Distinctive Context Sensitive and Hellinger Convolutional Learning for Privacy Preserving of Big Healthcare Data

Sujatha K[1]
Research Scholar
School of computing and IT
REVA University, Bangalore, India

Udayarani V[2]
Senior Associate Professor
School of computing and IT
REVA University, Bangalore, India

***Abstract*—The collection and effectiveness of sensitive Big Data have grown with Information Technology (IT) development. While using sensitive Big Data to acquire relevant information, it becomes indispensable that irrelevant sensitive data are reduced to safeguard personal information in healthcare sector. Many privacy-preserving strategies have been applied in the recent years using quasi-identifiers (QI) for applications like health services. However, privacy preservation over quasi-identifiers is still challenging in the context of Big Data because most datasets were of huge volume. Existing methods suffer from higher time consumption and lower data utility because of dynamically progressing datasets. In this paper, an efficient Distinctive Context Sensitive and Hellinger Convolutional Learning (DCS-HCL) is introduced to ensure privacy preservation and achieve high data utility for big healthcare datasets. First, Distinctive Impact Context Sensitive Hashing model is designed for the given input Big Dataset where both the distinctive and impact values are identified and applied to Context Sensitive Hashing. With this, similar QI-classes are mapped to evolve the computationally efficient anonymyzed data. Second, Hellinger Convolutional Neural Privacy Preservation model is presented to preserve the privacy of the sensitive unstructured data. This is performed by hashing QI-class values, weight updation and bias in CNN to increase the accuracy and to reduce the information loss. Evaluation results demonstrate that with proposed method with large-volume unstructured datasets improved performance of run time, data utility, information loss and accuracy significantly over existing methods.**

Keywords*—Big data; information technology; distinctive; impact; context sensitive hashing; quasi-identifier; Hellinger; convolutional neural

I. INTRODUCTION

Privacy-preservation issues have made an appearance with the growing magnitudes of data being issued together with sensitive, private information pertaining to individual persons and also business establishments. To address such issues, several strategies of minimizing risk connected with data being published have been designed. One of the remedies is protecting sensitive data via quasi identifier. Equivalence Classes with Cuckoo Filter (ENCC) [1] utilized anatomy alternative for suppression to design more effective l-diversity algorithm with the objective of preserving the privacy of those datasets. Moreover, a Cuckoo filter was utilized to approximate set-membership tests for enhancing the efficiency involved in data processing. With the application of l-diversity algorithm, the running time was found reduced than when compared to traditional re-anonymization techniques.

In addition, filter mechanism was used to maintain privacy of dynamically progressing datasets. Despite maintaining privacy and reducing the running time with the absence of strong data-anonymization models, data utility was not focused. To address this issue, in this work, Distinctive Impact Context Sensitive Hashing model is designed that evolves with computationally efficient quasi-identifiers with minimal time and higher data utility.

A novel privacy model utilizing integrated anonymization and reconstruction was proposed in [2] for making the strong assumption. The separation of quasi-identifiers (QIDs) was carried out from sensitive attributes. A sensitive QID using l-diversity and t-closeness was designed. It was in novel privacy model, anonymization and reconstruction was possible while maintaining the high quality of data within stipulated time period.

Though high data quality within stipulated time period was maintained, the accuracy and information loss was not concentrated. To address this issue in this work, Hellinger Convolutional Neural Privacy Preservation model is proposed to protect both the sensitive data by designing a significant privacy preservation model considering both the distance by means of Hellinger and improving the accuracy by updated weight and bias via convolutional neural learning.

A. Contributions

The main contributions of this paper to the literature are summarized as follows:

- A, Distinctive Context Sensitive and Hellinger Convolutional Learning (DCS-HCL) method is designed with the purpose of preserving the privacy of big healthcare data along with high data utility and minimum information loss.
- Distinctive Impact Context Sensitive Hashing model is developed for performing sensitive hashing to surpass the defined limitations and focuses on the run time and therefore improving the data utility.

- Hellinger Convolutional Neural Privacy Preservation model is a new privacy preservation model used for identifying quasi-identifiers to improve accuracy and reduce information loss.
- Privacy preservation methods are compared with the conventional privacy preservation ones. Experimental results demonstrated that proposed method showed comparatively better performance in terms of run time, accuracy and loss error.

B. Organization Structure

The organization of this paper is as follows. Section II reviews the development of privacy preservation techniques concerning big data. The details of the proposed method Distinctive Context Sensitive and Hellinger Convolutional Learning (DCS-HCL) is presented in Section III. The experimental analysis of the proposed method is discussed in Section IV. The result discussion with the other well-known privacy preservation methods is presented in Section V. Finally, the conclusion is given in Section VI.

II. Related Works

In recent years, the escalating issue of Internet phishing has been menacing the secure proliferation of sensitive data over the web, including several domains like healthcare data, video surveillance, Internet trafficking and so on. Therefore, privacy preservation has become a major challenge resulting in imprecise distribution of data.

A global survey on privacy preservation for big data was investigated in [3]. But, the information loss was not minimized. With big healthcare data to enhance patient outcomes, to predict pandemic outbreaks in early stage, keep away from avertable diseases, the security and privacy concerns were discussed in [4]. But, the runtime consumption was not minimized. An encryption algorithm using honey encryption algorithm was proposed in [5] to address the issues related to data security. However, the dimensionality issues were not minimized. A comprehensive focus was made for identical data types using quasi-identifiers called identical generalization hierarchy (IGH). An optimal solution was designed based on globally optimized k-anonymity [6] for minimizing the overall convergence time to a greater extent. But, accuracy level was not taken into consideration.

The privacy preservation in big data utilizing solution towards data warehousing was proposed in [7] using nearest similarity based clustering (NSB) with Bottom-up generalization. The susceptibility with respect to sensitivity was addressed and ensured privacy for user data. But, the computational cost was not minimized. A survey of privacy preservation techniques was investigated in [8]. However, security level was not improved. A review of privacy preservation for resource constrained sensors was proposed in [9]. However, attribute disclosure prevention were not met.

