain level. Hence it allows data mining technique to use perturb data sets and reveal hidden information without disclosing identity of anyone. PPDM also have several important algorithms which is being used to build various classification models for predictive measures and for finding association rules among various attributes of data set with a significant accuracy [3].

EHR provides various features such as "Home health monitoring system" in which a patient could use the services of monitoring system to measure and analyze his disease statistics from his own place every day. This system use internet to provide connectivity between user's personal computer and hospital system and fed the health data from hospital to user system and vice-versa. The data is in form of various disease attributes such as patient's blood pressure, diabetic measures such as glucose level and other measures of chronic diseases. This system fed health data on the web application which is running on patient's system and transfer to the hospital's system. The transferred data is used by healthcare attendant or Doctor for monitoring and diagnosis of health condition of patient. Researcher in healthcare sector requires these patient's health data to find out new discovery in preventive and curative of deadly diseases. This individual patient's health data could be integrated and transformed in large data sets which could be used to perform various researches in medical field and also for data mining purpose.

The integrated data which is collectively made from individual information could be used for sharing among various research institutes and hospitals. But the problem in data sharing is that it would result in privacy violation of patients. So it needs to require ensuring privacy constraints before sharing it to others. Hospital managements sometime want to share their health records with researchers who are working independently, but with privacy constraint. To solve this problem Hospital management could use PPDM. In PPDM system, hospital received patient health record from "Home Monitoring Online system". It saves one copy to original database and sends another copy to randomizer. Randomizer possesses functionality to randomize or perturbation of those attributes of health record which could be disclosed the identity of any patient. The perturb data is then saved to database which is developed for research and mining purpose and which contains data with satisfying privacy constraints.

*B. Privacy Preserving Data Mining With Decision Tree*

In data mining research one of important issue is ensuring of privacy constraint. The aim of PPDM technique is to secure the input data and still retrieve useful knowledge from data miners. Various privacy preserving data mining technique have been recently proposed by many researchers. These techniques include some famous privacy preserving approach such as "k-anonymity", "l-diversity", "t-closeness" and so on which uses a statistical approach. The cryptography technique uses secure multi party computation to ensure privacy, but it suffers from their poor performance. Other hand statistical approach used to mine association rules [23-31], decision tree and clustering methods [32-33], this type of approach is very popular because of its good accuracy and high performance. This paper focuses on the statistical method to PPDM decision tree mining.




.

Figure 3: Framework of Privacy Preserving Decision Tree Mining.

Figure 2 has a fine and useful applications: the anonymized data can be analysed by arbitrarily various data miners using traditional mining algorithms. One side, an admin can protects its own data by releasing only the anonymized version. On the other side, a data miner equipped with a focus of the anonymization technique can derive a decision tree that is having a good accuracy, compared to the previous derived from the original dataset. This framework is very useful and accurate for many real life healthcare applications. The resulting PPDM Decision tree mining approach has the following properties.

- It is useful in the data recovery and repeated perturbation attacks.
- This approach provides assurance that the classification tree which uses anonymized data as input has high accuracy as compared to various other techniques.

## V. K-ANONYMITY

For the purpose of privacy preserving of individual record we release the micro data in which often remove explicit identifier such as P-id, name , social security numbers, but de-identifying of data generally not provide the guarantee of anonymization. Published data often contain other data known as quasi identifying attribute such as age, location and sex that can be merged with publicly available data and reveal data information that was not planned for release. This procedure is known as linking attack [1, 3]. K-anonymity proposed by L. Sweeney in 2002 provides solution for linking attack. K-anonymity defines every tuple in the database table released be distinct related to not less than k records means in other words in which a database is a table with number of rows and columns. Each row of the table shows related record and the record in the various rows need not be unique and for this purpose each quasi identifying set of values should appear at least k time in the related micro data [1, 3, 5]. The following table is an example of anonymized database consisting of patient records of a hospital.

