Evolutionary tree-based quasi identifier and federated gradient privacy preservations over big healthcare data

Sujatha Krishna, Udayarani Vinayaka Murthy
School of Computing and Information Technology, REVA University, Bengaluru, India

Article Info

Article history:

Received Apr 14, 2021
Revised Jul 16, 2021
Accepted Aug 4, 2021

Keywords:

Adaptive Lorentz
Big data
Evolutionary tree
Federated gradient arbitrary
Privacy preservation
Quasi identifier

ABSTRACT

Big data has remodeled the way organizations supervise, examine and leverage data in any industry. To safeguard sensitive data from public contraventions, several countries investigated this issue and carried out privacy protection mechanism. With the aid of quasi-identifiers privacy is not said to be preserved to a greater extent. This paper proposes a method called evolutionary tree-based quasi-identifier and federated gradient (ETQI-FD) for privacy preservations over big healthcare data. The first step involved in the ETQI-FD is learning quasi-identifiers. Learning quasi-identifiers by employing information loss function separately for categorical and numerical attributes accomplishes both the largest dissimilarities and partition without a comprehensive exploration between tuples of features or attributes. Next with the learnt quasi-identifiers, privacy preservation of data item is made by applying federated gradient arbitrary privacy preservation learning model. This model attains optimal balance between privacy and accuracy. In the federated gradient privacy preservation learning model, we evaluate the determinant of each attribute to the outputs. Then injecting Adaptive Lorentz noise to data attributes our ETQI-FD significantly minimizes the influence of noise on the final results and therefore contributing to privacy and accuracy. An experimental evaluation of ETQI-FD method achieves better accuracy and privacy than the existing methods.

Corresponding Author:

Sujatha Krishna
School of Computing and Information Technology, REVA University
Bengaluru, India
Email: sujathasjcit@gmail.com

1. INTRODUCTION

In the recent few years, the data volume has expanded in an exponential manner and among this data there is an increasing amount of personal information contained within. This sensitive data has fascinated the recognition of those interested in producing more customized and personalized applications. This in turn infringes the individual privacy and ushers to the concerns that personal data may be broken and falsified. As a consequence, this occurrence has ushered new ultimatums to safeguard the data privacy as a key issue in privacy preserving health care data.

Attribute centric anonymization scheme was proposed in [1] to safeguard from identity disclosure even posed with malicious users possessing certain amount of background knowledge or information. However, the execution time and memory incurred in anonymization was less focused. A robust anonymization and risk assessment scheme was designed in [2] that achieved four different objectives for bio medical data. Despite minimum execution time required for anonymization, a tradeoff between privacy and accuracy is said to be occurred. Best seed values were identified in [3] to minimize the information loss. However, learning from outliers and imbalanced data is still found to be one of the major drawbacks for privacy preservation. Among the several strategies presented to solve this issue, data preprocessing solutions [4], [5] are found to be effective both in solving and implementation. Privacy preservation methods [6], [7] were discussed.

A state of the art security and privacy challenges were discussed in detail in [8]. Yet another multi-label ensemble classification approach including decision tree algorithms was designed in [9]. A concrete survey on user privacy was investigated in [10]. A double decryption algorithm equipping the public key encryption with differential privacy was proposed in [11], therefore contributing to security. Yet another data warehouse solution called hive was presented in [12] using nearest similarity based clustering. A convolutional neural network (CNN) was customized for preserving the privacy via mapping [13] and deep learning [14] for recording electronic health sequences. However, all of these attributes are said to possess features that are said to be both sensitive and quasi attributes in general. A separate anonymization and reconstruction algorithm was designed in [15] using real dataset. Rao *et al.* [16], a survey of privacy preservation techniques for big data was investigated. Temuujin *et al.* [17], l-diversity algorithm along with cuckoo filter was proposed to enhance the data processing efficiency. However, information loss was not focused.

Yet another two step clustering method was designed in [18] with the aid of equivalence classes to reduce the information loss of anonymous datasets. Despite minimum information loss, tradeoff between data privacy and quality was identified. To address this issue, conditional probability distribution along with Gibbs sampling was proposed in [19] therefore retaining better data utility. Recommender systems was designed in [20] with the objective of minimizing the query response. A new conjugate gradient (CG) method was introduced in [21] to solve the optimization issues. Clustering methods was developed in [22] with aid of log data. Metaheuristic optimization in neural network model was introduced in [23] for time series modeling. A hybrid algorithm was developed in [24] for increasing the security in e-business systems. Data-mining classification algorithms were introduced in [25] to detect the lung and breast cancer diagnose. New approach was developed in [26] to decision-making based on the characterization of cognitive tasks. Big data privacy models were introduced by the means of data masking methods.

The major issues identified in the most of the existing privacy preservation mechanism tries to optimal a single objective, like either minimizing the information loss incurred during identification of quasi identifiers or enhancing the privacy preservation accuracy. However, single objective optimization may reduce the significance and efficient of healthcare data in general. In addition, the learning aspects involved in privacy preservation was less concentrated. So, to run such types of applications or tasks i.e., preserving privacy of big healthcare data with minimum information loss, communication overhead and higher privacy preservation accuracy is one of the challenging issues. To overcome the issue, evolutionary tree-based quasi-identifier and federated gradient (ETQI-FD) is developed for privacy preservations involving big healthcare data.