Two privacy models called enhanced identity-reserved diversity and enhanced identity-reserved anonymity were presented in [10] to minimize the error. Though the error was reduced, multiple sensitive attributes preservation remained unaddressed. To provide solution to this issue, bucketization principles were utilized in [11] for preserving the vulnerable records. But, the computational complexity was not minimized. A bidirectional personalized generalization model was designed in [12] for multi-record datasets. Through validating the quasi-identifier anonymity and ensuring diversity on equivalence groups, information loss was reduced to a large extent. However the privacy level gets varied for different users.

In [13], a privacy preservation model to prevent data loss using hash anomaly detection process was designed, therefore improving the data privacy along with the minimization of data portability cost. However, the time consumption was not minimized. In [14], local differential privacy was applied with the objective of providing significant accuracy. Though the accuracy level was improved, the computational cost was not minimized. In [15], a healthcare privacy preservation scheme called, Healthchain was designed on the basis of the blockchain technology. The healthcare data were initially encrypted for ensuring fine-grained access control. The users significantly had possibility of either revoking or including certain features for efficient key management. But, the runtime was not reduced for healthcare privacy preservation.

Tampering was avoided to keep away from contentions or altercations for ensuring both privacy and security. In [16], security and privacy issues concerning healthcare sector was surveyed and mechanisms were included in addressing the issues. The focus was specifically designed depending one anonymization and encryption. Moreover, the advantages and disadvantages of introducing the anonymization and encryption standards were also made. However, the accuracy level was not taken into consideration.

An in-depth concentration on privacy and security aspects in big data and differentiation between the privacy and security aspects in big data was presented in [17]. But, the information loss was not focused. A systematic approach was proposed in [18] for selecting the seed with the purpose of clustering the records by employing adaptive *k*-anonymity algorithm. But, privacy preservation performance was not improved considerably. Rough set approach was proposed in [19] to balance between quasi-identifier anonymity and sensitive attribute diversity. However, runtime performance was not at required level by designed approach.

A. Research Gap

As a part of information sharing information via internet, each business establishments print data that are considered to be highly sensitive or personal. In this advancing IT-era towards big data, user's privacy protection is becoming a major issue to be addressed. In the recent years, as prototype of medical services has transformed from therapy to safeguard, there arises the heightening interest in healthcare sector. Despite the data being valuable asset, serious privacy issue is said to occur with the leakage of sensitive information. These data have to be preserved. After reviewing the existing methods, there are still difficulties in data utility management and information loss.

In addition, the high information loss, high runtime consumption, high computational cost, high computational

complexity, less accuracy, less security and privacy were issues faced by the user's during data communication in healthcare sector. Therefore, Distinctive Context Sensitive and Hellinger Convolutional Learning (DCS-HCL) is introduced for support fine-grained access control with big healthcare data to ensure data utility with high accuracy and minimum information loss as well as runtime consumption.

III. METHODOLOGY

In this section, the quasi-identifier arrangement based x model and y model are formulated in detail. Section 'A' sketches out the system model. In Section 'B', Distinctive Impact Context Sensitive Hashing model is described for quasi-identifier detection from Big (unstructured) Data. Based on the established arrangements (i.e. detected quasi-identifier) via Quasi-Identifier Classes, Section 'C' elaborates the design and development of privacy preservation for unstructured data. Fig. 1 shows the block diagram of Distinctive Context Sensitive and Hellinger Convolutional Learning (DCS-HCL) method.

As shown in Fig. 1, large volume Big Data dataset of diabetic patients are provided as input. Attribute segregation is initially performed with the input Big Data dataset by means of Distinctive Impact Context Sensitive Hashing model. With this, unique QI-classes possessing unstructured data are mapped to detect the computationally efficient anonymous data (i.e., quasi attributes or quasi-identifiers) from Big Data.

A. System Model

Let us consider a big data dataset 'DS' extracted from Diabetes 130-US hospitals for years 1999-2008 Data Set [3] consisting of 50 different features or attributes '$Attr = a_1, a_2, \dots, a_n$' of '$n$' patients. Each attribute classifies the data columns 'C' into four different classes '$Cl = \{cl_1, cl_2, cl_3, cl_4\}$' referred to as quasi attributes '$Q = \{q_1, q_2, \dots, q_n\}$', external attributes '$E = \{e_1, e_2, \dots, e_n\}$', sensitive attributes '$S = \{s_1, s_2, \dots, s_n\}$' and non-sensitive attributes '$NS = \{ns_1, ns_2, \dots, ns_n\}$' respectively.

B. Distinctive Impact Context Sensitive Hashing Model

First quasi attributes are identified from Big Data using Distinctive Impact Context Sensitive Hashing (DI-CSH) model. Quasi attributes are attributes that reveal data of precise identifiers employing background knowledge. Several strategies have been presented by various research analysts to identify the quasi identifiers where resources are considered for executing privacy. However, these techniques are not free from limitations like higher time consumption and lower data utility. The proposed DI-CSH model controls the limitation by extracting base essential quasi attributes with minimum time complexity and higher data utility. In ENCC [1] method, anonymization has been applied on quasi identifiers to convert it into more diversified form, the privacy expanded to certain extent. But, the issue remains in identifying the optimal quasi attributes in big data dataset.

Many quasi attributes on one side decreases the data utility. On other hand, less quasi attributes results in privacy breach. The objective of Distinctive Impact Context Sensitive Hashing model is to identify the optimal quasi attributes in Big Data dataset in optimal time and complexity resulting in the improvement of performance in preserving the privacy with the optimal number of quasi attributes. Fig. 2 given above shows the sample format of Distinctive Impact Context Sensitive Hashing model.