TABLE I: ANONYMIZATION PROCESS USING K ANONYMITY TECHNIQUE.

| Age | Weight | Name | | Age | Weight | Name |
|---|---|---|---|---|---|---|
| 35 | 50 | Ramesh | ⇨ | [35,45] | [50,65] | Ramesh |
| 60 | 55 | Shweta | | [50,60] | [50,65] | Shweta |
| 65 | 50 | Shyam | | [55,65] | [50,65] | Shyam |

## VI. DECISION TREE

Decision trees are generally used in operations analysis, especially in decision making and useful to identify a strategy just like to reach an aim. Decision tree create from left to right and solve by right to left. In the representation of decision tree every non-leaf node represents as decision entity, each edge signifies to decision about attribute value and every leaf represent a class. Decision tree is mainly useful in medical diagnosis, credit risk research, business process. Decision tree node representation is described below three manners. Square (□) represents decision nodes. Circle (O) represents by chance nodes. Triangles (optional) represented by terminal nodes as shown in figure 4.

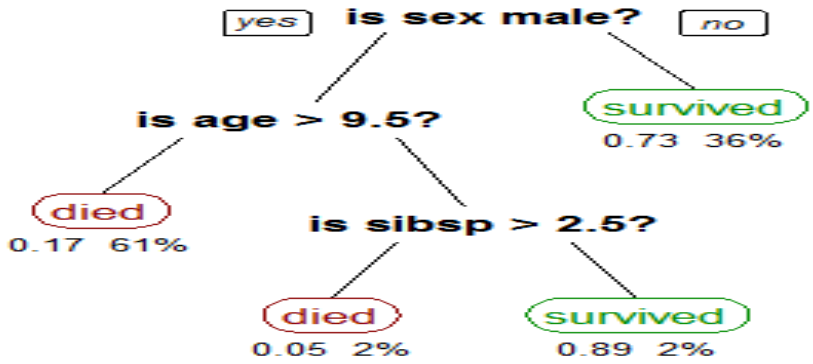



Figure 4: Example of Decision tree learning

In the above example every interior node belong to one of the input value, edge represent to each node for every possible value of the input value. Each and every leaf shows the value of the particular target value that is correspond to input value depicts by total sequence from top to bottom node. Following are the advantages of Decision Tree:

- This method is simple to use and understand. User friendly also.
- Possible value can be added.
- Various types of values can be easily determined according the scenario like that good or bad.
- This method generally used white box model.
- Might be merge with other decision tress techniques.
- With low hard data have even values.

### *A. Decision Tree Algorithm C (4.5)*

Improved and successor algorithm for the ID3 known as algorithm C4.5, which perform the better performance as comparison to other available algorithms. C4.5 algorithm is implemented either on rule sets or decision tree. In C4.5 algorithm the test sample is split based on the attribute that provides the highest information gain. After that each and every sub sample implemented by the initial split which is based on the various type of attribute, and repeated this process until the child node or sub sample cannot be split further. In the final process the terminal level splits are again evaluated, and those nodes that do not provide meaningful value to the model are pruned or removed. C4.5 has ability to produce two different variations of models [10,33].

Decision tree implemented here is providing description of the splits according the algorithm. In the classification tree every leaf node analysis a particular sub group of the training dataset and each and every case in this training datasets correspond to exactly one leave node. The C (4.5) classification tree algorithm requires calculating the two important metrics one is information entropy and other is information gain of the record

values in the database table. For calculation of information entropy, the following equation (1) is used, where probability is represented by $p_i$ of the particular class $c_i$ in the given table D [10].

$$\boldsymbol{Entropy(D)}=(\sum_{i=1}^{m} \boldsymbol{p_{i\times log_2}\ pi}) \quad \textbf{(1)}$$

The calculation of gain is derived by using the information entropy value; gain value is calculated by using the following equation (2) [36].

$$\boldsymbol{Gain(D)= Entropy(D) - \sum_{i=1}^{n} \frac{|D_i|}{|D|} \times Entropy(D_i)} \quad \textbf{(2)}$$

For all attribute value of the given table information gain is calculated, by which we are able to find out the attribute value with highest information. Accordingly attribute with highest or maximum gain, the node is split. This is the whole process of the decision tree C4.5 which provides accurate and efficient result compare to any other data mining technique.