In this paper, we have designed an evolutionary tree-based indexed quasi identification algorithm. Here, the numerical and categorical attributes are not merged into single data node. This may minimize the communication overhead involved in identification of quasi identifiers. The proposed ETQI-FD method also with the deployment of federated adaptive Lorentz privacy preservation algorithm minimizes the information loss involved in privacy preservation. Arbitrary privacy-preserving adaptation (APA) function is used to enhance the accuracy. The main novelty and contribution of the proposed method are summarized as follows:

- The main contribution of the proposed ETQI-FD method is introduced for finding the optimal quasi identifiers and thereby preserving the privacy of healthcare data. The contribution is achieved with the novelty of the ETQI model, and federated adaptive Lorentz privacy preservation algorithm.
- On the contrary to existing works, the ETQI-FD method is introduced with the novelty of the evolutionary tree-based indexed quasi identification model to achieve the quasi identifiers for big healthcare data with lesser execution time. The new idea of the information loss function is employed independently to map among sample sets and attribute values for categorical and numerical attributes. The generalization and suppression process is carried for numerical and categorical attributes through information loss function, therefore minimizing the communication overhead. Next, an anonymization process is carried to reduce the communication overhead via quasi identifier identification.
- In order to learn the features for preserving the privacy of big healthcare data via identified quasi identifiers, the ETQI-FD method is introduced with the novelty of the federated adaptive Lorentz privacy preservation algorithm. First, the new idea of gradient descent function is used for performing the linear regression to determine the quasi identifiers with all patients. Second, the new idea of the APA function is utilized for improving the accuracy. Third, with the new idea of injecting adaptive Lorentz distribution (ALD), preservation of privacy is done based on the threshold value. This helps to improve the accuracy and privacy of each patient big healthcare data.

The rest of the paper is organized as follows. Section 2 presents the research method. The proposed ETQI-FD method is described in section 3. Section 4 analyses the performance and discussion of the proposed method. Finally, the conclusion is presented in Section 5.

2. RESEARCH METHOD

Privacy preservation is essential when certain user's data are provided to a third party for processing of any other distinct objective. With the upsurge of big data, in recent years, several health and medical institutions have obtained huge medical data. As a result, safeguarding the private big healthcare data of an individual becomes a paramount research topic. In this work, we have proposed a method called ETQI-FD for privacy preservations over big healthcare data. Figure 1 shows the block diagram of ETQI-FD method. As shown in Figure 1, we first formulate the ETQI model in detail. Based on the quasi identifiers learnt, privacy preservation mechanism using federated gradient arbitrary learning model is proposed. The elaborate description of the ETQI-FD method is given as follows.

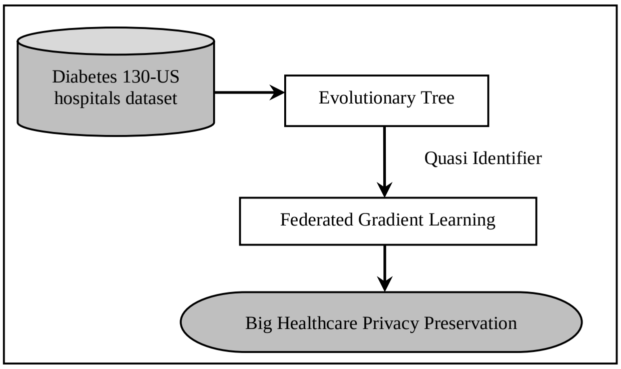
Diabetes 130-US hospitals dataset
Evolutionary Tree
Quasi Identifier
Federated Gradient Learning
Big Healthcare Privacy Preservation

Figure 1. Block diagram of evolutionary tree-based quasi identifier and federated gradient

2.1. Evolutionary tree-based indexed quasi identification model

In this section, we sketch out our evolutionary tree-based indexed quasi identification model to learn the quasi identifiers with minimum execution time required for anonymization. At first, large volume Diabetes 130-US hospitals dataset is considered as input. Figure 2 shows the block diagram of the quasi identification process.

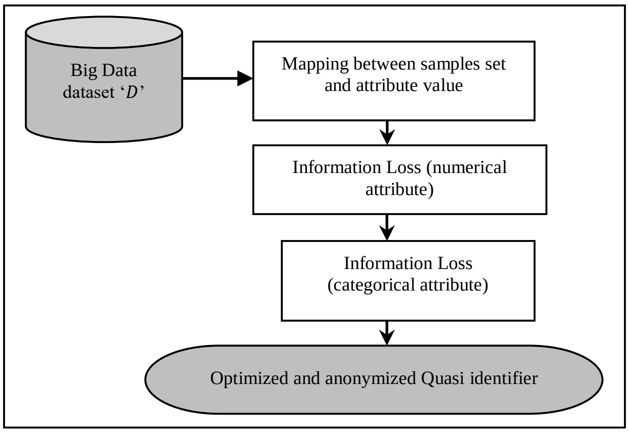
Big Data dataset 'D'
Mapping between samples set and attribute value
Information Loss (numerical attribute)
Information Loss (categorical attribute)
Optimized and anonymized Quasi identifier

Figure 2. Block diagram of evolutionary tree-based indexed quasi identification process

Figure 2 shows the evolutionary tree-based indexed quasi identification model. Let us consider the big dataset as input. Information loss function is used separately to map between sample sets and attribute value for categorical and numerical attributes. As given in the Figure 2, let us consider a Table 1 with columns ‘$a_1, a, \dots, a_m$’ and rows named ‘$r_1, r_2, \dots, r_n$’ with the column’s name representing the feature or attribute and row representing the instance or record. Now let us denote by $'A'$ the set of attributes and $'s'$ the set of samples. To each sample $'s'$ corresponds a tuple ‘$(v_1, v_2, \dots, v_n)$’, where ‘v_i’ refers to the value of the attribute ‘a_i’ for the underlying sample. Let ‘V_i’ denote the set of all values of the feature or attribute $'A'$. Then, a function $'f'$ is defined that maps $'s'$ to $'V'$ via the equation as in (1).

$$f(s) = [a_1(s), a_2(s), \dots\dots, a_n(s)] \qquad (1)$$

From the (1), the function $'f'$, for the corresponding sample $'s'$ is derived based on the attribute value of the respective sample ‘$a_1(s)$’. With the above function, a four tuple formation for our work is represented as ‘(S, A, V, f)’. The information is specified as a Table 1 given for diabetes 130-US hospitals dataset, where $'V'$ and $'f'$ are discarded, by considering only ‘(S, A)’.

