Personalization of Healthcare System using RNN and Blockchains

Kahandal Kalyani[1] Simran Sonawane[2] Shah Salik[3] Prof. S.K.Shinde[4]
[1,2,3,4]Department of Computer
[1,2,3,4]KJCOEMR, Pune, India

Abstract— The maintenance of health is one of the central aspects of a good life. There have been significant advancements in the paradigm of Health care in recent years which are assistive to doctors and medical practitioners. One such paradigm is the PHR or personalized Health Records that allows the creation of a repository of the various parameters of a patient and the treatment offered to such patients for his/her relief. The PHR paradigm is a highly useful reference for the doctors as it allows them to effectively diagnose the symptoms and offer accurate treatment. The problem arises when a patient with a unique set of symptoms and parameters is encountered. This is a problematic occurrence as without any history the doctor is forced to try the trial and error technique to identify the symptoms. Therefore, for this purpose, there is a need for an effective technique for the secure sharing of PHR data between the medical institutions. For this purpose, various data vendors facilitate the exchange of data between hospitals and medical organizations. To secure the highly sensitive PHR's there is a need for an effective and secure access control mechanism that is described in this publication. The proposed methodology secures the PHR data through the implementation of the Distributed Blockchain framework which is assisted by the X means clustering and Decision Tree along with the introduction of an effective access control mechanism through the use of Recurrent Neural Networks.

Keywords: Access Control Mechanism, Blockchains, RNN, Personalized health care system, Decision Tree

I. INTRODUCTION

Human health is one of the most important and central pillars of the survival of the human race. Ever since the Stone Age, man has been in the pursuit of a much more fulfilling and Secure Life Style. And hence humans have been working towards this goal since millennia. PHR or Public Health records are records of various diagnosis and treatments offered to a patient regarding their element or illness. These records are highly valuable as they contain valuable information about various treatments that are performed in the event of observation of similar symptoms. This is a valuable repository that can help the doctor to refer every type a case is encountered with symptoms that are similar to the ones encountered before. This reduces any problems that will happen if a trial error system is used. Therefore, the PHR or Public Health record system is one of the most useful and valuable resources of a particular medical Institution. This can help reduce the time taken for the recovery of a patient as well as reduce the time wasted by a medical professional in diagnosing illness. The paradigm of Public Health records would gain significantly from the introduction of the blocks in Framework as it would allow for the implementation of effective access control in the system. The blockchain networks also highly tamper-proof and would not allow any tampering that can be done on the public health records. This ensures the safety and security of the public health records by the distributed ledger blockchain system.

In this publication section 2 deeply explains the Implemented methodology techniques, whereas section 3 Discusses the obtained results. The future traces and conclusion of this publication is described in section 4.

II. IMPLEMENTED METHODOLOGY

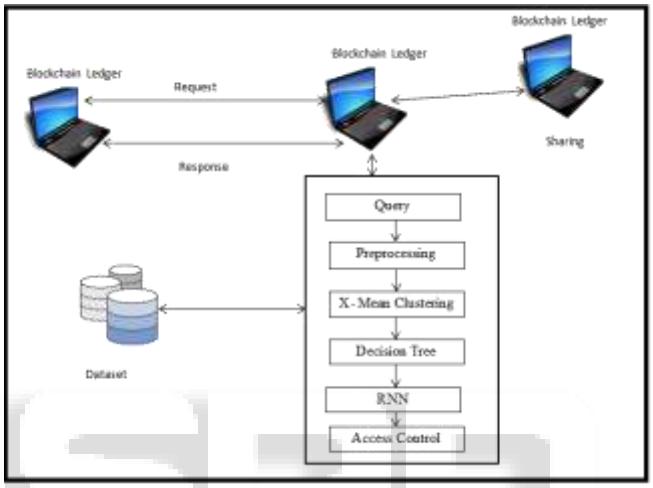

Fig. 1: Proposed model System Overview

The presented technique for the implementation of an efficient and secure access control mechanism for tamper-proof PHR data sharing with the help of the blockchain distributed framework is detailed in the system overview diagram given in figure 1. There is a collection of steps that need to be ensured to achieve the prescribed goals in the proposed methodology. These steps are discussed in detail in the section given below. There are 5 steps that achieve the complete methodology.