## VII. PROPOSED METHODOLOGY

The Figure 5 shows the outline of the plan of work which clearly shows that the result has been achieved in steps in which data pre-processing takes place at the first step and while performing pre-processing the domain knowledge of medical experts has been considered because the values of different attributes of healthcare dataset has a specific significance which is known to medical experts. By the term anonymization using k- anonymity algorithm it is shown that we preserve privacy for datasets by using generalization and suppression method. In this step we apply generalization on quasi identifying attribute to avoid linking attack. In generalization process we converted original attribute value with in some range value. After that we calculate the mean value of the range and create new dataset in CSV format. For apply next step decision tree we converted this CSV format dataset into .arff format, there after we apply decision tree C4.5 and create privacy preserve decision tree which is useful in the clinical trials and medical research. We also calculated the performance of the proposed methodology in term of accuracy, error rate, F- measure etc. For justification of our result we compare our proposed methodology result to the original dataset without using privacy constraints result and we conclude that our proposed method result is approximately same as the original dataset result.

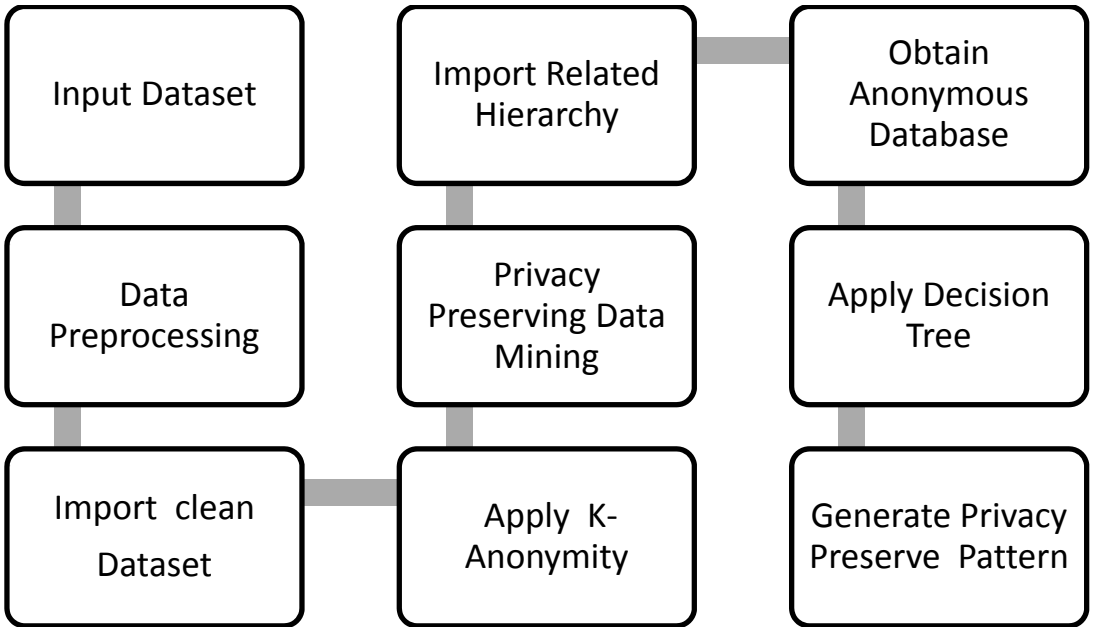



Figure 5: Proposed framework for PPDM In Electronic Health Record

For implementation of k-anonymity and decision tree we used Java language in the eclipse IDE. We follow following step to achieve a classifier that can predict accurate pattern while preserving individual privacy.

Step 1:- We begin with apply k –anonymity in original medical datasets, by which we ensure for anonymous data mining result. K-anonymity provides identity protection with generalization and suppression method. K-anonymity proposed by L. Sweeney in 2002 provides solution for linking attack. K-anonymity defines every tuple in the database table released be distinct related to not less than k records means in other words in which a database is a table with number of rows and columns. Each row of the table shows related record and the record in the various rows need not be unique and for this purpose each quasi identifying set of values should appear at least k time in the related micro data. We exemplify with some medical datasets and improve the accuracy of k-anonymity with decision tree C (4.5) and validate that anonymization algorithm is useful with classification tree C4.5 algorithm to ensure k-anonymous result. This process include following steps:-

- Firstly import medical data.
- Create privacy level hierarchy for each quasi identify attribute in healthcare dataset.
- Apply K-anonymity algorithm on healthcare dataset for identity disclosure and also prevent from linking attack.
- By which we achieve anonymous dataset. After that we perform on that anonymous dataset step 2 processing.