Table 1. Example of diabetes 130-US hospitals dataset

Patient Number	Race	Age	Gender	Time in Hospital	Hba1c
1	African American	15	Male	11	None
2	Other	25	Male	13	>7
3	African American	40	Female	21	>7
4	Other	65	Female	22	>6
5	African American	60	Male	16	None

From the Table 1 set of examples, ‘$S = \{1, 2, 3, 4, \dots, 50\}$’ and the attribute set is ‘$A = \{\text{Patient number, race, age, gender, time in hospital, Hba1c}, \dots.\}$’, then ‘$V_{\text{Patient number}} = 1, 2, 3, 4, 5$’, ‘$V_{\text{race}} = \text{AfricanAmerican, Other, AfricanAmerican, Other, AfricanAmerican}$’, ‘$V_{\text{Age}} = 15, 25, 40, 65, 50$’, ‘$V_{\text{Gender}} = \text{Male, Male, Female, Female, Male}$’, ‘$V_{\text{Time in hospital}} = 11, 13, 21, 22, 16$’, ‘$V_{\text{Hba1c}} = \text{None}, > 7, > 7, > 6, \text{None}$. Then, for the first record when the attribute is ‘Patient number’ the value of ‘f’ is ‘$[f(1) = \text{AfricanAmerican}, 15, \text{Male}, 11, \text{None}]$’, therefore ‘$\text{Race}(1) = \text{AfricanAmerican}$’, ‘$\text{Age}(1) = 15$’, ‘$\text{Gender}(1) = \text{Male}$’, ‘$\text{Time in hospital } (1) = 11$’, ‘$\text{Hba1c}(1) = \text{None}$’. Let ‘ET’ represents the evolutionary tree as illustrated in figure for ‘Time in hospital’.

The evolutionary tree given in the Figure 3 is utilized in our work is used to generalize the value of each categorical and numerical attribute. An ‘$\gamma - \text{dissimilar}$’ quasi identifier is a subset of attributes that becomes a pivotal element when at most a ratio ‘$1 - \gamma$’ of samples is discarded. Moreover, a subset of attributes partitions two samples ‘s_1’ and ‘s_2’ if ‘s_1’ and ‘s_2’ have different values of at least one attribute of that subset. A ‘$\gamma - \text{separation}$’ quasi identifier is a subset of attributes that separates at least a ratio ‘γ’ of all probable instance pairs.

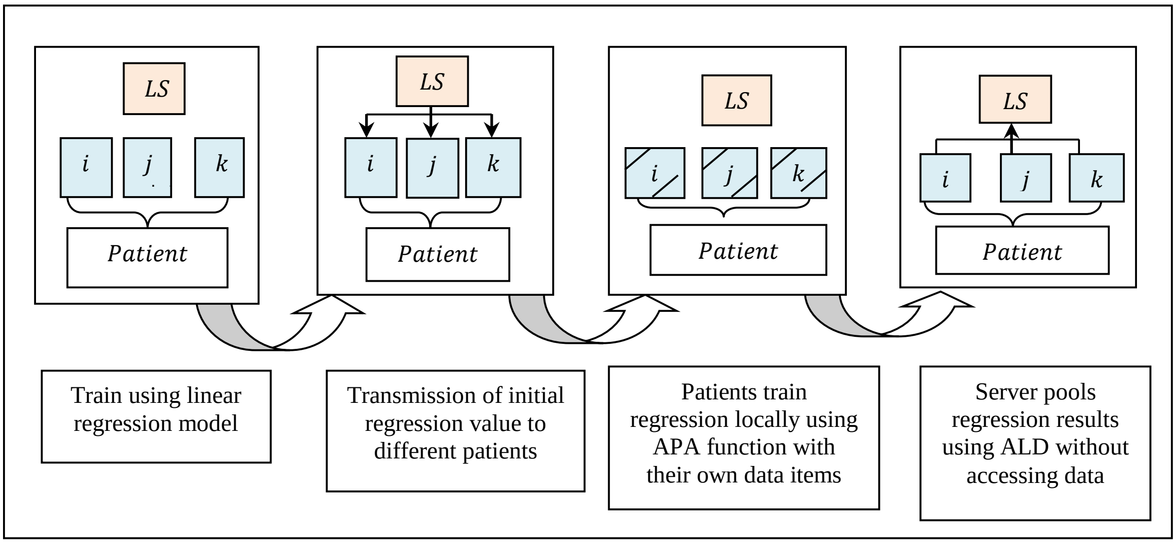
LS
i
j
k
Patient
LS
i
j
k
Patient
LS
i
j
k
Patient
LS
i
j
k
Patient
Train using linear regression model
Transmission of initial regression value to different patients
Patients train regression locally using APA function with their own data items
Server pools regression results using ALD without accessing data

Figure 3. Evolutionary tree of time in hospital

Let '$\gamma = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$' where '$\alpha$' symbolizes a group. Then, the information loss '$\mathrm{IL_{num}}$' for the numerical attribute representation of a given dataset 'D' through generalization and suppression is mathematically evaluated as in (2).

$$\mathrm{IL_{num}} = |\gamma| \sum_{i=1}^{m} \frac{\mathrm{Gi_h} - \mathrm{Gi_l}}{\mathrm{Ti_h} - \mathrm{Ti_l}} \tag{2}$$

From (2), the information loss for numerical attributes '$\mathrm{IL_{num}}$' is evaluated based on the highest '$\mathrm{Gi_h}$' and least value '$\mathrm{Gi_l}$' of the tuples in group, highest '$\mathrm{Ti_h}$' and least '$\mathrm{Ti_l}$' values of the tuples in dataset 'D' respectively. In a similar manner, the information loss for categorical attributes '$\mathrm{IL_{cat}}$' is evaluated as shown in (3) based on the height of the column values '$\mathrm{H}(\mathrm{c_j})$' and the height of the evolutionary tree of the column '$\mathrm{H}\left(\mathrm{ET}(\mathrm{c_j})\right)$' respectively.

$$\mathrm{IL_{cat}} = |\gamma| \sum_{j=1}^{n} \frac{\mathrm{H}(\mathrm{c_j})}{\mathrm{H}\left(\mathrm{ET}(\mathrm{c_j})\right)} \tag{3}$$

Finally, the quasi identifiers based on the resultant values of information loss arrived via numerical attributes '$\mathrm{IL_{num}}$', information loss arrived via categorical attributes '$\mathrm{IL_{cat}}$' is obtained as (4).

$$\mathrm{QID} = \ a \ \in \mathrm{Index}\ [\mathrm{A}|\gamma_{\mathrm{A}}(\mathrm{IL_{num}}) - \gamma_{\mathrm{A}}(\mathrm{IL_{cat}})] \tag{4}$$

The pseudo code representation of evolutionary tree-based indexed quasi identification is given as follows.