As shown in the above Fig. 2, with the input diabetic dataset provided as input, the objective of designing Distinctive Impact Context Sensitive Hashing model remains in extracting the quasi-attributes with minimum time complexity and high data utility. The distinctive value 'DV' is evaluated based on the number of distinct values 'DV' in column 'C_i' and the total number of different values in column 'TV' respectively. The distinctive value is expressed as given below.

$$DV = \frac{\sum_{i=1}^{n} DV[C_i]}{TV} \quad (1)$$

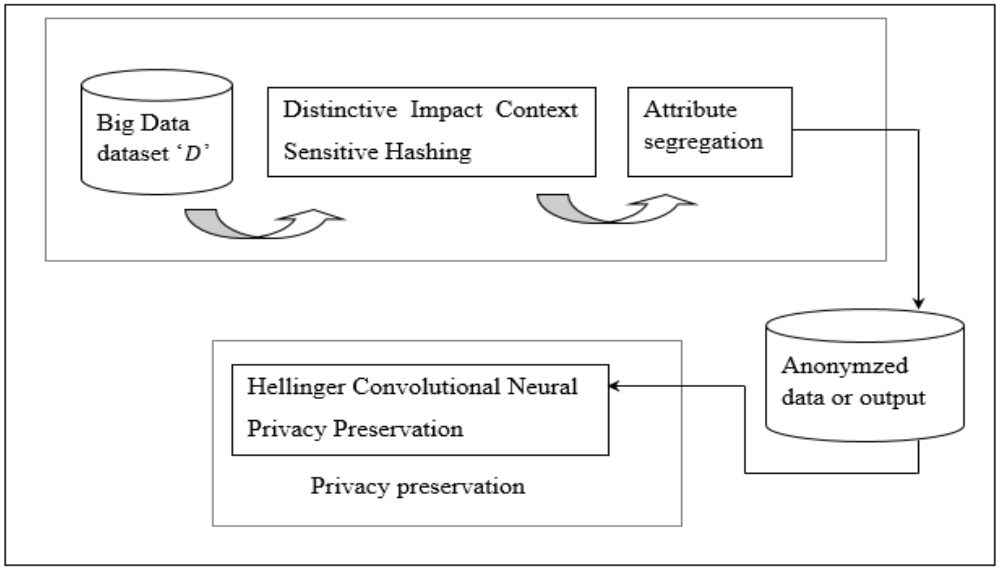

Fig. 1. Block Diagram of Distinctive Context Sensitive and Hellinger Convolutional Learning Method.

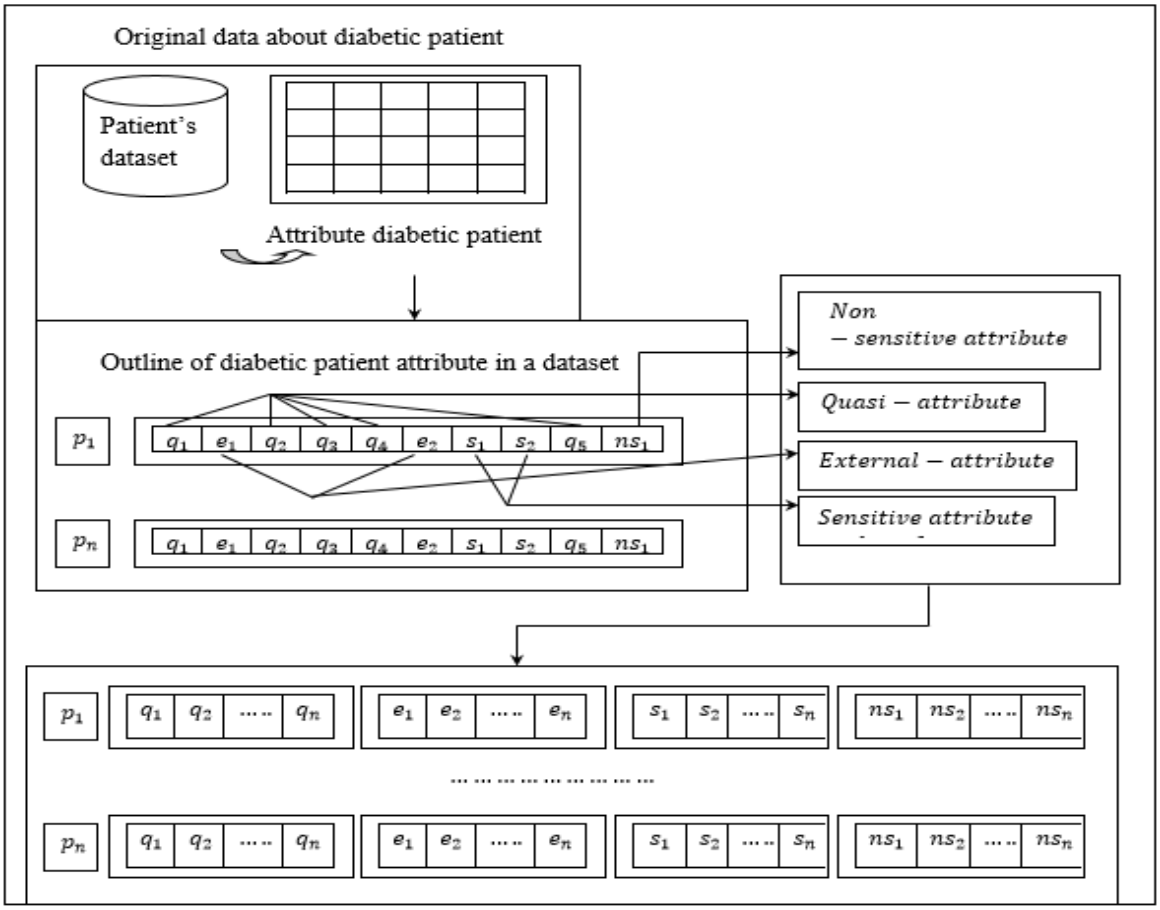

Fig. 2. Sample Distinctive Impact Context Sensitive Hashing.