Step 2:- We applied classification tree C4.5 in our k- anonymous results which provide better accuracy rather than other data mining technique. For efficient and accurate result we used the improved and successor of the ID3 algorithm known as algorithm C4.5. C4.5 algorithm is implemented either on rule sets or decision tree. In C4.5 algorithm the test sample is split based on the attribute that provides the highest information gain. After that each and every sub sample implemented by the initial split which is based on the various type of attribute, and repeated this process until the child node or sub sample cannot be split further. In the final process the terminal level splits are again evaluated, and those nodes that do not provide meaningful value to the model are pruned or removed. C4.5 has ability to produce two different variations of models. For achieve the accurate result we anonymizing the data first and after that apply decision tree C4.5 on anonymous datasets later. This process include following steps:-

- It presents decision tree algorithms based C4.5, which ensure k-anonymous output and this method is more accurate than existing methodology.
- It shows attribute generalization within decision tree.
- It analysis the privacy implications of using record with missing values for decision tree C4.5.

Step 3:- Perform comparatively analysis between the obtained classification accuracy with applying anonymous healthcare datasets to the original classification accuracy without using k-anonymity. Study the privacy/accuracy tradeoff in the context of information loss. The C4.5 algorithm requires calculating the two important metrics one is information entropy and other is information gain of the record values in the database table. Information entropy and gain are calculated by using equation 1 and 2.

## VIII. EXPERIMENTAL RESULT AND ANALYSIS

We have created a large personal healthcare datasets for patient's records and test the proposed approach with this data. The dataset contains the name of the patient, age, gender, Blood Pressure, Address, Disease, Health History, Medications, Allergies, Diet, Height, Mobile Number, Marital Status. These are the techniques for getting secure data base for patient privacy. We select the dataset randomly from the web sites. In this approach, we make the record similar so they cannot identify the actual record.

### *A. Implementation Detail of K-anonymity*

Step 1:- Implement java code for k anonymity in Eclipse IDE. In the running process of k –anonymity, firstly in which create a project and save in particular location.

Step 2:- Import dataset as shown in figure 6.

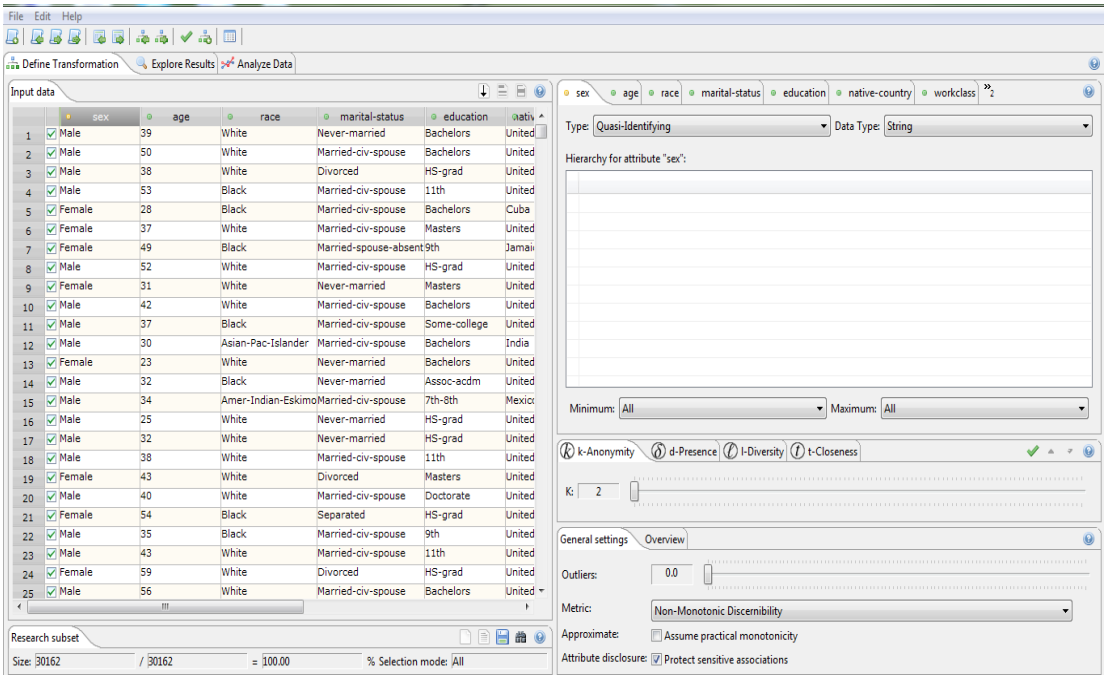



Figure 6: Snapshot of Import dataset

Step 3:- Before apply k-anonymity process we create hierarchy of each attributes which define the privacy level, means how much range we want to generalize in record table for each attribute.