Step 1: Data Seeking and Data preprocessing –As the doctor encounters a patient with a collection of symptoms that can be an indication of a variety of diseases, the doctor references the PHR of the patient to look for more clues. When the PHR of that person is not available, the doctor seeks the data from the prescribed methodology by passing a Query to the system.

The query contains various parameters of the patient along with the symptoms being experienced by the patient which needs to be preprocessed and conditioned before being passed to the next step. The proposed methodology preprocesses the query by adhering to the 4 steps like Special Symbol Removal, Tokenization, Stopword Removal and Stemming, and this can be seen in figure 2.

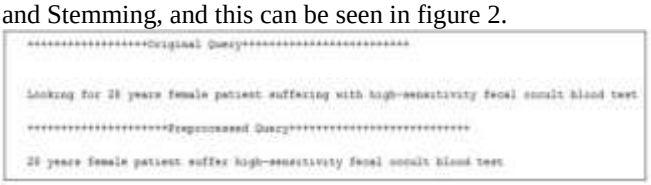

Fig. 2: String Preprocessing

Step2: X Means Clustering- The dataset is downloaded and clustered through the preprocessed query obtained in the above steps. The PHR dataset utilized for this purpose is downloaded from the URL: https://data.world/arvin6/medical-records-10-yrs.
The dataset extracted from the above URL contains the PHR data with extensive parameters such as symptoms, diseases, age and sex in addition to the treatment offered to the respective patients. The soapnotes attribute in the dataset contains the treatment details of the subject which is processed through the use of the query that is passed by the data seeker. A double dimension list called count list is created with the dimensions such as encounter id and the count to query.

The X Means Clustering is then performed on the count list for the creation of the clusters, which is achieved through the some Steps that are detailed below. Distance Evaluation- The count list created in the previous step is utilized for distance alculation and the results are added to the end of the particular row. The distance calculation is done through the use of all the other rows in the list and then the mean of the rows is stored at the end of the respective rows. The distance is calculated by performing the Euclidean Distance through the Equation given in 1.

$$AD = \sqrt{(a1 - a2)^2 + (b1 - b2)^2} \quad \text{-------} \quad (1)$$

$$AVG_{ED} = \sum_{k=0}^{n} AD/n \quad \text{_____} \quad (2)$$

Where
a1, a2, b1, b2 are the values of the Rows.
AD- Euclidean distance of a particular row
AVG_{ED} = Mean Euclidean Distance
n = Number of Rows

Sorting and Data Point Selection- The sorting of the Euclidean distance in the list obtained in the previous step is done through the utilization of the Bubble Sort algorithm. The X is the number of clusters to be formed in the proposed methodology, which is selected by taking a sorted list obtained in the previous step and selecting X random rows form it.

Boundary Evaluation and Cluster Formation- The data points are utilized for the centroid evaluation through the row Euclidean distance obtained in the previous steps. The boundaries of the clusters are determined by utilizing the centroid Euclidean distance obtained in the previous step. The lower limit of the cluster boundary is calculated by subtracting the centroid Euclidean distance with the mean Euclidean distance of the entire list. In the same way, by the addition of the mean Euclidean distance to the centroid distance, the upper boundary of the cluster is achieved. Therefore, Upper and Lower boundaries are stored in a boundary list created through the use of a double dimension list.

Cluster Formation- This is the last step in the X means clustering where the rows are distributed into their respective clusters utilizing the Row Euclidean distances. The X clusters are formed by comparison of the cluster boundary range with the row Euclidean Distances such as min and max values.