Next, the impact value 'IV' is evaluated based on the equivalent class 'EC', total number of different values in column 'TV' and the 'ith' column in consideration 'C_i' respectively.

$$IV = 1 - \frac{EC(TV - C_i)}{EC(TV)} \quad (2)$$

To improve the performance of arrangements of Quasi Identifier Classes (QI-classes), heterogeneous and unstructured data were used in terms of missing values or inconsistent record, the hash function '$H(qeid)$' map similar QI-classes in place of arbitrarily map QI-classes. When QI-classes are nearer in terms of their quasi-identifiers, segregating QI-classes while preserving privacy become more ease when compared with dynamically evolving datasets [1] dispersed in a dynamic manner.

For instance, Context Hashing and hash QI-classes are integrated for every distinctive and impact values generated for total number of different values in column. The distance designates proximity between two quasi-identifiers (i.e. quasi encounter identifiers) '$qeid^x$' and '$qeid^y$'. It is described in order to incorporate Context Hashing. Let distance between two quasi-identifiers denoted as '$Dis(qeid^x, qeid^y)$' and moreover '$qeid^x = (q_1^x, q_2^x, \dots, q_n^x)$' and '$qeid^y = (q_1^y, q_2^y, \dots, q_n^y)$' respectively. Then, the distance is mathematically expressed as given below.

$$Res = Dis(qeid^x, qeid^y) = \sqrt{\sum D^2\left(q_i^x, q_i^y\right)} \quad (3)$$

Equation (3), '$D\left(q_i^x, q_i^y\right)$', represent the distance between 'q_i^x' and 'q_i^y'. With the resultant distance obtained from the above equation (3), a Context Hash function is associated to map homogeneous QI-classes. Let us assume that 'd_i' and 'd_j' are two distances, then the hash function is evolved using '(d_i, d_j)' for any two quasi-identifiers '$qeid^x, qeid^y$' via duality principle and probability function as stated in the pseudo code. The pseudo code representation of Distinctive Impact Context Hash is given below.

As given in the above Distinctive Impact Context Hash Quasi-Identifier algorithm, three steps are incorporated. First, with the big data dataset (i.e., diabetic dataset) provided as input, distinctive and impact value for total number of different values in columns are identified. Then, similar QI-classes are mapped by means of a Context Hashing distance function. Finally, with the aid of the duality principle by mapping similar QI-classes, computationally efficient similar quasi-identifiers are obtained. With application of this algorithm, optimal and computationally efficient quasi-identifiers are identified. Therefore, maximum of attributes are not selected as quasi-identifiers and only optimal attributes are selected as quasi-identifiers to improve data utility performance.

Algorithm 1: Distinctive Impact Context Hash Quasi-Identifier

Input: Patients '$P = P_1, P_2, \dots, P_n$', big data dataset 'DS', attributes '$Attr = a_1, a_2, \dots, a_n$'
Output: Computationally efficient and optimized quasi-identifiers
Step 1: **Initialize**'$qeid^x$' and '$qeid^y$' Step 2: **Initialize** classes '$Cl = \{cl_1, cl_2, cl_3, cl_4\}$', column '$C_i$' Step 3: **Begin** Step 4: **For** each big data dataset 'DS' with 'n' attributes '$Attr = a_1, a_2, \dots, a_n$' and Patients '$P$' Step 5: **For** two quasi-identifiers (i.e., quasi encounter identifiers)'$qeid^x$' and '$qeid^y$' Step 6: Evaluate distinctive value using equation (1) Step 7: Evaluate impact value using equation (2) Step 8: Evaluate distance between two quasi-identifiers using equation (3) Step 9: **If** '$Res(qeid^x, qeid^y) \leq d_j$ Step 10: **Then** '$Prob[H(qeid^x) = H(qeid^y)]$' Step 11: **End if** Step 12: **If** '$Res(qeid^x, qeid^y) \geq d_j$' Step 13: **Then** '$Prob[H(qeid^x) = H(qeid^y)]$' Step 14: **End if** Step 15: **End for** Step 16: **End for** Step 17: **Return** quasi attributes '$p = Q = \{q_1, q_2, \dots, q_n\}$' Step 18: **End**

C. Hellinger Convolutional Neural Privacy Preservation Model

With the computationally efficient quasi-identifiers retrieved, Distinctive Impact Context Hash Quasi-Identifier algorithm is used to learn features from unstructured data and initialize the CNN arrangement. Hellinger Convolutional Neural Privacy Preservation model is used to reduce significant amount of information loss while identifying quasi-identifiers and preserving it for ensuring privacy.

In this work, Hellinger Distance values are determined in each equivalence class (i.e. class other than QI-classes) to quantify the distance. After that, the cautious scrutiny is paid to QI-classes with minimum distance values. By quantifying the distance, information loss is said to be minimized and accuracy level gets increased. Then, the learned Distinctive Impact Context Hash Quasi-Identifier is utilized to train a CNN for privacy preservation. The proposed privacy-preserving data analysis architecture is illustrated in Fig. 3.

As illustrated in the above Fig. 3, with the separation between QI-classes and non QI-classes, let us assume that '$X = \{x_1, x_2, \dots, x_n\}, x_i \in R^m$', where '$X = Q = \{q_1, q_2, \dots, q_n\}$', where '$n$' represents the number of samples (i.e. other than quasi identifiers obtained in QI-classes) and 'm' represents the length of non-quasi identifiers, '$Y = SIGMOID(Wa' + b)$'. 'W' represents the weight and 'b' represents the bias respectively. With these two, activation function is mathematically expressed as given below.

$$H_{w,b} = H(x_i, W, b) = SIGMOID(Wx_i + b) \quad (4)$$

In equation (4), the sigmoid of the weight along with the bias is utilized at the average activation. The origination hypothesis is then mathematically formulated as given below.

$$P_{init} = \sum_{j=1}^{l} HD(\alpha || \alpha_j) \quad (5)$$

From the above equation (5), 'l' refers to the number of samples remained in Big Data dataset after the application of quasi-identifier detection and '$HD(.)$' refers to the Hellinger distance, quantifying the similarity between two probability distributions. This is mathematically formulated as given below.