For this purpose import hierarchy for particular attribute in this step and also define the privacy level k for anonymization process. For example we import hierarchy of age attribute and perform generalization process on this by using k-anonymity process. In which privacy level k=3

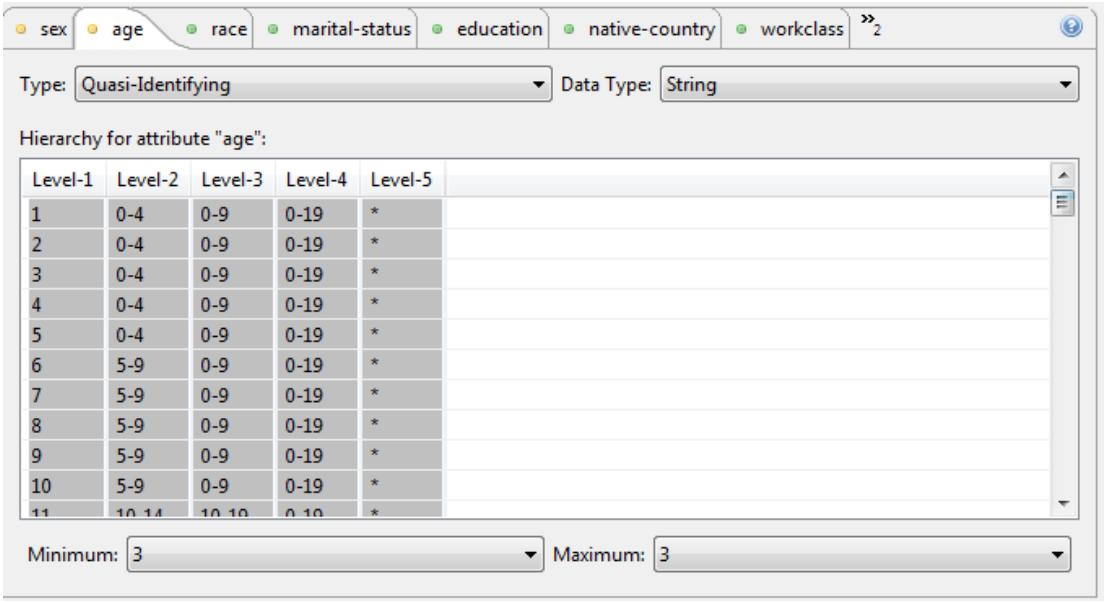



Figure 7: Snapshot of generalization process of age attributes

Step 4:- After import hierarchy of age attribute perform anonymization by using k-anonymity algorithm and get privacy preserve dataset which is useful for make identity private and also prevent for linking attack in medical research. In which age attribute is generalize by using range of 10.