Algorithm 1. Evolutionary tree-based indexed quasi identification

```
Input: patients '$P = P_1, P_2, \dots, P_n$', big data dataset 'DS', attributes '$a_1, a, \dots, a_m$',
Output: optimized and anonymized quasi identifiers
Begin
For each big data dataset 'DS' with 'n' attributes '$A = a_1, a_2, \dots, a_n$' and Patients 'P'
    For each function 'f' defined that maps 'S' to 'V' as given in (1)
        Evaluate information loss for numerical attributes using (2)
        Evaluate information loss for categorical attributes using (3)
        Estimate quasi identifiers using (4)
    End for
End for
End
```

As given in the evolutionary tree-based indexed quasi identification algorithm the objective remains in learning the quasi identifiers in a timely manner with minimum communication overhead incurring while maintaining the links between evolutionary tree. This is achieved by first performing a mapping function based on two tuples. Next, with the generalization and suppression process performed via information loss function, therefore contributing to execution time involved in anonymization process.

Finally, with the aid of index function to the attribute information loss, the numerical and categorical attributes are not merged into single data node to minimize communication overhead, instead while one data node is locally merged while the rest are associated with the representative data node. In this manner, a significant amount of communication overhead is said to be reduced while performing anonymization process during the quasi identifier identification.

2.2. Federated gradient arbitrary privacy preservation learning model

Nowadays, patient's privacy is an analytical circumstance in big health care data. However, conventional machine learning techniques that purely depend on patient's log files and behavioral aspects are not adequate to preserve it. Hence, the health care data security should have numerous considerations to take into account supplementary information to safeguard patient's data.

In this work, federated learning is a privacy preservation algorithm is implemented that incorporates a collaborative learning model with centralized approach without the necessity of uploading the local dataset into one server. With this, robust machine learning is said to be ensured thus permitting to address issues such as data privacy.

Privacy-preserving federated machine learning process is specifically designed on the concept of Differential Privacy. Let us consider a dataset 'DS' quasi identifiers 'QI' about some set of 'n' patients are stored. This database can be queried by '$\mathrm{q_1, q_2, \dots, q_n}$' authorized users, among which there may be several malicious users trying to analyze data. Let ‘$q_1, q_2, \ldots, q_n$’ be their queries, ‘$ans_1, ans_2, \ldots, ans_n$’ representing the answers for these queries.

The main idea behind the design of differential privacy is to bestow with the answers to queries that it was impractical to differential the existence or non-existence of information. Thus, if there are two databases ‘DS’ and ‘DS'’, differing by only one record ‘$QI_1, QI_2, \ldots QI_3, \ldots QI_n$’, ‘$QI_1, QI_2, \ldots QI_{i'}, \ldots QI_n$’, then the probability distributions of ‘$Prob(DS)$’ and ‘$Prob(DS')$’ must be very close to each other. Figure 4 shows the block diagram of federated gradient arbitrary privacy preservation learning model.

As shown in the Figure 4, each patient ‘P_i’ owns ‘n’ data items ‘(P_i, Q_i)’, where ‘$i \in [1, n]$’. Each data item is first initialized with ‘m’ attributes and ‘n’ labels, i.e., ‘$p_{i1}, p_{i2}, \ldots, p_{in}, q_{i1}, q_{i2}, \ldots, q_{in}$’ based on linear regression. Then, to optimize the learning process, a gradient descent function is utilized and mathematically expressed as in (5).

$$\nabla g(D_i^n, w_i^n) = \partial IL_{num} IL_{cat}[Q_i, f(D_i^n, w_i^n)] \tag{5}$$

From (4), the gradient descent function ‘g()’ with set of data item ‘D_i^n’, weight matrix of patient ‘i’ after ‘n’ iteration is obtained based on the derivative loss function of numerical attribute ‘IL_{num}’, categorical attribute ‘IL_{cat}’, extent of variability between the predicted value ‘$f(D_i^n, w_i^n)$’ and the actual value ‘Q_i’ respectively. Next, with the fraction of local nodes (patients) selected to undergo training configuration is performed by upgrading weight and is formulated as in (6).

$$w_i^{n+1} = w_i^n - \eta_i \nabla g(D_i^n, w_i^n) \tag{6}$$

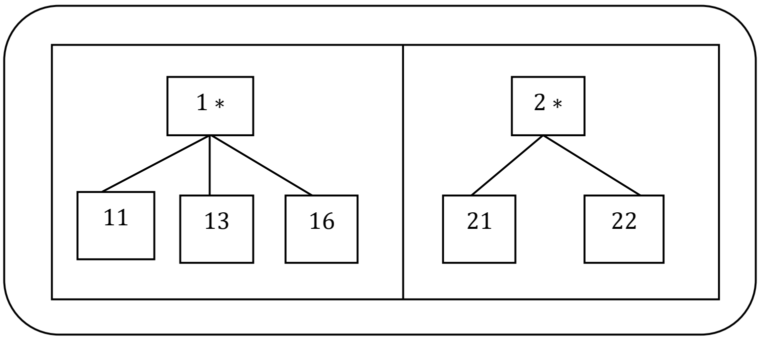

Figure 4. Block diagram of federated gradient arbitrary privacy preservation learning model

From (6), the weight upgrade ‘w_i^{n+1}’ for each patient ‘P_i’ is arrived at based on the existing weight ‘w_i^n’ and a learning factor ‘η_i’. Next, with the configured resultant value, an arbitrary privacy-preserving adaptation (APA) function is utilized with the objective of improving the accuracy of big healthcare data. For this, the determinant of attribute is evaluated. By extracting the determinant of the same attribute or feature from tuple, the mean determinant of every attribute or feature ‘$D_j(p_i)$’ to the output is evaluated as in (7).