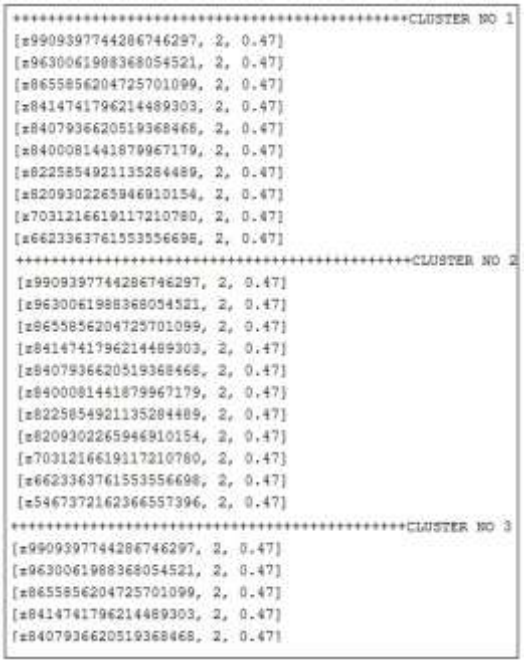

Fig. 3: X-Mean clusters

Step3: Decision Making – For the estimation of the best cluster containing the data requested through the query is obtained through the application of the logarithmic transformations on each cluster. The query match count is utilized on the clusters obtained, if the count obtained is more than 2, then it is considered as positive count and denoted as P. The negative count in the cluster is set as N and the total number of rows are set as T. Therefore, the negative count can be calculated as N (N =T-P). On the application of the logarithmic equation as depicted in equation 3 for Decision Tree analysis of Shannon information gain , which extracts the value of L_V that falls between 0 to 1 for every cluster. The value of L_V closer to 1 denotes that the cluster is of the best quality for implementation in the process of RNN. The RNN input list is created through the use of the obtained logarithmic value and adding it into an array after sorting the values in descending order. This array is then transferred to the RNN module which is the next step.

$$L_V - \frac{P}{T}\log\left(\frac{P}{T}\right) - \frac{N}{T}\log\left(\frac{N}{T}\right) \quad \text{_____} \quad (3)$$

This process of Decision Tree is can be depicted through the below mentioned algorithm1.

ALGORITHM 1: Decision Tree through Entropy Analysis

//Input : X-Mean Clusters XC_L
//Output: Entropy List Of clusters E_L
1: Start
2: FOR i=0 to Size of XC_L
3: $S_{CL} = XC_{L\,[i]}$ [S_{CL} = Single Cluster]
4: $I_{LST} = \emptyset$ [I_{LST} = Interim List]
5: count=0
6: FOR j=0 to Size of S_{CL}
7: $T_L = S_{CL\,[j]}$ [T_L = Temporary List]
8: CT= $T_{L\,[1]}$ [Query Count]
9: IF(CT > 2)
10: count++

11: End FOR
12: P=count, T= Size of $S_{CL,}$ N=T-P
13: E= (-P/T) log(P/T) (-N/T) log(N/T)
14: $I_{LST[0]}$=I , $I_{LST[1]}$=E
15: $E_{L=}$ E_L+ I_{LST}
16: END For
17: return E_L
18: Stop

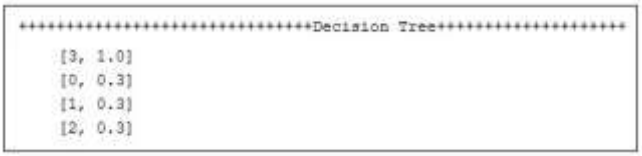

Fig. 4: Cluster Indices and respective Scores of Entropy

Step 4: Access Control Mechanism: The best match of the clusters according to the Query are the uppermost clusters in the RNN input list obtained from the previous step out of which the top two are selected. The two clusters are then utilized to create a single RNN list that constitutes only the best two clusters produced. Through the evaluation of the maximum and minimum values in the query count Target 1 and target 2 variables for the RNN are selected.