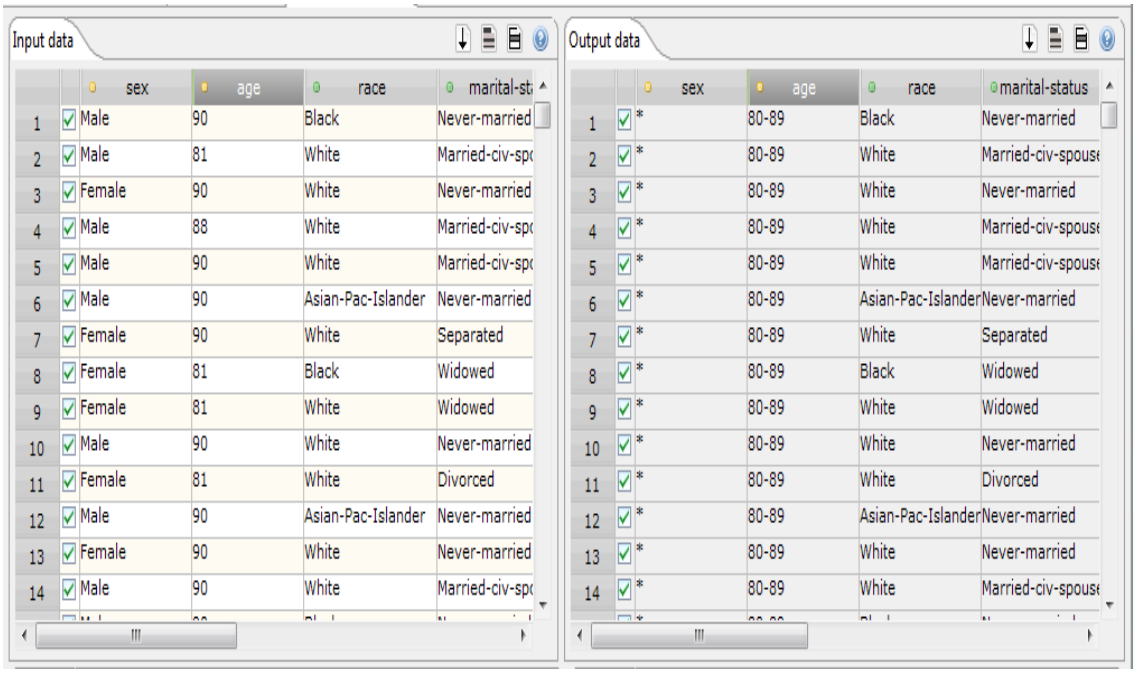



Figure 8: Snapshot of k-anonymity anonymization result

Step 5:- Repeat anonymization process on quasi identifier attribute to prevent record form linking attack using k-anonymity.

Step 6:- Perform k-anonymity algorithm in quasi identifier attribute we found out anonymous datasets. Export anonymous dataset on particular disk location for further processing which is also in CSV format.

*B. Implementation process of Decision tree*

Apply decision tree C4.5 on the anonymous dataset for generating pattern. Decision tree C4.5 is also implemented in Java with eclipse IDE.

Step 1:- Read anonymous healthcare dataset.

Step 2:- Consider the quasi identifying attribute and identify the mean values of every range. Here nine ranges have been chosen.

Step 3:- Converted .csv to .arff format.

Step 4:- Create Training example and test example.

Step 5:- Create decision tree. In which define correctly classified dataset and incorrectly classified dataset accordingly performance evolution is calculated for example accuracy, error rate, F-measure etc.

Step 6:- We visualize decision tree by clicking option button generate tree.

Table 2 and figure 9 show the result of comparison of the classifier with the original dataset and dataset with k-anonymity.

TABLE II. RESULT COMPARISON

| Performance Measure | Original Dataset + Decision Tree | Dataset with K-anonymity and Decision Tree |
|---|---|---|
| Accuracy | 82.9 | 82.8 |
| Error Rate | 17.1 | 17.2 |
| Precision | 0.6213 | 0.6168 |
| Recall | 0.6921 | 0.6934 |
| F-Measure | 0.5636 | 0.5555 |
| Kappa Value | 0.5124 | 0.5079 |

From the table, it is observed that the proposed approach has gained comparable performance i.e, when we applied privacy preserving approach, it does not deteriorate the performance of classifier. So, with the proposed approach we maintain privacy of the patient as well as obtain better results.