$$H^2(P, Q) = \int \left(\sqrt{dP} - \sqrt{dQ}\right)^2 \quad (6)$$

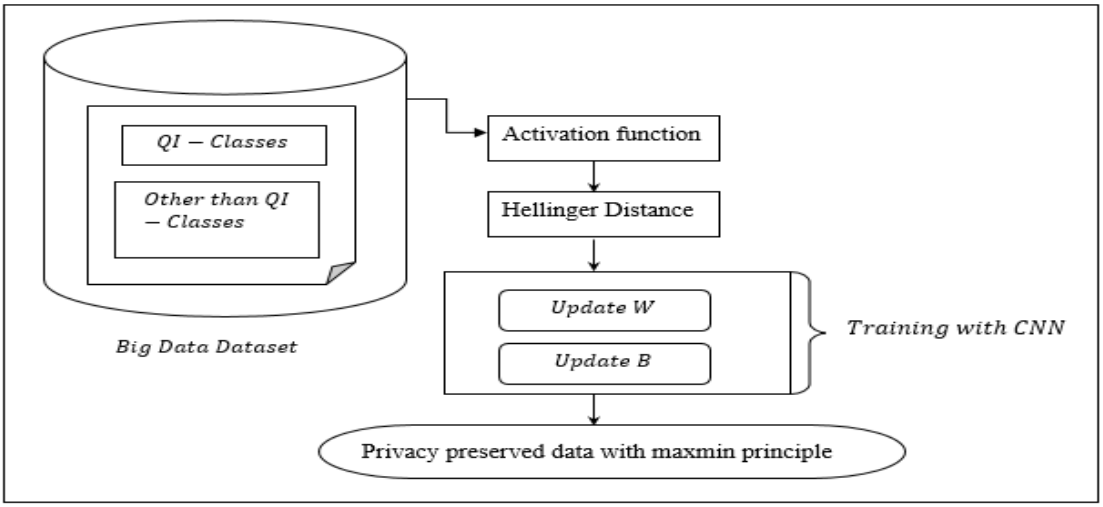

Fig. 3. Architecture of Proposed Privacy-Preserving Data Analysis.

From the above equation (6), 'P' and 'Q' refers to the two probability measures (i.e., quasi encounter identifiers and the non-quasi encounter identifiers) are continuous with respect to the third measure with a different probability measure with respect to which both 'P' and 'Q' are continuous. Next, the learned Distinctive Impact cost function is mathematically expressed as given below.

$$C_{CI}(W,b) = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\left(H_{W,b}(x_i) - (y_i)\right)^2\right) + H^2(P,Q)\right] \quad (7)$$

From the above equation (7), the cost function 'C_{CI}' is arrived at based on resultant activation function '$H_{W,b}$', its Hellinger distance '$H^2(P,Q)$' and the input vector 'x_i'. The parameters 'W_{ij}' and 'b_i' are updated and formulated as,

$$W_{ij} = W_{ij}(l) - LR\frac{\partial}{\partial W_{ij}(l)} \quad (8)$$

$$b_i = b_i(l) - LR\frac{\partial}{\partial b_i(l)} \quad (9)$$

Finally, the mean square error of the Distinctive Impact cost function is evaluated as given below.

$$C(W,b) = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\left(H_{w,b}(x_i) - (y_i)^2\right)\right) + \frac{y}{2}\sum_{i=1}^{n}\sum_{j=2}^{n}\sum_{l=1}^{n} W_{ij}(l)\right] \quad (10)$$

Finally, the efficiency of the proposed Distinctive Impact cost function is verified by original and transformed data from the Big Data dataset to train CNN for classification and to ensure privacy of the data. The pseudocode representation of Hellinger Convolutional Neural Privacy Preservation is given below.

Algorithm 2: Hellinger Convolutional Neural Privacy Preservation

Input: Input Vector '$X = \{x_1, x_2, \dots, x_n\}, x_i \in R^m$'
Output: Accurate and minimum loss privacy preserved identifiers
Step 1: **Initialize** Weight 'W', Bias 'B' Step 2: **Begin** Step 3: **For** each Input Vector 'X' Step 4: Mathematically formulate activation function using equation (4) Step 5: Obtain origination hypothesis using equation (5) Step 6: Evaluate similarity between two probability distribution using equation (6) Step 7: Mathematically formulate learned Distinctive Impact cost function using equation (7) Step 8: Update parameters weight and bias using equation (8) and (9) Step 9: Evaluate mean square error of the Distinctive Impact cost function using equation (10) Step 10: **Return** (privacy preserved identifiers) Step 11: **End for** Step 12: **End**

As given in the above Hellinger Convolutional Neural Privacy Preservation algorithm, three steps are followed. At first, non QI-classes are provided as input. After that, the process remains in generating maxmin principle (i.e., maximizing accuracy and minimizing information loss) to ensure privacy preservation for unstructured data for protecting the sensitive data. An activation function is derived by hashing QI-classes and then using Hellinger distance to minimize the information loss with minimum distance values. After that, it is provided as input for learning with updated CNN, i.e., updating weight and bias by Distinctive Impact cost function. In this manner, the accuracy level and the information loss is improved. Therefore, privacy preservation of sensitive unstructured data is carried out in efficient manner.

IV. Experimental Analysis

In this section, a detailed analysis of experimental results has been presented to evaluate the performance of Distinctive Context Sensitive and Hellinger Convolutional Learning (DCS-HCL) method for privacy preserving of sensitive unstructured big healthcare data through quasi-identifier. Based on recent state-of-the-art methods in the literature, an evaluation of privacy preservation of big healthcare data using quasi identifiers is performed in terms of run time, accuracy and information loss with respect to number of patients. The proposed DCS-HCL method is compared with two existing privacy preservation methods, Equivalence Classes with Cuckoo Filter (ENCC) [1] and integrated anonymization and reconstruction [2]. The result analysis shows that DCS-HCL method ensures data utility with higher accuracy and minimum information loss as well as runtime consumption for support fine-grained access control with big healthcare data when compared to state-of-the-art works.