The Random class belonging to the Java programing language is used for the allocation of ten Random weights in between the ranges 0 to 1 in the form of variables such as W1, W2, W3, W4, W5, W6, W7, W8, b1, b2. Here the values b1 and b2 are assigned as bias weights.

The RNN list is utilized for the estimation of the hidden layer for all of the rows using the equation detailed below.

$$X= (AT1* W1) +(AT2*w2) +b1 \quad \text{_____ (4)}$$

$$H_{LV}= \frac{1}{1 + \exp(-X)} \quad \text{_____ (5)}$$

Where AT1 is the query matching count and AT2 is the Euclidean distance. Then the sigmoid function is given by Equation 5 of the RNN. Target1 and Target2 variables are then utilized to calculate the target error. Thus, the output layer can be calculated through these two variables that are obtained. The probability values for each row are obtained through the evaluation of the two hidden values along with the respective weight entities. The rows obtained in the previous step are optimized using the best matching rows selected using the access control mechanism which achieves 99% accuracy of optimization. Each of the rows and their prediction value frequency is utilized to decide the access control level based on the different attributes. The selected attributes and their characters are then replaced by empty characters or by the '*' character. The data must be transformed before sending it to the data vendor through the utilization of the access control mechanism levels calculated by the proposed methodology.

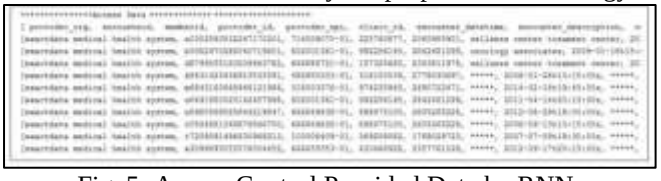

Fig. 5: Access Control Provided Data by RNN

Step 5: Blockchain – The blockchain framework is one of the most important implementations in this methodology. The data obtained in the previous step is treated as the block in this methodology. The block contains the access-controlled data that is then utilized to create a hash key through the use of the MD5 algorithm.

The hash key is then significantly transformed and reduced to a shareable length that can be utilized for sharing to the data vendor. The data vendor on receiving the hash key containing the access-controlled data is then transferred to the data seeker through an email. The data seeker can then utilize this hash key by entering it into the system to access the requested data securely through the tamper-proof distributed blockchain framework.

III. RESULTS AND DISCUSSIONS

The proposed methodology for establishing an effective access control mechanism has been developed in the Java programming language on the NetBeans IDE. For the execution of the presented technique, three development machines are utilized which are furnished with an Intel i5 processor for the processing requirements along with 4GB of primary memory as RAM and 500 GB of storage. The database responsibilities are fulfilled by the MySQL database server along with a D-Link WIFI router for networking purposes.

Extensive Experimentation was conducted to calculate the performance metrics of the presented technique. To calculate the accuracy of the proposed methodology, the Precision and Recall analysis technique was utilized which can accurately outline the performance metrics of the proposed methodology. The performance metrics were calculated to ascertain that the access control mechanism based on the distributed framework of the blockchain platform presented in this paper is executing properly.

A. Performance Evaluation based on Precision and Recall

Precision and Recall can provide insightful information associated with the performance of the presented technique. The precision and recall parameters are one of the most accurate and useful parameters that are applied to enable the extraction of the absolute performance of the system. Precision in this experimentation evaluates the relative accuracy of the proposed methodology by extracting the accurate values of the level of precision achieved in the system.

Precision in this experimentation is being calculated as the ratio of the combined sum of all the queries that have been matched to the number of accurate predictions achieved for granting access control mechanism. Therefore, the evaluation of the values of precision results in an in-depth analysis of the accuracy of the presented technique.

The Recall parameters are used for evaluation of the absolute accuracy of the system which is incomparable to the precision parameters.