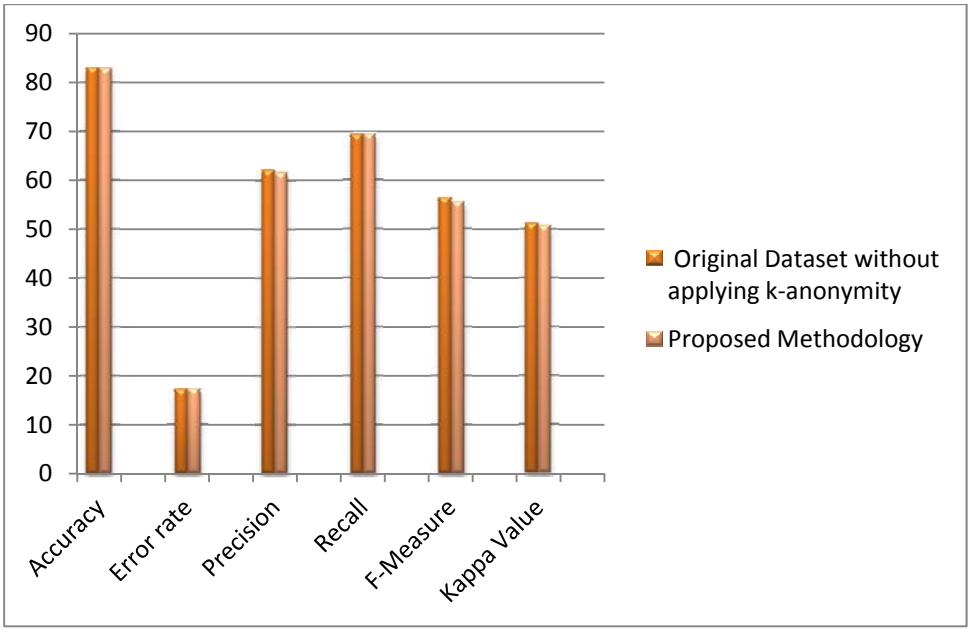



Figure 9: Performance Comparison

## IX. CONCLUSION

In this research work, we have proposed a method that satisfies privacy constraint and its utility for data mining in EHR. We have used k- anonymity algorithm to ensure data privacy in EHR and C4.5 decision tree, which provides more accurate and error free measure of classification for research in healthcare systems. The experimental results are analyzed by using some parameters such as accuracy, error rate measure, precision, recall, F- measure and Kappa values. We have compared the results between original datasets without applying k anonymity and our privacy preserve proposed methodology. From the experimental results, it is clear that the result for proposed methodology with privacy constraints approximately similar to the original datasets without using privacy constraints and we are successfully provide trade-offs between utility and privacy.

In future, k-anonymity algorithm could be utilized with other data mining techniques. Other future scope of this work is to implement a website of developed software, where according the anonymous dataset we are able to predict future disease of various age group patients that is a very useful for medical research or clinical trials purpose without revealing the identity of the patient.

## REFERENCES

[1] A. A. AlShwaier and A. Z. Emam, “Data Privacy On E-Health Care System”, International Journal of Engineering, Business and Enterprise Applications, (2013).
[2] Li, Tiancheng, and Ninghui Li. "Towards optimal k-anonymization." Data & Knowledge Engineering 65, no. 1, pp.22-39, (2008).
[3] Xu, Yang, Tinghuai Ma, Meili Tang, and Wei Tian. "A survey of privacy preserving data publishing using generalization and suppression." Appl. Math 8, no. 3, pp. 1103-1116, (2014).
[4] Singh, Neera, Sonali Agarwal, and Ramesh C. Tripathi. "A Data Mining Perspective on the Prevalence of Polio in India." International Journal on Computer Science and Engineering 3, no. 2, pp.580-585, (2011).
[5] Wong, Raymond Chi-Wing, Jiuyong Li, Ada Wai-Chee Fu, and Ke Wang. "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing." In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 754-759. ACM, (2006).
[6] Tomar, Divya, and Sonali Agarwal. "A survey on pre-processing and postprocessing techniques in data mining." International Journal of Database Theory & Application 7, no. 4, (2014).
[7] Xiao, Xiaokui, and Yufei Tao. "Personalized privacy preservation." InProceedings of the 2006 ACM SIGMOD international conference on Management of data, pp. 229-240. ACM, (2006).
[8] Sweeney, Latanya. "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, no. 05, pp.557-570, (2002).
[9] Bayardo, Roberto J., and Rakesh Agrawal. "Data privacy through optimal k-anonymization." In Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on, pp. 217-228. IEEE, (2005).
[10] Han, Jiawei, Micheline Kamber, and Jian Pei. “Data mining: concepts and techniques: concepts and techniques”. Elsevier, (2011).
[11] Agarwal, Sonali, G. N. Pandey, and M. D. Tiwari. "Data mining in education: data classification and decision tree approach." International Journal of e-Education, e-Business, e-Management and e-Learning, 2 (2), pp.140-144, (2012).
[12] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5, no. 5, pp.241-266, (2013).
[13] Agarwal, Sonali, and G. N. Pandey. "SVM based context awareness using body area sensor network for pervasive healthcare monitoring." In Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia, pp. 271-278. ACM, (2010).
[14] Ghani, Mohd Khanapi Abd, and Lee Chew Wen. "The design of flexible Pervasive Electronic Health Record (PEHR)." In Humanities, Science and Engineering (CHUSER), 2011 IEEE Colloquium on, pp. 249-254. IEEE, 2011.