$$D_j(p_i) = \frac{1}{n}\sum_{i=1}^{n}(p_{ij}[p_i]), j \ \in [1, p] \tag{7}$$

With (7) determinant value obtained, two adaptation components ‘c_1’ and ‘c_2’ are introduced, where ‘$c_1 \epsilon [0,1]$’ and ‘$c_2 \epsilon [0,1]$’ respectively, where ‘c_1’ denotes a threshold whether the attribute to the output is high or low and its value is defined by patients. In other words, if mean determinant of every attribute exceeds the threshold ‘c_1’ possess greater determination to output. Then, we inject ALD or Lorentz noise to all these attributes. On the other hand, while the determinant value obtained is lesser than the threshold, ‘c_1’, original data with probability ‘$1 - prob$’ is selected and to inject adaptive Lorentz Distribution to some attributes with probability ‘$prob$’. This is mathematically expressed as in (8).

$$Rp_{ij} = \begin{cases} p_{ij}\ \alpha \ \geq \ c_1 \\ p'_{ij}\alpha \ < c_1 \end{cases} \tag{8}$$

From (8), ‘α’ refers to the ratio of determinant results ‘$\alpha = \frac{|D_j|}{\sum_{j=1}^{n}|D_j|}$’. However, when ‘$\alpha < c_1$’, the mathematical representation is formulated as in (9).

$$p'_{ij} = \begin{cases} p_{ij} \text{ with probability prob} \\ p_{ij} \text{with probability } 1 - \text{prob} \end{cases} \tag{9}$$

With the obtained probability measures, a small amount of adaptive Lorentz distribution is injected into the attributes as in (10).

$$p'_{ij} = p_{ij} + \frac{1}{|D_i^n|} f[p; p_0, \gamma] \tag{10}$$

From (10), ‘$f[p; p_0, \gamma]$’ represent the ALD or Lorentz noise which is estimated as in (11).

$$f(p, p_0, \gamma) = \frac{1}{\pi\gamma \left[1 + \left(\frac{p - p_0}{\gamma}\right)^2\right]} = \frac{1}{\pi\gamma}\left[\frac{\gamma^2}{(p - p_0) + \gamma^2}\right] \tag{11}$$

From (11), ‘p_0’ specifies the input parameter of consideration, ‘γ’ represents the scaling factor with a maximum value being specified as ‘$\frac{1}{\pi\gamma}$’, for each patient ‘$p = p_0$’ respectively. The pseudo code representation of federated adaptive Lorentz privacy preservation is given as follows.

Algorithm 2. Federated adaptive Lorentz privacy preservation

```
Input: patients '$P = P_1, P_2, \dots, P_n$', big data dataset 'DS', attributes '$a_1, a, \dots, a_{m-}$', quasi
identifiers '$QI = QI_1, QI_2, \dots QI_3, \dots QI_n$'
Output: accurate and adaptive privacy preserved identifiers
Initialize fraction of local nodes or patients '$p_{i1}, p_{i2}, \dots, p_{in}, q_{i1}, q_{i2}, \dots, q_{in}$'
Initialize '$c_1$' and '$c_2$'
Begin
For each big data dataset 'DS' with attributes 'A'
For each Quasi Identifiers '$QI = QI_1, QI_2, \dots QI_3, \dots QI_n$'
Perform linear regression based on gradient descent function using (5)
//configuration
Upgrade weight using (6) for each patient '$P_i$'
Evaluate mean determinant of every attribute using (7)
Estimate adaptive Lorentz Distribution using (8), (9) and (10)
Return (privacy preserved data items)
End for
End for
End
```

As given in the federated adaptive Lorentz privacy preservation algorithm, the objective remains in accurate and adaptive privacy preserved identifies with higher accuracy. First linear regression based on gradient descent function is evolved for each patient with the obtained quasi identifiers. Next, for each patient, weight is upgraded by utilizing APA function therefore contributing to accuracy. Finally, preservation of privacy for each quasi identifier is made by injecting adaptive Lorentz distribution according to the threshold value. With this, the accuracy and privacy are said to be ensured for each patient big healthcare data.

3. EXPERIMENTAL SETTINGS

In this section, a detailed analysis of experimental results has been presented to evaluate the performance of ETQI-FD for privacy-preserving of big healthcare data via quasi-identifier. The efficiency of the ETQI-FD method is determined along with the metrics such as execution time, communicational overhead, accuracy, and information loss by using diabetes 130-US hospitals dataset. Using this dataset, privacy preserving experiments are conducted via Python**.** The implementation is conducted with the hardware specification of Windows 10 Operating system, core i3-4130 3.40 GHZ processor, 4 GB RAM, 1 TB (1000 GB) hard disk, ASUSTek P5G41C-M motherboard, internet protocol. For accomplishing the experimental evaluation, the ETQI-FD considers a number of patient data in the range of 500-5000 from the diabetes 130-US hospitals.

4. RESULTS AND DISCUSSION

4.1. Analysis of communicational overhead

The communicational overhead refers to the overhead incurred during maintenance of links while designing evolutionary tree. This is mathematically expressed as in (12).

$$CO = \sum_{i=1}^{n} P_i * MEM\ [QID] \tag{12}$$

From (12), the communicational overhead 'CO' is measured based on the number of patients involved in simulation process 'P_i' and the memory consumed during the identification of quasi identifier (MEM-QID). It is measured in terms of kilobytes (KB). Results of paired tests for comparing the communication overhead until a migration produced by the algorithms with privacy preservation tasks. The results of the comparison of the communication overhead for the proposed method ETQI-FD and existing attribute centric anonymization [1], robust anonymization and risk assessment [2] are graphically depicted in Figure 5.

The experimental results on the communication overhead on the diabetes 130-US hospitals dataset is depicted in Figure 5. To conduct our experiments, the number of patients provided as input was selected in the range of 500 to 5000. However, with a simulation involving '500' patients and the communication link established while designing evolutionary tree for identification of quasi identifier being '2 KB' using ETQI-FD, '3 KB' using [1] and '4 KB' using [2], the overall communication overhead was observed to be 1000 KB, 1500 KB and 2000 KB respectively. The reason behind the improvement is owing to the application of evolutionary tree-based indexed quasi identification algorithm in proposed ETQI-FD. Therefore, the communication overhead involved during privacy preservation is said to be reduced. The average communication overhead result of ETQI-FD is reduced by 32% when compared to [1] and 49% compared to [2].