A. Dataset Description

The Diabetes 130-US hospitals for years 1999-2008 Data Set [20] is used for conducting the experiments. The dataset comprises 10 years of clinical care obtained from 130 US hospitals and integrated delivery networks and covers 50 features denoting patient and hospital outcomes. Certain attributes present in dataset are patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, numbers of outpatient, inpatient, and emergency visits in year before hospitalization, etc. With the aid of this dataset experiments for privacy preserving is conducted using Python. In this section, performance metrics, namely run time, accuracy and information loss with respect to number of patients are considered for privacy preservation.

1) Run time evaluation: With the big healthcare data being shared between the patients and in public domain, the run time involved should be minimum otherwise the data is said to be loss or privacy is said to be compromised. However, a significant amount of time is said to be consumed while preserving privacy of big healthcare data. The run time involved is mathematically expressed as given below.

$$RT = \sum_{i=1}^{n} P_i * Time\ [PP] \quad (11)$$

From the above equation (11), the run time 'RT' involved in preserving the privacy of big healthcare data using quasi identifiers is evaluated based on the number of patients considered during simulation 'P_i' and the time involved in preserving the privacy '$Time\ [PP]$'. It is measured in terms of milliseconds (ms).

2) Accuracy evaluation: The accuracy maintenance for quasi identifiers is an important issue in preserving privacy of big healthcare data. The accuracy refers to the amount of sensitive data being preserved during the process involved in privacy preservation using quasi identifiers. The accuracy measure is mathematically expressed as given below.

$$A = \sum_{i=1}^{n} \frac{P_{AP}}{P_i} * 100 \quad (12)$$

From the above equation (12), the accuracy 'A' is measured on the basis of the number of patients 'P_i' considered for simulation and the patients data accurate preserved 'P_{AP}'. It is measured in terms of percentage (%).

3) Information loss evaluation: During the privacy preservation of big healthcare data, certain amount of information gets lost. However, the information loss should be lesser so that higher amount of information is said to be preserved. The information loss is mathematically evaluated as given below.

$$IL = \sum_{i=1}^{n} \frac{P_{dc}}{P_i} * 100 \quad (13)$$

From the above equation (13), the information loss 'IL' is obtained on the basis of the number of patients considered for conducting simulation 'P_i' and the number of patient data compromised 'P_{dc}' during privacy preservation. It is expressed in terms of percentage (%).

V. Result and Discussions

In this section, a series of experiments are conducted to verify the significance of the proposed method Distinctive Context Sensitive and Hellinger Convolutional Learning (DCS-HCL) using Diabetes 130-US hospitals dataset. Then, three commonly used evaluation metrics, run time, accuracy and information loss are used to compare the performance of the privacy preservation with two existing methods, Equivalence Classes with Cuckoo Filter (ENCC) [1] and integrated anonymization and reconstruction [2].

A. Performance Measure of Run Time

First, the performance analysis of run time is carried out. Table I shows the run time comparison of the proposed DCS-HCL with the existing methods, ENCC [1] and integrated anonymization and reconstruction [2] using 10 different values of 'P_i'. The rise in 'P_i' value causes an increase in the run time for all the three methods due to the increase in the records and their corresponding similar quasi-identifiers. The proposed method run time values are lesser than existing methods [1] and [2] in most cases because the proposed method selects only the optimized identifiers as the quasi-identifiers.

TABLE I. Analysis Results of Runtime using DCS-HCL, ENCC [1] and Integrated Anonymization and Reconstruction [2]

Number of patients	Run time (ms)		
	DCS-HCL	ENCC	Integrated anonymization and reconstruction
500	42.5	57.5	72.5
1000	75.35	105.35	125.35
1500	90.25	125.45	140.55
2000	105.35	140.55	175.55
2500	125.45	195.35	225.35
3000	140.55	215.25	255.85
3500	175.35	225.35	315.55
4000	190.15	240.55	335.25
4500	200.35	280.15	350.55
5000	225.55	315.55	385.55

Fig. 4 given shows the run time values of the proposed DCS-HCL method and its comparison with the existing two methods [1] and [2] on Diabetes 130-US hospitals dataset. From the figure, it is inferred that the run time linearly increases with the increase in number of patients during privacy preservation. With the simulation conducted for '500' numbers of patients for preserving the privacy of big healthcare data using quasi identifiers, the run time involved for preserving single patient is '0.085ms' by DI-CSH model. The overall run time for '500' patients was found to be '42.5ms', '57.5ms' and '72.5ms' using DCS-HCL, [1] and [2] respectively. From the results, the run time using DCS-HCL is comparatively lesser than [1] and [2]. The reason behind the improvement is the application of Distinctive Impact Context Sensitive Hashing (DI-CSH) model. By applying this model, the base essential quasi attributes are identified by mapping the similar QI-classes hash function when compared to the arbitrarily mapped QI-classes. With this, the run time involved in preserving the privacy of big healthcare data using DCS-HCL is comparatively lesser than 28% compared to [1] and 42% compared to the [2], respectively.

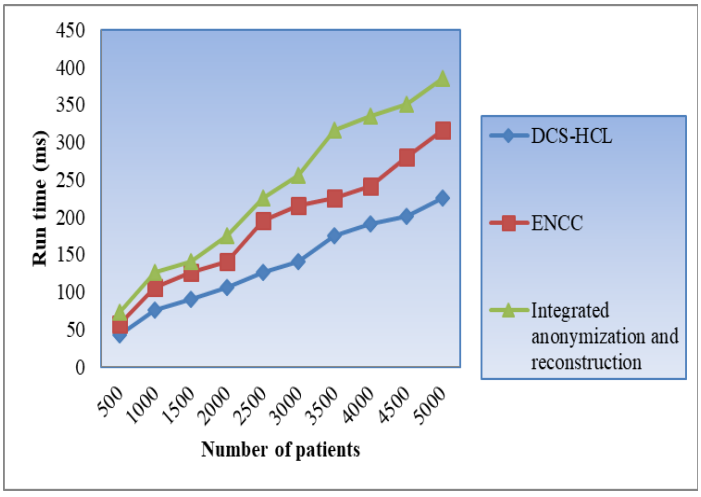

Fig. 4. Graphical Representation of Run Time.