The Recall parameters are evaluated by calculating the ratio of the number of accurate predictions for the given query matched versus the total number of inaccurate

predictions for the given query matched for granting the data access control. This analysis technique provides valuable information as it evaluates the absolute accuracy of the system. Precision and recall are mathematically explained in the equations given below.

Precision can be concisely explained as below

A = The number of accurate Access control applied for the given query using RNN

B= The number of inaccurate Access control applied for the given query using RNN

C = The number of accurate Access control not applied for the given query using RNN

So, precision can be defined as

$$\text{Precision} = (A / (A+ B)) *100$$

$$\text{Recall} = (A / (A+ C)) *100$$

The equations detailed above are utilized for executing extensive experimentation on the presented technique through the extraction of the results of the Recurrent Neural Network module. The experimental results are tabulated in Table 1, given below.

No of Queries	Accurate Access Control Applied (A)	Inaccurate Access Control applied (B)	Accurate Access Control not applied (C)	Precision	Recall
131	120	4	7	96.77419355	94.48818898
32	25	2	4	92.59259259	86.20689655
116	99	8	9	92.52336449	91.66666667
59	47	7	5	87.03703704	90.38461538
89	79	4	6	95.18072289	92.94117647

Table 1: Precision and Recall Measurement Table for the performance of RNN

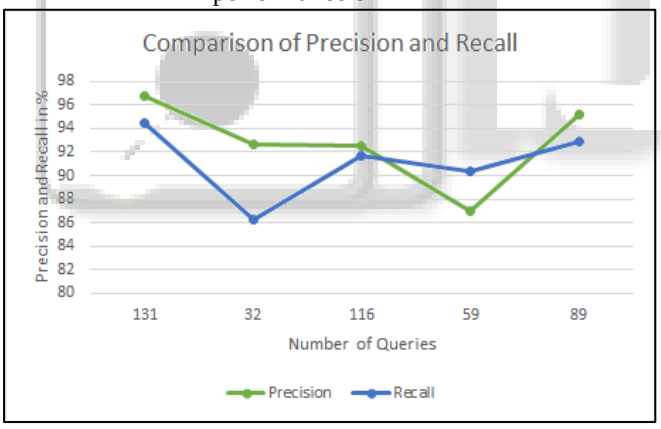

Fig. 6: Comparison of Precision and Recall for the performance of ANN

The graph depicted above illustrates that the Recurrent Neural Network in the presented system achieves satisfactory parameters of precision and recall for the enabling of access control mechanism. The model achieved the precision of 92.82% and Recall of 91.13% and this precision and recall parameters illustrate that the RNN module performs with very high accuracy and efficiency in the presented technique.

These experimentation results explain that the RNN analysis module is being executed accurately in the presented technique and is highly accurate in its performance. The Recurrent Neural Network is one of the central concepts in the presented technique and the accurate and successful demonstration of this module improves the performance of the proposed Access control mechanism significantly.

IV. Conclusion and Future Scope

The PHR or Personal Health records are one of the most valuable resources of the medical institutions which helps the doctors achieve significant improvement in the diagnosing and treatment capabilities. The PHR allows the streamlining of the treatment process by effectively reducing the pain and suffering of the patient by a large margin. The main problem associated with the PHR is that the data contained in the PHR can often include highly sensitive and personally identifiable information that cannot be accessed by any other individual. This introduces a problem that does not allow the free exchange and sharing of PHR which can be detrimental to a patient's well-being. Therefore, this paper outlines an effective technique for the implementation of a secure access control system through the use of X Means Clustering and Decision Tree along with the use of Recurrent Neural Network. The extensive experimental evaluation of the presented technique concludes that this methodology has been a significant improvement over the traditional PHR sharing techniques.

For future research, the presented technique can be deployed in a real-life scenario with real-time data in a medical institution. The proposed access control mechanism can also be deployed in various fields such as Law, Agriculture, etc