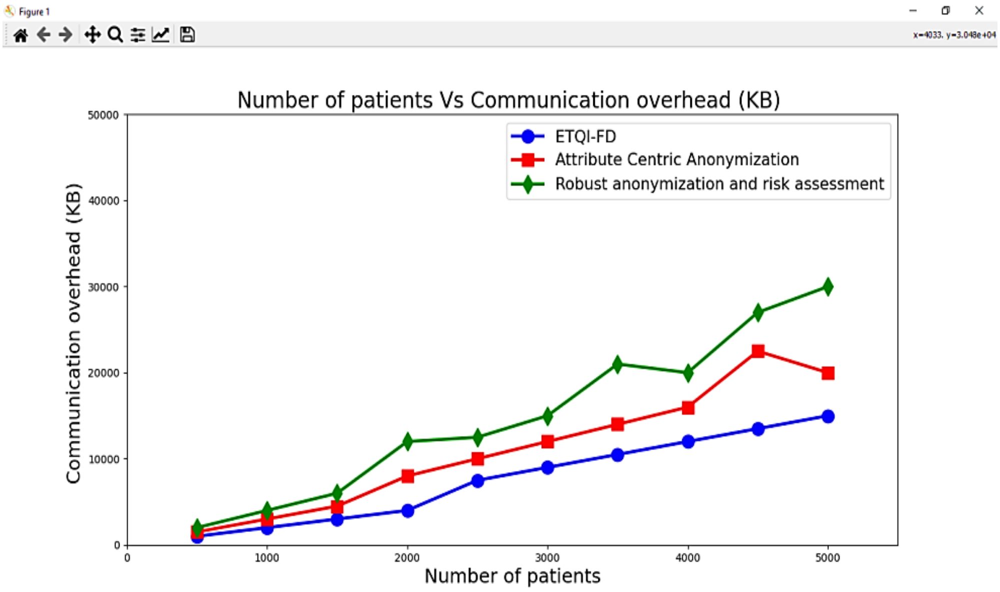

Figure 5. Comparison of ETQI-FD, attribute centric anonymization [1], robust anonymization and risk assessment [2] with respect to communication overhead

4.2 Analysis of accuracy

In this section, privacy preservation accuracy in our work refers to the accuracy maintained during quasi identifier identification and also the privacy preserved in big healthcare data. This is mathematically expressed as (13).

$$A = \sum_{i=1}^{n} \frac{P_{AP}}{P_i} * 100 \tag{13}$$

From (13), the accuracy 'A' is measured based on the percentage ratio of accurate identification of quasi identifier and privacy preservation in big healthcare data 'P_{AP}' to the number of patients involved during simulation process 'P_i'. It is measured in terms of percentage (%). The results of the comparison of the accuracy factor for proposed method ETQI-FD and existing attribute centric anonymization [1], robust anonymization and risk assessment [2] are graphically depicted in Figure 6.

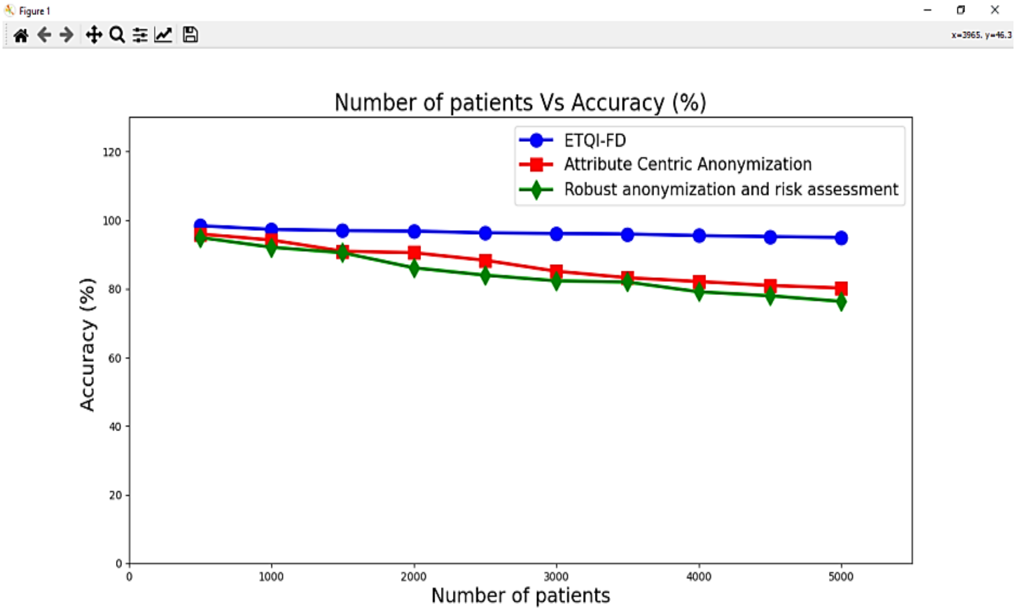

Figure 6. Comparison of ETQI-FD, attribute centric anonymization [1], robust anonymization and risk assessment [2] with respect to accuracy

Followed by the experimental results of communication overhead on the diabetes 130-US hospitals dataset, the accuracy rate is depicted in Figure 6. The experiments conducted to estimate the accuracy were obtained in the range of 500 to 5000. Let us consider '1000' patients data taken from the dataset for conducting the experiments. By applying the ETQI-FD, '487'patient's data are correctly recognized hence the accuracy is 97.35%. Whereas '471'and '461'patient's data are correctly detected by using [1], [2] and their accuracy percentages are 94.25% and 92.15% respectively. This is owing to the implementation of information loss detection function separately for numerical and categorical attributes via generalization and suppression. By applying this function separately, for each feature, a separate evolutionary tree was constructed. Followed by quasi identifier were identified to preserve the privacy for improving the accuracy. The average comparison results demonstrate that the accuracy of the proposed ETQI-FD is considerably improved by 11% and 14% during privacy preservation when compared to existing [1] and [2] respectively.