B. Performance Measure of Accuracy

Secondly, the performance analysis of accuracy is investigated. Furthermore, the results were performed to compare the accuracy of the proposed DCS-HCL with two existing methods, ENCC [1] and integrated anonymization and reconstruction [2] using 10 different values of 'P_i'. Accuracy results are shown in Table II. From the results, it is clear that as 'P_i' value is increased, the accuracy values of all the three methods decreased. The proposed method yields higher accuracy when compared to the existing privacy preservation methods in most cases by controlling the information loss via distance quantification. In contrast, the existing privacy preservation methods not apply the concept of distance quantification to control the information loss and attain relatively lesser accuracy.

TABLE II. ANALYSIS RESULTS OF ACCURACY USING DCS-HCL, ENCC [1] AND INTEGRATED ANONYMIZATION AND RECONSTRUCTION [2]

Number of patients	Accuracy (%)		
	DCS-HCL	ENCC	Integrated anonymization and reconstruction
500	97	95	92
1000	96.35	92.15	90.25
1500	96.15	90.55	88.35
2000	96	88.35	86.15
2500	95	86.25	84.35
3000	94.35	85.15	82.15
3500	94.15	84.35	80
4000	94	82.15	78.85
4500	93.25	81.55	75.35
5000	92	80	75

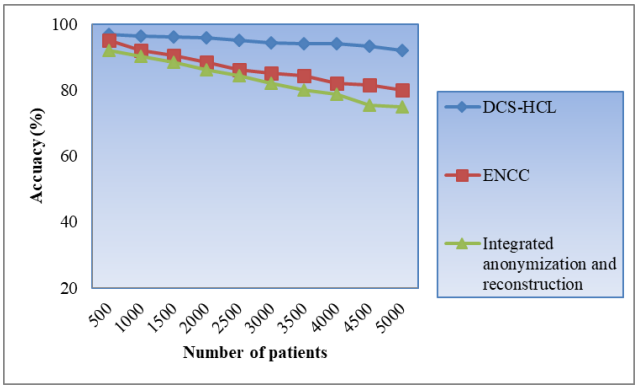

Fig. 5. Graphical Representation of Accuracy.

Fig. 5 illustrated above shows the average accuracy values obtained using three different methods, DCS-HCL, [1] and [2]. It is observed that, rise in the number of patient causes decrease in the accuracy due to less modification in the QI original values. The rationale regarding better accuracy of the proposed method compared to the existing methods [1] and [2] is derived from the fact that minimum distance consistency is maintained in the quasi-identification process. With '500' number of patients considered for simulation to evaluate the privacy preservation of big healthcare data and '485' number of patients data accurately preserved, the overall accuracy using DCS-HCL was found to be '97%', '95%' using [1] and '92%' using [2]. This is because of applying Hellinger Convolutional Neural Privacy Preservation algorithm in proposed model. A maxmin principle is applied for unstructured data. With this objective, an activation function is derived by hashing QI-classes. Hellinger distance maximizes the accuracy involved in preserving the privacy. In this manner, the accuracy of privacy being preserved for big healthcare data is said to be improved using DCS-HCL by 10% compared to [1] and 14% compared to [2], respectively.

C. Performance Measure of Information Loss

Finally, the information loss involved is presented in this section. To further demonstrate the effectiveness of the proposed method, information loss values have been measured and compared with the result of the two existing privacy preservation methods [1] and [2]. Results are shown in Table III. The proposed DCS-HCL method has produced lesser information loss value than other privacy preservation methods, [1] and [2]. The proposed method applies the concept of Hellinger Distance in privacy preservation process and maintains the QI's values consistency resulting in higher data utility to reduce information loss.

TABLE III. ANALYSIS RESULTS OF INFORMATION LOSS USING DCS-HCL, ENCC [1] AND INTEGRATED ANONYMIZATION AND RECONSTRUCTION [2]

Number of patients	Information loss (%)		
	DCS-HCL	ENCC	Integrated anonymization and reconstruction
500	3	5	8
1000	3.5	6.25	9.35
1500	4	6.55	10
2000	4.25	6.85	10.55
2500	4.45	7	10.85
3000	4.85	7.25	11.35
3500	5	7.45	11.85
4000	6.35	7.85	12.45
4500	8	9	14
5000	8.15	10.15	15.35

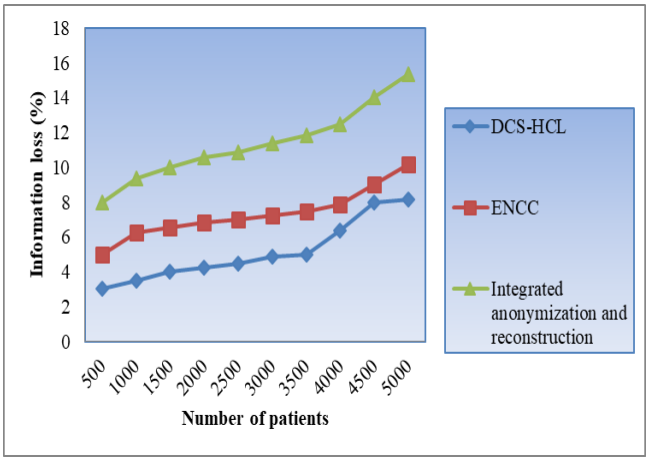

Fig. 6. Graphical Representation of Information Loss.

Fig. 6 shown above provides the graphical representation of information loss using three different methods. From the figure it is inferred that the information loss is linearly increased with the increase in the number of patients. This is owing to the fact that with the increase in the number of patients, the attributes involved in preserving the privacy also increases and obviously compromising the sensitive unstructured data. However with the simulations conducted for preserving the privacy with '500' number of patients '15' number of patients data compromised during the process and the overall information loss using DCS-HCL was observed to be '3', '5' using [1] and '8' using [2], respectively. From the results, it is inferred that the information loss using DCS-HCL is found to be comparatively lesser when compared to [1] and [2]. The improvement is due to the application of Hellinger Convolutional Neural Privacy Preservation model. Distinctive Impact cost function is used to update weight and loss for contributing to higher data utility. With this function, data utility is said to be improved and results in the minimization of information loss. The information loss using DCS-HCL is said to be reduced by 31% compared to [1] and 56% compared to the [2], respectively.