[15] Hsu, William, Ricky K. Taira, Suzie El-Saden, Hooshang Kangarloo, and Alex AT Bui. "Context-based electronic health record: toward patient specific healthcare." Information Technology in Biomedicine, IEEE Transactions on 16, no. 2, pp.228-234, (2012).
[16] Agarwal, Sonali, Neera Singh, and G. N. Pandey. "Implementation of Data Mining and Data Warehousing In E-Governance." International Journal of Computer Applications 9, no. 4, pp. 18-22, (2010).
[17] Tomar, Divya, Shubham Singhal, and Sonali Agarwal. "Weighted least square twin support vector machine for imbalanced dataset." International Journal of Database Theory and Application 7, no. 2, 25-36, (2014).
[18] Agarwal, Sonali, and Divya Tomar. "A feature selection based model for software defect prediction." International Journal of Advanced Science and Technology, vol. 65, pp.39-58, (2014).
[19] Nergiz, Mehmet Ercan, Maurizio Atzori, and Chris Clifton. "Hiding the presence of individuals from shared databases." In Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pp. 665-676. ACM, (2007).
[20] Rathore, Nisha, Divya Tomar, and Sankalp Agarwal. "Predicting the survivability of breast cancer patients using ensemble approach." In Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on, pp. 459-464. IEEE, 2014.
[21] Nath, Sayantan, Sonali Agarwal, and Qasima Abbas Kazmi. "Image histogram segmentation by multi-level thresholding using hill climbing algorithm." Int. J. Comput. Appl 35, no. 1 (2011).
[22] Chawla, Shuchi, Cynthia Dwork, Frank McSherry, Adam Smith, and Hoeteck Wee. "Toward privacy in public databases." In Theory of Cryptography, pp. 363-385. Springer Berlin Heidelberg, (2005).
[23] Agarwal, Sonali. "Weighted support vector regression approach for remote healthcare monitoring." In Recent Trends in Information Technology (ICRTIT), 2011 International Conference on, pp. 969-974. IEEE, 2011.
[24] Tomar, Divya, and Sonali Agarwal. "Hybrid Feature Selection Based Weighted Least Squares Twin Support Vector Machine Approach for Diagnosing Breast Cancer, Hepatitis, and Diabetes." Advances in Artificial Neural Systems 2015 (2015).
[25] Rastogi, Vibhor, Dan Suciu, and Sungho Hong. "The boundary between privacy and utility in data publishing." In Proceedings of the 33rd international conference on Very large data bases, pp. 531-542. VLDB Endowment, 2007.
[26] Agrawal, Shipra, and Jayant R. Haritsa. "A framework for high-accuracy privacy-preserving mining." In Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on, pp. 193-204. IEEE, 2005.
[27] Tomar, Divya, and Sonali Agarwal. "A comparison on multi-class classification methods based on least squares twin support vector machine." Knowledge-Based Systems 81 (2015): 131-147.
[28] Evfimievski, Alexandre, Johannes Gehrke, and Ramakrishnan Srikant. "Limiting privacy breaches in privacy preserving data mining." In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 211-222. ACM, (2003).
[29] Tomar, Divya, Bakshi Rohit Prasad, and Sonali Agarwal. "An efficient Parkinson disease diagnosis system based on Least Squares Twin Support Vector Machine and Particle Swarm Optimization." In Industrial and Information Systems (ICIIS), 2014 9th International Conference on, pp. 1-6. IEEE, 2014.
[30] Tomar, Divya and Sonali Agarwal, "Predictive model for diabetic patients using hybrid twin support vector machine." 5th International Conferences on advances in communication Network and Computing (CNC-2014), (2014).
[31] Khanna, Subham, and Sonali Agarwal. "An Integrated Approach towards the prediction of Likelihood of Diabetes." In Machine Intelligence and Research Advancement (ICMIRA), 2013 International Conference on, pp. 294-298. IEEE, 2013.
[32] Merugu, Srujana, and Joydeep Ghosh. "Privacy-preserving distributed clustering using generative models." In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, pp. 211-218. IEEE, (2003).
[33] Agrawal, Rakesh, and Ramakrishnan Srikant. "Privacy-preserving data mining." In ACM Sigmod Record, vol. 29, no. 2, pp. 439-450. ACM, (2000).