4.3. Analysis of information loss

The information loss is referred to as the amount of loss incurred during privacy preservation. This is mathematically estimated as (14).

$$IL = \sum_{i=1}^{n} \frac{P_{dl}}{P_i} * 100 \qquad (14)$$

From (14), information loss 'IL' is measured on the basis of the patients involved in simulation during privacy preservation 'P_i' and the amount of patient data lost 'P_{dl}'. It is measured in terms of percentage (%). The results of information loss comparing the performance of ETQI-FD with existing attribute centric anonymization [1] and robust anonymization and risk assessment [2].

Figure 7 illustrates the variation in information loss for different numbers of patients obtained at different time intervals. However, with a simulation involving '500' patients and the amount of patient data lost being '12'using ETQI-FD, '17' using [1] and '30' using [2], the overall information loss was observed to be 2.4%, 3.4% and 6% respectively. The reason behind the minimization of information loss using ETQI-

FD is due to the application of federated adaptive Lorentz privacy preservation algorithm. This is helps to reduce the information loss. Finally, average of ten results indicates that the information loss is consderably minimized by 32% and 60% when compared to existing [1], [2] respectively.

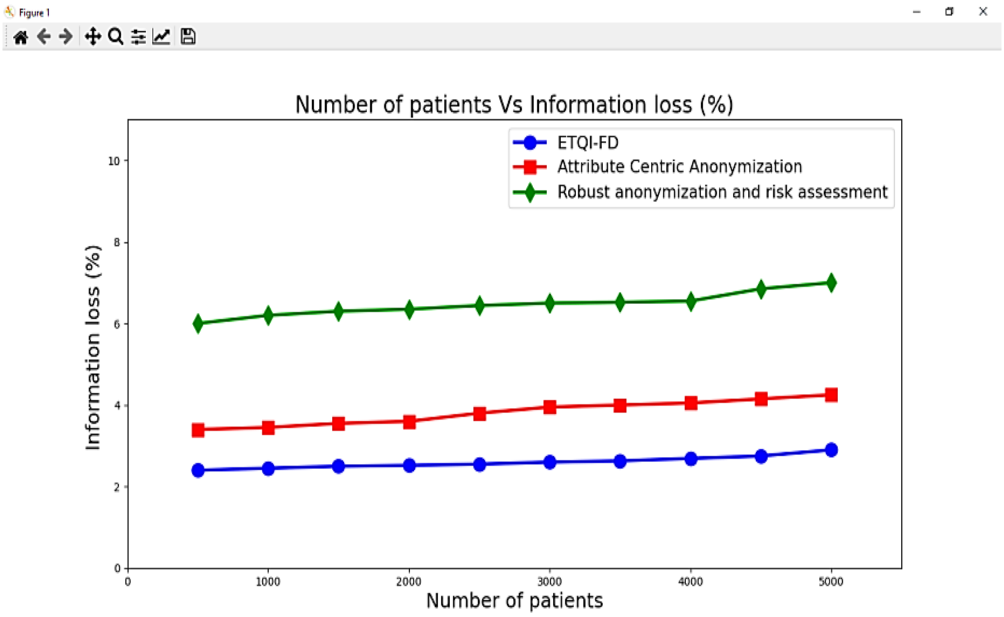

Figure 7. Comparison of ETQI-FD, attribute centric anonymization [1], robust anonymization and risk assessment [2] with respect to information loss

5. CONCLUSION

A machine learning privacy preservation method has been proposed for big healthcare data for privacy preservation of healthcare data in case of a high level of anonymization. Evolutionary tree-based indexed quasi identification model is introduced to map between sample sets and attribute values according to numerical and categorical attributes separately. Also, it integrates the privacy preservation model with the federated learning, therefore injecting noise based on a threshold value. It takes advantage of the existing features and generates only a few attributes as quasi identifiers. The proposed method is compared with the two existing methods (attribute centric anonymization, robust anonymization and risk assessment). The proposed ETQI-FD method achieves better privacy and accuracy. The proposed work is further suggested to use new research work with the implementation of cryptographic techniques for privacy preserving.

REFERENCES

[1] A. Majeed, “Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data,” in *Journal of King Saud University-Computer and Information Sciences*, vol. 31, no. 4, pp. 426-435 Jul. 2019, doi: 10.1016/j/jksuci.2018.03.014.

[2] F. Prasser, H. Spengler, R. Bild, J. Eicher, and K. A. Kuhn, “Privacy-enhancing ETL-processes for biomedical data,” in *International Journal of Medical Informatics*, vol. 126, pp. 72-81, 2019, doi: 10.1016/j.ijmedinf.2019.03.006.

[3] K. Arava and S. Lingamgunta, “Adaptive k-anonymity approach for privacy preserving in cloud,” in *Arabian Journal for Science and Engineering*, *Springer*, vol. 45, pp. 2425-2432, 2020, doi: 10.1007/s13369-019-03999-0.

[4] N. Nnamoko and I. Korkontzelos, “Efficient treatment of outliers and class imbalance for diabetes prediction,” in *Artificial Intelligence in Medicine*, vol. 104, pp. 1-12, Apr. 2020, doi: 10.1016/j.artmed.2020.101815.

[5] C. W. Soh, L. L. Njilla, K. K. Kwiat, and C. A. Kamhoua, “Learning quasi-identifiers for privacy-preserving exchanges: A rough set theory approach,” in *Granular Computing*, vol. 5, pp. 71-84, 2018, doi: 10.1007/s41066-018-1027-0.

[6] R. Mendes and J. P. Vilela, “Privacy-preserving data mining: Methods, metrics, and applications,” in *IEEE Access*, vol. 5, pp. 10562-10582, 2017, doi: 10.1109/ACCESS.2017.2706947.

[7] O. Kwabena, Z. Qin, T. Zhuang, and Z. Qin, “MSCryptoNet: Multi-scheme privacy-preserving deep learning in cloud computing,” in *IEEE Access*, vol. 7, pp. 29344-29354, 2019, doi: 10.1109/ACCESS.2019.2901219.