VI. Conclusion

In this paper, the quasi-identifiers is used in big healthcare datasets to ensure the privacy requirements and to achieve high data utility simultaneously with minimum run time and information loss. Distinctive Impact Context Sensitive Hashing (DI-CSH) model is used for privacy preservation by extracting base essential quasi attributes. The designed model access only a part of attributes in data asset rather than access all data records as required by existing methods. To further enhance the performance of privacy preserving mechanism, Hellinger Convolutional Neural Privacy Preservation model is used to preserve the data via maxmin principle. Thus, the number of data nodes across QI-group gets reduced considerably with minimum information loss. Evaluation results with Diabetes 130-US hospitals dataset have demonstrated in DI-CSH model in terms of run time, accuracy and information loss over existing methods for privacy preservation on big healthcare data set. In future, the accuracy level can be further enhanced by using deep learning algorithms. In addition, cryptosystem can be included in order to enhance the security level during data communication in healthcare and other applications.

References

[1] O. Temuujin, J. Ahn and D. Im, "Efficient L-Diversity Algorithm for Preserving Privacy of Dynamically Published Datasets," IEEE Access, vol. 7, pp. 122878-122888, September 2019.

[2] Y. Sei, H. Okumura, T. Takenouchi and A. Ohsuga, "Anonymization of Sensitive Quasi-Identifiers for l-Diversity and t-Closeness," IEEE Transactions on Dependable and Secure Computing, vol. 16, no. 4, pp. 580- 593, August 2019.

[3] M. Binjubeir, A. A. Ahmed, M. A. B. Ismail, A. S. Sadiq and M. K. Khan, "Comprehensive Survey on Big Data Privacy Protection," IEEE Access, vol. 8, pp. 20067– 20079, January 2020.

[4] K. Abouelmehdi, A. B. Hessane and H. Khaloufi, "Big healthcare data: preserving security and privacy," J Big Data, Springer, vol. 5, no. 1, pp. 1-18, July 2018.

[5] G. Kapil, A. Agrawal, A. Attaallah, A. Algarni, R. Kumar and R. A. Khan, "Attribute based honey encryption algorithm for securing big data: Hadoop distributed file system perspective," Peer J Computer Science, vol. 6, pp. 1-27, February 2020.

[6] W. Mahanan, W. A. Chaovalitwongse and J. Natwichai, "Data anonymization: a novel optimal k-anonymity algorithm for identical generalization hierarchy data in IoT," Service Oriented Computing and Applications, Springer, vol. 14, pp. 89-100, February 2020.

[7] P. S. Rao and S. Satyanarayana, "Privacy preserving data publishing based on sensitivity in context of Big Data using Hive," J Big Data, Springer, vol. 5, no. 20, pp. 1-20, August 2018.

[8] P. R. M. Rao, S. M. Krishna and A. P. S. Kumar, "Privacy preservation techniques in big data analytics: a survey," J Big Data, Springer, vol. 5, no. 33, pp. 1-12, July 2018.

[9] I. Ali, E. Khan and S. Sabir, "Privacy-preserving data aggregation in resource-constrained sensor nodes in Internet of Things: A review," Future Computing and Informatics Journal, Elsevier, vol. 3, no. 1, pp. 41-50, June 2018.

[10] J. Wang, K. Du, X. Luo and X. Li, "Two privacy-preserving approaches for data publishing with identity reservation," Knowledge and Information Systems, Springer, vol. 60, pp.1039–1080, June 2018.

[11] R. Khan, X. Tao, A. Anjum, H. Sajjad, S. R. Malik, A. Khan and F. Amiri, "Privacy Preserving for Multiple Sensitive Attributes against Fingerprint Correlation Attack Satisfying c-Diversity," Wireless Communications and Mobile Computing, Hindawi Publishing Coorporation, vol. 2020, pp. 1-18, January 2020.

[12] X. Li and Z. Zhou, "A generalization model for multi-record privacy preservation," J Ambient Intell Human Comput, Springer, vol. 11, pp. 2899–2912, 2020.

[13] C. Dhasarathan, V. Thirumal and D. Ponnurangam, "A secure data privacy preservation for on-demand cloud service," Journal of King Saud University – Engineering Sciences, Elsevier, vol. 29, no. 2, pp. 144-150, April 2017.

[14] J. W. Kim, B. Jang and H. Yoo, "Privacy-preserving aggregation of personal health data streams," PLoS ONE, vol. 13, no.11, pp. 1-15, November 2018.

[15] J. Xu, K. Xue, S. Li, H. Tian, Ji. Hong, P. Hong and N. Yu, "Healthchain: A Blockchain-Based Privacy Preserving Scheme for Large-Scale Health Data," IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8770 – 8781, October 2019.

[16] K. Abouelmehdi, A. B. Hessane and H. Khaloufi, "Big healthcare data: preserving security and privacy," J Big Data, Springer, vol. 5, no. 1, pp. 1-18, February 2018.

[17] P. Jain, M. Gyanchandani and N. Khare, "Big data privacy: a technological perspective and review," J Big Data, vol. 3, no. 25, pp.1-25, September 2016.

[18] K. Arava, S. Lingamgunta, "Adaptive k-Anonymity Approach for Privacy Preserving in Cloud," Arab J Sci Eng, Springer, vol. 45, pp. 2425–2432, July 2019.

[19] 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," Granular Computing, Springer, vol. 5, pp. 71-84, August 2018.

[20] B. Strack, J. P. D. Shazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios and J. N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, Hindawi Publishing Corporation, vol. 2014, pp. 1-11, April 2014.