[8] K. Abouelmehdi, A. B. Hessane, and H. Khaloufi, “Big healthcare data: preserving security and privacy,” in *Journal of Big Data*, vol. 5, no. 1, pp 1-18, 2018, doi: 10.1186/s40537-017-0110-7.

[9] N. K. Parachurcotha and M. Sokolova, “Multi-label learning in classification of patients’ quasi-identifiers,” *Progress in Artificial Intelligence*, vol. 4, pp. 37-48, 2015, doi: 10.1007/s13748-015-0064-y.

[10] M. Binjubeir, A. A. Ahmed, M. A. B. Ismail, A. S. Sadiq, and M. Khurram Khan, "Comprehensive survey on big data privacy protection," in *IEEE Access*, vol. 8, pp. 20067-20079, 2020, doi: 10.1109/ACCESS.2019.2962368.

[11] P. Li, T. Li, H. Ye, J. Li, X. Chen, and Y. Xiang, "Privacy-preserving machine learning with multiple data providers," in *Future Generation Computer Systems*, vol. 87, pp 341-350, 2018, doi: 10.1016/j.future.2018.04.076.

[12] P. S. Rao and S. Satyanarayana, "Privacy preserving data publishing based on sensitivity in context of big data using Hive," in *Journal of Big Data*, vol. 5, no. 20, 2018, doi: 10.1186/s40537-018-0130-y.

[13] Q. Zhang *et al.*, "Deep learning with attention supervision for automated motion artifact detection in quality control of cardiac T1-mapping," in *Artificial Intelligence in Medicine, Elsevier*, vol. 110, p. 101955, 2020, doi: 10.1016/j.artmed.2020.101955.

[14] S. M. Lauritsen *et al.*, "Early detection of sepsis utilizing deep learning on electronic health record event sequences," in *Artificial Intelligence in Medicine*, vol. 104, pp. 1-11, 2020, doi: 10.1016/j.artmed.2020.101820.

[15] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, "Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness," in *IEEE Transactions on Dependable and Secure Computing*, vol. 16, no. 4, pp. 580-593, 2019, doi: 10.1109/TDSC.2017.2698472.

[16] P. R. M. Rao, S. M. Krishna, and A. P. S. Kumar, "Privacy preservation techniques in big data analytics: a survey," in *Journal of Big Data, Springer*, vol. 5, no. 33, pp. 1-12, 2018, doi: 10.1186/s40537-018-0141-8.

[17] O. Temuujin, J. Ahn, and D. Im, "Efficient l-diversity algorithm for preserving privacy of dynamically published datasets," in *IEEE Access*, vol. 7, pp. 122878-122888, 2019, doi: 10.1109/ACCESS.2019.2936301.

[18] F. Song, T. Ma, Y. Tian, and M. Al-Rodhaan, "A new method of privacy protection: random k-anonymous," in *IEEE Access*, vol. 7, pp. 75434-75445, 2019, doi: 10.1109/ACCESS.2019.2919165.

[19] C. Liu, S. Chen, S. Zhou, J. Guan, and Y. Ma, "A novel privacy preserving method for data publication," in *Information Sciences*, vol. 501, pp. 421-435, 2019, doi: 10.1016/j.ins.2019.06.022.

[20] Y. Saleem, M. H. Rehmani, N. Crespi, and R. Minerva, "Parking recommender system privacy preservation through anonymization and differential privacy," in *Engineering Reports*, vol. 3, no. 2, pp. 1-30, 2020, doi: 10.1002/eng2.12297.

[21] S. M. Ibrahim, N. A. Bakar, M. Mamat, B. A. Hassan, M. Malik, and A. M. Ahmed, "A new hybrid conjugate gradient algorithm for optimization models and its application to regression analysis" *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, volume 23, no. 2, pp. 1100-1109, Apr. 2021, doi: 10.11591/ijeecs.v23.i2.pp1100-1109.

[22] A. Triayudi, W. O. Widyarto, L. Kamelia, I. Iksal, and S. Sumiati, "CLG clustering for dropout prediction using log-data clustering method," in *IAES Internatinal Journal of Artificial Intelligence (IJ-AL)*, vol. 10, no. 3, pp. 764-770, 2021, doi: 10.11591/ijai.v10.i3.pp764-770.

[23] M. S. Farahani, S. H. R. Hajiagha, "Forecasting stock price using integrated artificial neural network and metaheuristic algorithms compared to time series models" in *Soft Computong*, vol. 25, pp. 8483–8513, 2021, doi: 10.1007/s00500-021-05775-5.

[24] L. Sumaryanti, D. H. Kusuma, R. Widijastuti, and M. N. Muzaki, "Improvement security in e-business systems using hybrid algorithm," in *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 19, no. 5, 2021, doi: 10.12928/telkomnika.v19i5.20403.

[25] B. T. Ahmed, "Data mining techniques for lung and breast cancer diagnosis: A review," in *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 10, no. 2, 2021, doi: 10.11591/ijict.v10i2.pp93-103.

[26] T. M. Escobar, L. C. Reyes, C. M. Trejo, C. G. Santillán, N. R. Valdez and H. F. Huacuja, "An interactive recommendation system for decision making based on the characterization of cognitive tasks" in *Math. Comput. Appl.*, vol. 26, no. 2, 2021, Art. no. 35, doi: 10.3390/mca26020035.

BIOGRAPHIES OF AUTHORS

Sujatha Krishna received the B.E. and M. Tech degrees in Computer Science and Engineering from Visvesvaraya Technological University, Karnataka, India. She is currently pursuing the Ph.D. degree in the Computer Science and Engineering from REVA University, Karnataka, India. Her research interests include big data, data mining, machine learning and privacy preserving algorithms. She can be contacted at email: sujathasjcit@gmail.com.

Udayarani Vinayaka Murthy received her Ph.D. degree in Computer Science from Mother Teresa University, Kodaikanal, Tamil Nadu, India, in 2014. She is currently a Senior Associate Professor at REVA University, Bangalore, India. Her research interest includes data mining, machine learning, big data analytics, and genetic algorithms. She can be contacted at email: udayarani.v@reva.edu.in.